 doi:10.20944/preprints202206.0390.v1

Article

Multi-scale Object Detection with Pixel Attention Mechanism in Complex Background

Jinsheng Xiao [1,*], Haowen Guo [1], Shuhao Zhang [1], Yuntao Yao [1], Jian Zhou [2] and Zhijun Jiang [3,4]

[1] School of Electronic Information, Wuhan University, Wuhan, China, 430064
[2] State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan, China, 430079
[3] Aerospace System Development Research Center, China Aerospace Science and Technology Corporation, Beijing, China, 100094
[4] Qian Xuesen Laboratory of Space Technology, Beijing, China, 100094
* Correspondence: xiaojs@whu.edu.cn; Tel.: +86-18971297802

Abstract: The object detection task is usually affected by complex backgrounds. In this paper, a new image object detection method is proposed, which can perform multi-feature selection on multi-scale feature maps. By this method, a bidirectional multi-scale feature fusion network is designed to fuse semantic features and shallow features to improve the detection effect of small objects in complex backgrounds. When the shallow features are transferred to the top layer, a bottom-up path is added to reduce the number of network layers experienced by the feature fusion network, reducing the loss of shallow features. In addition, a multi-feature selection module based on the attention mechanism is used to minimize the interference of useless information on subsequent classification and regression, allowing the network to adaptively focus on appropriate information for classification or regression to improve detection accuracy. Because the traditional five-parameter regression method has severe boundary problems when predicting objects with large aspect ratios, the proposed network treats angle prediction as a classification task. The experimental results on the DOTA dataset, the self-made DOTA-GF dataset and the HRSC 2016 dataset show that, compared with several popular object detection algorithms, the proposed method has certain advantages in detection accuracy.

Keywords: Object detection; Feature fusion network; Multiple feature selection; Angle prediction; Pixel Attention Mechanism

1. Introduction

Object detection in remote sensing and UAV (Unmanned Aerial Vehicle) imagery is important in a variety of sectors, including resource monitoring, national defense, and urban planning[1,2]. Unlike typical optical images, optical remote sensing images always have their own unique qualities, such as numerous sizes of objects, arbitrary object direction, and complex backgrounds that take up the majority of the image. Many remote sensing image object detection algorithms borrow ideas from text detection algorithms like RRPN[3] because the arbitrariness of the object direction in remote sensing images has a lot in common with text detection[4]. However, due to the peculiar nature of remote sensing images, directly applying text detection algorithms to remote sensing image object detection frequently yields unsatisfactory results.

For scale-differences between classes, the feature pyramid network (FPN) [5] is commonly utilized in object detection of various remote sensing images. Shallow features in FPN, on the other hand, must transit through numerous layers to reach the top layer, resulting in significant information loss. To improve the detection effect of small objects, certain algorithms[6–8] optimize the structure of FPN. The traditional technique to counteract the arbitrariness of object orientation in remote sensing images is to raise the regression parameters to estimate the angles[9,10], which has a severe problem of boundary

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 June 2022 doi:10.20944/preprints202206.0390.v1

discontinuities[11]. To tackle the boundary problem, the IoU constant factor is added to the smooth L_1 loss to make correct angle predictions. Because the complex background contains a lot of noise[12], [13] uses a multi-scale feature extraction method to enhance each feature map with a visual attention mechanism to lessen the impact of background noise on object detection. After using the region proposal network (RPN) to acquire regional suggestions, [14] uses the location-sensitive score map to anticipate the target's local location, and specifies that it can only be classified as a given category after reaching a certain local feature similarity. To some extent, this strategy can also eliminate the influence of the background.

In summary, the main issues with remote sensing image object detection are numerous scales, complex backgrounds, and poor angle prediction. This paper proposes a new remote sensing image object detection algorithm to address these issues, and the framework is shown in Fig. 1.

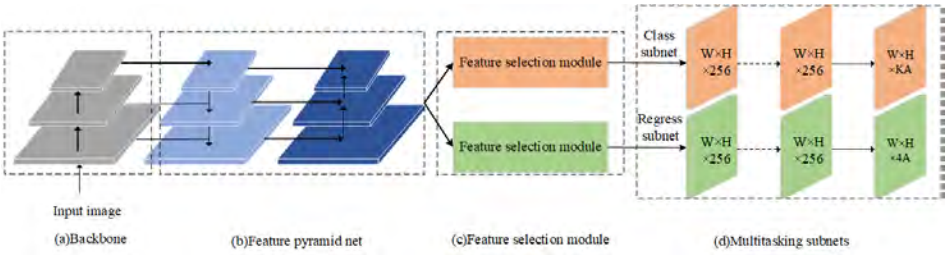


Figure 1. The network structure of proposed method. It can be divided into four parts: (a)Input image, (b)Feature pyramid net, (c)Feature selection module, (d) Multitasking subnets

We use a single-stage rotation detector for multi-scale objects to retain good detection accuracy and speed. The first step is to build a bidirectional multi-scale feature fusion network. To prevent information loss during the transfer of shallow features to the top layer, a bottom-up path is added to merge high-level semantic information and shallow features. Second, a multi-feature selection module based on the attention mechanism is designed to reduce the complex background's influence on object detection. The visual attention mechanism allows the network to focus on more significant information while avoiding background noise, and choose appropriate features for classification and regression tasks. Third, to increase the accuracy of direction prediction, the proposed network treats angle prediction as a classification problem. The distribution vectors of the category labels are smoothed using the circular smooth label, which divides the angles into 180 categories. The majority of the data in open-source remote sensing image object detection datasets comes from Google Earth, with only a minor amount coming from domestic satellites. And there is a lack of military targets. As a result, we gathered some GF-2 and GF-6 images and created a new dataset named DOTA-GF. On DOTA [15] dataset and DOTA-GF dataset, the proposed method is compared to many popular remote sensing image object detection algorithms. This work makes the following contributions:

- A bidirectional multi-scale feature fusion network is built for high-precision multi-scale object detection in remote sensing images. It is the first work that we are aware of that achieves high-precision object detection in complex backgrounds.
- The multi-feature selection module (MFSM) based on attention mechanism is designed to reduce the influence of useless features in feature maps in complex backgrounds with a lot of noise.
- We proposed a novel remote sensing image object detection algorithm that included a bidirectional multi-scale feature fusion network and a multi-feature selection module. With extensive ablation experiments, we validate the effectiveness of our approach on the standard DOTA dataset and a customized dataset named DOTA-GF. Our proposed method achieves a mAP of 65.1% with ResNet50 backbone in DOTA dataset and 64.1% with ResNet50 backbone in DOTA-GF dataset when compared to state-of-the-art methods.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 June 2022 doi:10.20944/preprints202206.0390.v1

2. Related work

2.1. Object Detection Algorithms Based on Deep Learning

Object detection algorithms based on deep learning are mainly divided into two categories, one-stage algorithms and two-stage algorithms. The series of algorithms of R-CNN are typical two-stage method,including R-CNN, Fast R-CNN, and Faster R-CNN [16]. Fast R-CNN proposed RoIpooling and used convolution network to achieve regression and classification, while Faster R-CNN used the RPN (RegionProposal Network) to replace selective search and shared feature map with the subsequent classification network. The one-stage methods extract feature maps and predict the categories and locations simultaneously. The SSD and YOLO are two typical one-stage methods [17]. Different from the two-stage methods, the one-stage methods are influenced by the problem of category imbalance during detection. To tackle such problem, focal loss [18] is proposed to suppress category imbalance in one-stage methods.

2.2. Arbitrary-oriented object detection

Arbitrary-oriented object detection has been widely used in remote sensing image, aerial image, natural scene text, etc. These detectors also use rotated bounding boxes to describe positions of objects, which are more accurate than those using horizontal bounding boxes. Recently, many detectors have been proposed. For example, RRPN [3] used rotating anchors to improve the qualities of region proposals. R2CNN is a multi-tasking text detector that identifies both rotated and horizontal bounding boxes at the same time. However, object detection in remote sensing images is more difficult, due to multiple categories, multiple scales, complex backgrounds. So many Arbitrary-oriented object detection in remote sensing images has been proposed. R3Det [10] proposed an improved one-stage rotated object detector for accurate object localization by solving the feature misalignment problem. SCRDet [19] proposed an IoU-smooth L_1 loss to solve the loss discontinuity caused by the angular periodicity. [20] proposed a Anchor-free Oriented Proposal Generator (AOPG) that abandoned the horizontal boxes-related operations from the network architecture. The AOPG produced coarse oriented boxes by Coarse Location Module in an anchor-free manner and refined them into high-quality oriented proposals. [21] proposed an effective oriented object detection method, termed Oriented R-CNN. Oriented R-CNN is a general two-stage oriented detector. In the first stage, the oriented Region Proposal Network directly generates high-quality oriented proposals in a nearly cost-free manner. The second stage is oriented R-CNN head for refining oriented regions of interest and recognizing them.

3. The proposed algorithm

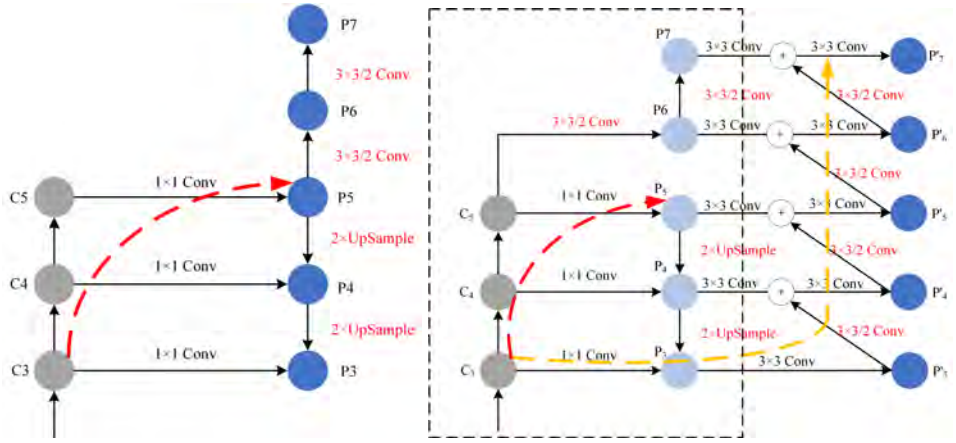
We give an overview of our algorithm as sketched in Figure 1. It consists of four parts: the backbone, the bidirectional multi-scale feature fusion network, the multi-feature selection module based on attention mechanism and the multi-task subnets. We use the ResNet50 [22] as our backbone. The bidirectional multi-scale feature fusion network is responsible for fusing the high-level semantic information and the shallow features output by the backbone. The multi-feature selection module based on the attention mechanism can select features that are appropriate for classification and regression. After feature selection, the multi-scale feature maps are sent into the classification and regression sub-networks, respectively. Only the center points, width, and height of the bounding boxes are predicted by the regression subnet in this case. Through the classification subnet, the categories and angles are predicted.

3.1. Bidirectional multi-Scale feature fusion network

In the early object detection algorithms, such as Faster R-CNN [23], the subsequent classification and regression are usually performed on the feature map of the last layer of the backbone, which is less computationally expensive. But for the multi-scale object detection, the information of a single-layer feature map is not enough. In 2017, He et

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 June 2022 doi:10.20944/preprints202206.0390.v1

al. proposed FPN [18], which fuses high-level features and low-level features, and uses multi-scale fusion feature maps for subsequent detection. RetinaNet [24] also follows the idea of FPN to build a feature pyramid net, as shown in Figure 2(a).



(a) Feature pyramid structure in RetinaNet network (b) Bidirectional multi-scale feature fusion network

Figure 2. The network structure of feature fusion network. The red dotted line : the bottom-up path of the shallow information transmitted to the high level, the yellow dotted : the new bottom-up path, $1 \times 1Conv$: convolution operation with 1×1 convolution kernel, $2 \times UpSample$: the double upsampling operation by bilinear interpolation, $3 \times 3/2Conv$: convolution operation with 3×3 convolution kernel and a stride of 2, $3 \times 3Conv$: convolution operation with 3×3 convolution kernel and a stride of 1

Compared with the features extracted only through the last layer of convolution, FPN can use more high-level semantic information and detailed information. The red dotted line in Figure 2(a) indicates that in FPN, because of the bottom-up path, shallow features need to pass through multilayer networks to reach the top layer, and the information loss is more serious. Taking ResNet50 as an example, the transfer of the C_3 layer to the C_5 layer needs to go through 27 layers of convolution operations, as shown in Figure 3. The shallow details contained in P_5, P_6 and P_7 are lacking to be used for subsequent detection. With the addition of the bottom-up fusion path, the detailed texture features of the C_3 layer can be transferred to P_5', P_6' and P_7' with only a few layers, as indicated by the yellow dotted line in Figure 2(b). Therefore, the loss of shallow features is reduced.

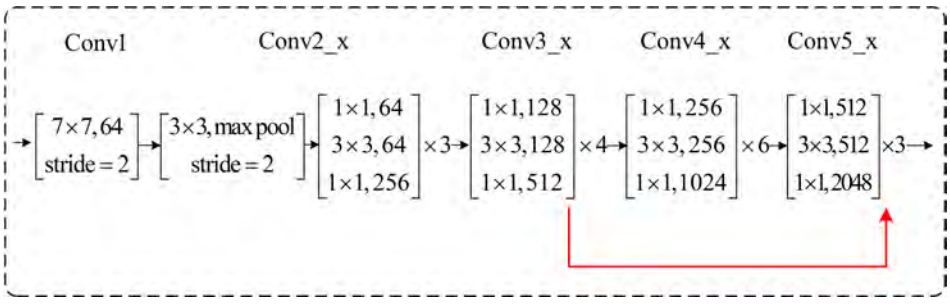


Figure 3. ResNet50 network structure, the red arrow indicates the path from C_3 to C_5.

Therefore, we design a new feature fusion network, and a bottom-up path is added to reduce the number of network layers experienced when the shallow features are transferred to the top layer, thereby reducing the loss of shallow features. The detailed information of the network is shown in Figure 2(b).

As shown in Figure 2(b), $1 \times 1Conv$ represents using 1×1 convolution kernel to perform convolution operations and change the number of channels in the feature map.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 June 2022 doi:10.20944/preprints202206.0390.v1

2×UpSample represents the double upsampling operation of the feature map using bilinear interpolation. $3 \times 3/2Conv$ means using a 3×3 convolution kernel to perform a convolution operation with a stride of 2, reducing the size of the feature map to half of the original size. The output of the backbone is $C_i(i \in 3-5)$, and the feature map after feature fusion is $P_i(i \in 3-7)$. Using 1×1 convolution to reduce the dimension of C_5 to get P_5, C_5 is double downsampled to get P_6, P_6 is double downsampled to get P_7. The result of double upsampling of P_5 is fused with C_4 to obtain P_4. The result of double upsampling of P_4 is fused with C_3 to obtain P_3. $P_i(i \in 3-7)$ combines the information of C_3, C_4, and C_5 at the same time, and contains low-level detailed information and high-level semantic information. Although it has a strong characterization ability for multi-scale objects, the transmission path of shallow features to higher layers is too long, and the feature loss is severe. Therefore, we add a bottom-up path, as shown in the yellow dotted line in Figure 2(b). $3 \times 3Conv$ represents a convolution operation with a stride of 1 and a 3×3 convolution kernel. Perform a 3×3 convolution operation on P_3 to obtain P_3'. The result of P_4' after 3×3 convolution and the result of double downsampling of P_3' are fused to obtain P_4'. Then P_5', P_6' and P_7' are obtained in the same way.

3.2. Multi-Feature selection module based on attention mechanism

The complex background of satellite remote sensing images occupies a large area of the whole image. The images taken by domestic satellites, such as GF-2 and GF-6, are not as clear as Google Earth images, which leads to more complex backgrounds of the images, unclear object textures and sometimes interference from cloud and fog. Directly inputting feature maps of different scales into the subsequent classification and regression sub-networks often fails to obtain ideal results. In recent years, the attention mechanism has achieved great success in computer vision tasks, such as image classification [24] and semantic segmentation [25]. Here we designed a MFSM. MFSM uses the pixel attention mechanism to select the features suitable for classification and regression, respectively, to reduce the influence of useless information in the feature maps. Different from the spatial attention mechanism, which learns the degree of dependence on different locations in space[26] , the pixel attention mechanism learns the degree of dependence on each pixel, and adjusts the feature map according to the degree of dependence.

The general one-stage object detection algorithms directly input $P_i'(i = 3,4,5,6,7)$ into classification subnet and regression subnet. The classification subnet is to predict the category of the bounding box. The regression subnet is primarily responsible for predicting the specific position of the bounding box. The purposes of the two subnets are different. It is dinappropriate to use the same feature maps to perform classification and regression tasks at the same time. Therefore, we design the MFSM. As shown in Figure 4, the multi-scale feature maps are obtained through the feature fusion network, and then are input into two feature selection modules respectively. Finally, the feature maps after feature selection are input into the classification subnet and regression subnet.

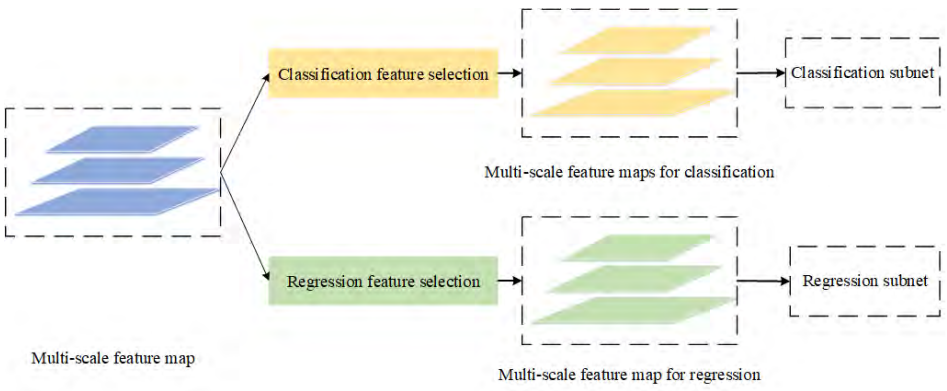


Figure 4. Multi-feature selection of multi-scale feature maps

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 June 2022 doi:10.20944/preprints202206.0390.v1

The network details of the feature selection module for classification and the feature selection module for regression are the same, as shown in Figure 5.

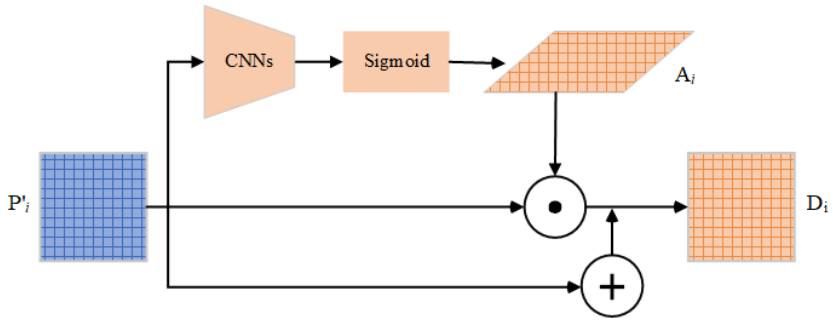


Figure 5. Detailed information of the multi-feature selection module. CNNs: four layers of 3×3 convolution, $\odot$: hadamard product,$\oplus$ Matrix addition.

The input of the module is the multi-scale feature maps input $P'_i(i = 3,4,5,6,7)$ output by the feature fusion network, and the output of the module is a series of feature maps input $D_i(i = 3,4,5,6,7)$ with the same dimensions as the input. The processing process for each input input $P'_i(i = 3,4,5,6,7)$ is shown in Figure 5 and Equation 1-2:

$$A_i = \sigma[\phi_i(P'_i)] \tag{1}$$

$$D_i = A_i \odot P'_i + P'_i \tag{2}$$

where $\phi_i(P'_i)$ means performing four layers of 3×3 convolution on P'_i. σ is the sigmoid function which converts the value of $\phi_i(P'_i)$ into [0-1] to get A_i, so that it can converge faster during training. Finally, the result of multiplying the corresponding elements of P'_i and A_i is added to P'_i. The multiplication operation can make the value of the functional information in P'_i larger and the value of the useless information smaller. The addition operation refers to the idea of the residual network [22], which can make the network converge faster. This design can make the network adaptively select features suitable for classification or regression.

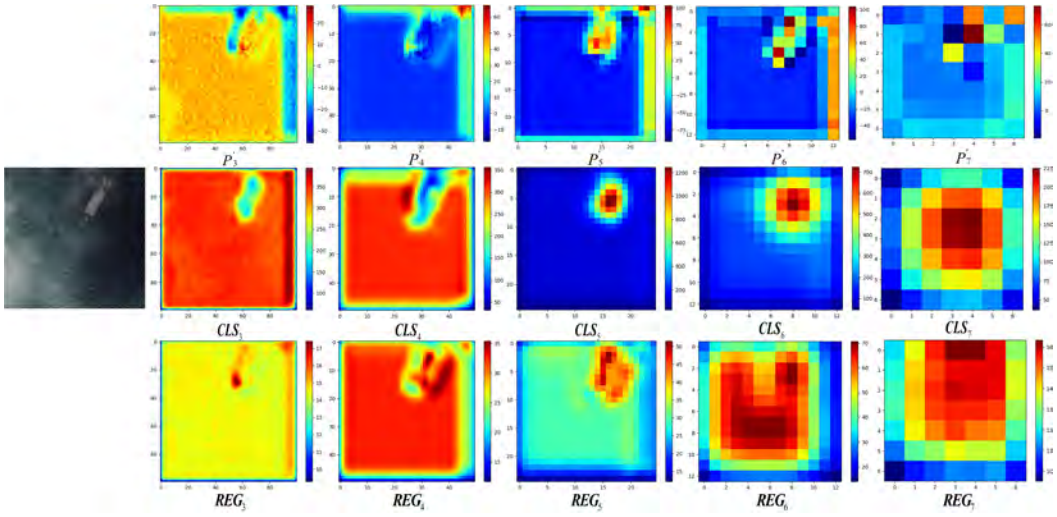


Figure 6. Visualization results of multi-scale feature maps.From top to bottom, there are the multi-scale feature maps $P'_i(i = 3,4,5,6,7)$, the multi-scale feature maps $CLS_i(i = 3,4,5,6,7)$ used for classification tasks, and the multi-scale feature map $REG_i(i = 3,4,5,6,7)$ used for regression tasks.

Figure 6 shows a remote sensing image with cloud interference and visualization results of its feature maps. The feature map $P'_i(i = 3,4,5,6,7)$ is obtained by the feature

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 June 2022 doi:10.20944/preprints202206.0390.v1

fusion network. P_i' is input into the multi-feature selection network, and the feature map $CLS_i(i = 3,4,5,6,7)$ for the classification prediction task and the feature map $REG_i(i = 3,4,5,6,7)$ for the bounding box prediction task are obtained. In Figure 6, three rows from top to bottom are P_i', $CLS_i(i = 3,4,5,6,7)$ and $REG_i(i = 3,4,5,6,7)$. Five columns from left to right are feature maps of the 3rd, 4th, 5th, 6th and 7th layers respectively. For the ship in Figure 6, the P_3' and P_4' in the multi-scale feature maps have a greater response. From P_3', P_4', CLS_3, CLS_4 and REG_4, REG_4, we can see that after feature selection, the feature map has a stronger response in the object area. It shows that the multi-feature selection module based on the attention mechanism can select features suitable for classification tasks and regression tasks from multi-scale feature maps and improve the detection accuracy.

3.3. Accurate acquisition of target direction based on angle classification

At present, most mainstream algorithms use the idea of regression for angle prediction, and the bounding box is determined by five parameters. The five-parameter regression method has problem of boundary discontinuities [11], which will make prediction box inaccurate. Figure 7 shows the results of the prediction angle based on the five-parameter regression method. As can be seen from the red boxes in the figure, there is a significant difference between the angles of the detected bounding boxes and the angles of the actual objects, including the large vehicles on the left and the ships on the right.

(a) Large vehicle (b) Ship

Figure 7. The regression inaccuracy of the five-parameter method. RetinaNet is the base model. The cars and ships in the red box have not been accurately detected, and the angles between the prediction boxes and the ground truth are much different.

Aiming at the loss discontinuity of five-parameter regression, this paper treats the angle prediction as a classification task [27]. The angles are divided into 180 categories. We find that directly dividing the angle into 180 categories will lead to low fault tolerance of adjacent angles. Therefore, the circular smooth label (CSL)[27] is used in this paper. The CSL expression is as follows:

$$S(x) = \begin{cases} f(x), & if\ \theta + \tau \leq x \leq \theta + \tau, \\ 0, & otherwise \end{cases} \tag{3}$$

where γ denotes the radius of the window. θ is the angle of the current ground truth. The circular smooth label is different for each ground truth. $f(x)$ is the window function, and the Gaussian function is used here, as shown in Equation 4:

$$Gaussian(x) = ae^{\frac{-(x-b)^2}{2c^2}} \tag{4}$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 June 2022 doi:10.20944/preprints202206.0390.v1

where a, b and c are constants ($a > 0$), in this paper, $a = 1$, $b = 0$, and c is equal to the radius of the window function which is set to be 6. The CSL [27] can increase the error tolerance to adjacent angles.

In the paper, the angles of the bounding box are divided into 180 categories. If the angle of a ground truth is -90°, the traditional label of the angle is as follows:

$$label = (1, 0, 0, 0, 0, 0, 0, 0 \cdots 0, 0, 0, 0) \tag{5}$$

The circular smooth label of the angle is as follows:

$$label_{csl} = (1, 0.86, 0.71, 0.57, 0.43, 0.29, 0.14, 0 \cdots 0, 0, 0.14, 0.29, 0.43, 0.57, 0.71, 0.86) \tag{6}$$

The detector has two prediction results. In the traditional method, $softmax$ is used to calculate the probabilities of different classes. The corresponding labels are as follows:

$$\begin{cases} label_1 = (0.03, 0.4, 0.03, 0.03, 0.03, 0.03, 0.03, 0.03 \cdots 0.03, 0.03, 0.03) \\ label_2 = (0.03, 0.03, 0.03, 0.03, 0.03, 0.03, 0.4, 0.03 \cdots 0.03, 0.03, 0.03) \end{cases} \tag{7}$$

In the proposed method, $sigmoid$ is used to calculate the probabilities of different classes. The corresponding labels are as follows:

$$\begin{cases} label_1^* = (0.1, 0.8, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1 \cdots 0.1, 0.1, 0.1, 0.1, 0.1) \\ label_2^* = (0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.8, 0.1, 0.1 \cdots 0.1, 0.1, 0.1, 0.1, 0.1) \end{cases} \tag{8}$$

The predicted angle corresponding to $label_1$ and $label_1^*$ is -89°, and the predicted angle corresponding to $label_2$ and $label_2^*$ is -84°. Taking the cross-entropy loss function as an example. In the traditional method, the losses of $label_1$ and $label_2$ to the real label are as follows:

$$\begin{cases} loss_1 = -(1 \times log(0.03) + 0 \times log(0.4) + 0 \times log(0.03) + \cdots) = -log(0.03) \\ loss_2 = -(1 \times log(0.03) + \cdots + 0 \times log(0.4) + 0 \times log(0.03) + \cdots) = -log(0.03) \end{cases} \tag{9}$$

It is found that $loss_1 = loss_2$, that is, the $label_1$ and $label_2$ have the same loss to the ground truth. However, the predicted angle obtained by $label_1$ is -89°, which is only 1° different from the true angle. While the predicted angle obtained by $label_2$ is -84°, which is 6° different from the true angle. The first prediction result is obviously more accurate. The analysis shows that directly dividing the angle into 180 categories will lead to low fault tolerance of adjacent angles. In the proposed method, the losses of $label_1$ and $label_2$ to the real label are as follows:

$$\begin{cases} loss_1^* = -(1 \times log(0.1) + 0.86 \times log(0.8) + 0.71 \times log(0.1) + \cdots) \approx 14.33 \\ loss_2^* = -(1 \times log(0.1) + 0.86 \times log(0.1) + 0.71 \times log(0.8) + \cdots) \approx 16.12 \end{cases} \tag{10}$$

It can be found that $loss_1^* < loss_2^*$, that is, the circular smooth label makes the losses of more accurate labels smaller and increase the error tolerance to adjacent angles.

3.4. Loss function

The total loss function is as equation 11:

$$L = \frac{1}{N} \sum_{n=1}^{N} t'_n \sum_{j \in \{x,y,w,h\}} L_{reg}(v'_{nj}, v_{nj}) + \frac{\lambda_1}{N} \sum_{n=1}^{N} L_{cls}(p_n, t_n) + \frac{\lambda_2}{N} \sum_{n=1}^{N} L_{cls_\theta}(\theta'_n, \theta_n) \tag{11}$$

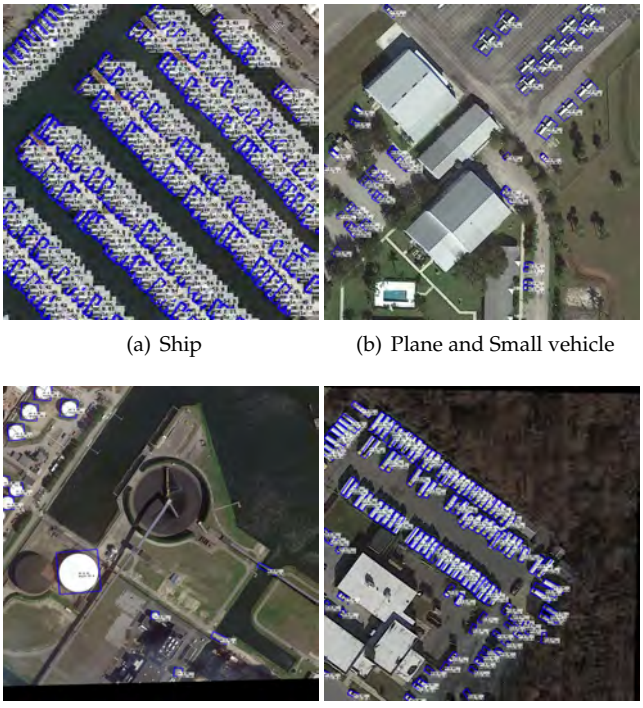
where N indicates the number of anchors, t'_n has two values, i.e., 0 and 1, respectively ($t'_n = 1$ for foreground and $t'_n = 0$ for background). v'_{ij} indicates the predicted offset vector.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 June 2022 doi:10.20944/preprints202206.0390.v1

And v_{ij} indicates the real offset vector. t_n indicates the label of object, p_n indicates the probability distribution of various classes calculated by sigmoid function. Hyperparameter λ_1 and λ_2 are trade-off factors, which control the weights of different loss functions, and their default values are both 1. L_{reg} indicates Smooth L_1 Loss [23]. L_{cls} represents the loss of classification in the object category prediction. L_{cls_θ} represents the loss of angle classification in the angle prediction. Both L_{cls} and L_{cls_θ} use Focal loss [18].

4. Experimental results and discussion

The GPU used in this paper is GTX1660Ti with 6G memory. The operating system we used is Ubuntu 16.04. The deep learning framework is Tensorflow. ResNet50 is used as the backbone of the network. Experiments are carried out on DOTA and the self-made dataset DOTA-GF. Some visual experiment results are shown in Figure 8.



(a) Ship (b) Plane and Small vehicle

(c) Bridge and Storage tank (d) Small vehicle and Large vehicle

Figure 8. Visual detection results of some typical objects

4.1. Ablation studies

In this section, we conduct detailed ablation on DOTA to evaluate the effctiveness of each module and illustrate the advance and generalization of the proposed method.

4.1.1. Bidirectional multi-Scale feature fusion network

To verify the effectiveness of the improved feature fusion network, using ResNet50 as the backbone and RetinaNet as the embodiment, to compare the detection result of the original FPN and the improved feature pyramid network (Improved-FPN) on the DOTA [15] dataset. We mainly consider the average precision (AP) and mean average precision (mAP) of six types of typical objects, including plane (PL), ship (SH), bridge (BG), small vehicle (SV), large vehicle (LV), and storage tank (ST). The experimental results are shown in Table 1.

It can be seen from Table 1 that the Improved-FPN can significantly improve the detection accuracy of typical objects in remote sensing images. Among them, the AP of the ship has the highest increase of 2.4%. That is because many ships in DOTA is small, and

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 June 2022 doi:10.20944/preprints202206.0390.v1

the shallow features have a greater impact on the detection results and the bidirectional multi-scale feature fusion network can make full use of the shallow features. The AP of the storage tank has the least increase, which is 0.6%. The mAP of 6 types of objects is increased by 1.4%. Experimental results show that the improved feature fusion network is more suitable for remote sensing image object detection than the original feature fusion network.

Table 1. The experimental results of the bidirectional multi-scale feature fusion network

Method	AP(%)						mAP(%)
	PL	SH	BG	SV	LV	ST	
FPN	83.4	62.2	32.3	65.7	48.3	74.9	61.1
our-FPN	84.5	64.6	34.0	67.2	49.2	75.5	62.5

4.1.2. Multi-Feature selection module based on attention mechanism

To further prove the effectiveness of the multi-feature selection module, the multi-feature selection module is added to RetinaNet [18] to conduct experiments on DOTA [22]. The comparative experiments of the MFSM with other attention mechanisms are supplemented too. The experimental results with MFSM, SE[28] and CBAM[26] are shown in Table 2.

Table 2. Experimental results of different attention mechanisms

Mthod	AP(%)						mAP(%)
	PL	SH	BG	SV	LV	ST	
Baseline	83.4	62.2	32.3	65.7	48.3	74.9	61.1
SE	83.6	64.3	33.4	66.1	50.1	74.1	61.9
CBAM	84.4	64.5	33.7	67.0	49.1	75.2	62.3
MFSM	84.7	63.4	33.6	67.3	49.5	76.1	62.4

Compared with RetinaNet [18], after adding the multi-feature selection module, the detection accuracy of the six types of typical objects has been significantly improved with an AP increase of 1.2% to 1.6%. The mAP has increased by 1.3%. Among them, the detection accuracy of small vehicle has the greatest improvement, and the AP increases by 1.6%. At the same time, MFSM has better detection performance than SE and CBAM. In SE and CBAM, an attention module is used to process the feature map, and the classification and regression subnets share the feature map. MFSM processes feature maps for classification and regression respectively, which can alleviate the conflict between classification tasks and regression tasks to a certain extent. Therefore, MFSM has a simpler structure, but has better performance.

4.1.3. Accurate acquisition of target direction based on angle classification

To further prove that turning the angle regression problem into a classification task can improve the remote sensing images detection effect, the angle prediction in RetinaNet is regarded a classification task with 180 categories, and CSL is used for smoothing. Comparative experiments are performed on the DOTA, and the experimental results are shown in Table 3. It can be seen from Table 3 that treating the angle prediction as a classification task can significantly improve the detection effect. Among the six types of typical targets, the AP of ships, bridges, small vehicles, and large vehicles increased by 2.7%, 2.2%, 1.9%, and 3.2% respectively. This is because the aspect ratios of these four types of objects are relatively

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 June 2022 doi:10.20944/preprints202206.0390.v1

large, and the use of regression to predict angles has more serious loss discontinuity. For planes and storage tanks with an aspect ratio close to 1, the AP also increased by 0.8% and 0.9%. The experimental results prove that treating the angle prediction as a classification task can effectively improve the detection accuracy of objects with larger aspect ratios.

Table 3. Experimental results of RetinaNet using classification and regression methods to predict angles

Mthod	AP(%)						mAP(%)
	PL	SH	BG	SV	LV	ST	
Regression	83.4	62.2	32.3	65.7	48.3	74.9	61.1
Classification	84.2	64.9	34.5	67.6	51.5	75.8	63.1

4.2. Results on DOTA

The DOTA [15] dataset contains 15 categories. This paper mainly analyzes six typical objects, including ships, planes, bridges, small vehicles, large vehicles, and storage tanks. The evaluation indicators used are AP and mAP. CSL [27], RRPN [3], RetinaNet [18] and Xiao [9] were selected as comparative algorithms. The comparison results of different algorithms are shown in Table 4.

Table 4. Comparison results of different algorithms on the DOTA dataset

Category	CSL	RRPN	RetinaNet	Xiao	Proposed
PL	84.2	83.9	83.4	78	85.7
SH	64.9	47.2	62.2	65	66.5
BG	34.5	32.3	32.3	38	37.5
LV	51.5	49.7	48.3	59	54.2
SV	67.6	34.7	65.7	37	69.2
ST	75.8	48.8	74.9	50	77.3
mAP(%)	63.1	48.0	61.1	55	65.1

The data in Table 4 shows that mAP of the proposed method is better than most of the mainstream object detection algorithms. The algorithm proposed has achieved the highest AP in four types of objects: planes, ships, small vehicles, and storage tanks. Besides, the APs of large vehicles and bridges are second only to the highest. The large vehicles in the DOTA dataset are often placed very closely, and adjacent objects have occlusion problems. This is also a problem that we will study in the future. These comparison results show that the algorithm proposed in this paper can effectively detect typical objects in remote sensing images.

4.3. Results on DOTA-GF

At present, the remote sensing images in public remote sensing datasets such as DOTA [15] and NWPU VHR-10 [29] are mainly derived from Google Earth, with only a small amount of data derived from domestic data and lack of military objects. Therefore, we collected 188 GF-2 Satellite images and GF-6 Satellite images with a resolution of 1000×1000 to 4000×4000 and labeled them using the four-point method.

138 domestic remote sensing images were added to the training set of DOTA as the DOTA-GF training set. The remaining 50 domestic remote sensing images are added to the DOTA testing set as the DOTA-GF testing set. Then select the data containing six types of

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 June 2022 doi:10.20944/preprints202206.0390.v1

objects: ships, planes, bridges, small vehicles, large vehicles, and storage tanks, and crop them to pieces of size 600 × 600 for training. To illustrate the effectiveness of the proposed algorithm, four representative object detection algorithms, CSL [27], RRPN [3], RetinaNet [18] and R3Det [10] were selected for comparison experiments. The detection results of different algorithms are shown in Table 5.

Table 5. Comparison results of different algorithms on the DOTA-GF dataset

Category	CSL	RRPN	RetinaNet	R3Det	Proposed
PL	83.6	81.7	83.2	85.2	84.6
SH	64.1	46.8	61.0	66.1	66.3
BG	35.3	34.8	32.5	35.5	37.2
LV	50.4	48.2	50.2	61.5	53.8
SV	64.7	33.8	64.5	59.8	68.6
ST	72.9	48.6	72.7	70.5	74.1
mAP(%)	56.5	49.0	60.7	63.1	64.1

It can be seen from Table 5 that compared with the four representative algorithms, the algorithm proposed in this paper has achieved the highest AP in four types of objects: ships, bridges, small vehicles, and storage tanks. The APs of planes and large vehicles are also close to the highest AP of the four types of algorithms. However, the network structure of R3Det is more complex. Both the training time and the testing time of a single image are longer than the proposed algorithm. Compared with the four comparison algorithms, the mAP of the six typical objects of the proposed algorithm is also the highest. The experimental results show that the algorithm proposed in this paper still has certain advantages on the self-made DOTA-GF dataset.

4.4. Results on HRSC 2016

HRSC 2016 [30] contains lots of remote sensing ships with a large aspect ratio, scales and arbitrary orientations. Our method achieves competitive performances on the HRSC2016 dataset. The comparison results are shown in Table 6.

Table 6. comparisons with different methods on the HRSC2016 dataset

Methods	Size	mAP (%)
R2CNN	800×800	73.7
RRPN	800×800	79.1
RetinaNet	800×800	81.7
RoI-Transformer	512×800	86.2
Proposed	800×800	87.1

From Table 6, it can be seen that compared with R2CNN [31], RRPN [3], RetinaNet [18], and RoI-Transformer [32], the algorithm in this paper achieves the best detection results, with a mAP of 87.1%. The experimental results verify the effectiveness of the proposed algorithm on HRSC 2016 dataset.

5. Conclusion

Aiming at the challenges such as multi-scale objects, complex backgrounds, and boundary problems, we propose a new remote sensing image object detection algorithm. In this algorithm, a bidirectional multi-scale feature fusion network is designed to combine

the semantic features and shallow detailed features to reduce the loss of information in the process of transferring shallow features to the top layer. A multi-feature selection module based on the attention mechanism is designed to make the network focus on valuable information and assist to select the feature maps appropriate for classification and regression tasks. To avoid boundary discontinuities problem in the regression process, we treat angle prediction as a classification task rather than a regression task. Finally, experimental results on the DOTA dataset, the DOTA-GF dataset and the HRSC 2016 dataset show that the proposed algorithm has certain advantages in remote sensing image object detection. However, our proposed method still has limitations in detecting dense objects. In the future, we will outlook the situation of dense object occlusion, and improve our network model to better detect dense objects. The results reported in this paper can be downloaded from the URL https://github.com/xiaojs18/ObjectDetection/tree/main/Remote-Detection.

References

1. Li, K.; Wan, G.; Cheng, G.; Meng, L.; Han, J. Object detection in optical remote sensing images: A survey and a new benchmark. *ISPRS Journal of Photogrammetry and Remote Sensing* **2020**, *159*, 296–307.
2. Fatima, S.A.; Kumar, A.; Pratap, Raoof, S.S. Object Recognition and Detection in Remote Sensing Images: A Comparative Study. In Proceedings of the 2020 International Conference on Artificial Intelligence and Signal Processing, AISP 2020, 2020.
3. Ma, J.; Shao, W.; Ye, H.; Wang, L.; Wang, H.; Zheng, Y.; Xue, X. Arbitrary-oriented scene text detection via rotation proposals. *IEEE Transactions on Multimedia* **2018**, *20*, 3111–3122.
4. Liu, X.; Meng, G.; Pan, C.a. Scene text detection and recognition with advances in deep learning: a survey. In Proceedings of the International Journal on Document Analysis and Recognition, 2019, pp. 143–162.
5. Lin, T.Y.; Dollar, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Image-to-image translation with conditional adversarial networks. In Proceedings of the Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2117–2125.
6. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path Aggregation Network for Instance Segmentation. In Proceedings of the Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2018, pp. 8759–8768.
7. Guo, C.; Fan, B.; Zhang, Q.; Xiang, S.; Pan, C. AUGFPN: Improving multi-scale feature learning for object detection. In Proceedings of the Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2020, pp. 12592–12601.
8. Ghiasi, G.; Lin, T.Y.; Le, Q.V. NAS-FPN: Learning scalable feature pyramid architecture for object detection. In Proceedings of the Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2019, pp. 7029–7038.
9. Xiao, J.; Zhang, S.; Dai, Y.; Jiang, Z.; Yi, B.; Xu, C. Multiclass Object Detection in UAV Images Based on Rotation Region Network. *IEEE Journal on Miniaturization for Air and Space Systems* **2020**, *1*, 188–196.
10. Yang, X.; Liu, Q.; Yan, J.; Li, A. R3Det: Refined Single-Stage Detector with Feature Refinement for Rotating Object. In Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence, 2021.
11. 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 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 8231–8240.
12. Zhang, Y.; Xiao, J.; Jinye, P.; Ding, Y.; Liu, J.; Guo, Z.; Xiaopeng, Z. Kernel Wiener Filtering Model with Low-Rank Approximation for Image Denoising. *Information Sciences* **2018**, *462*, 402–416.
13. Li, Q.; Mou, L.M.; Jiang, K.; Liu, Q.; Wang, Y.; Zhu, X. Hierarchical Region Based Convolution Neural Network for Multi-scale Object Detection in Remote Sensing Images. In Proceedings of the In Proceedings of 2018 IEEE International Geoscience and Remote Sensing Symposium, 2018, pp. 4355–4358.
14. Xie, H.; Wang, T.; Qiao, M.; Zhang, M.; Shan, G.; Snoussi, H. Robust object detection for tiny and dense targets in VHR aerial images. In Proceedings of the Proceedings - 2017 Chinese Automation Congress, 2017, pp. 6397–6401.
15. Xia, G.S.; Bai, X.; Ding, J.; Zhu, Z.; Belongie, S.; Luo, J.; Datcu, M.; Pelillo, M.; Zhang, L. DOTA: A Large-scale Dataset for Object Detection in Aerial Images. In Proceedings of the Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2018, pp. 3974–3983.
16. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2017**, *39*, 1137–1149.
17. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A. SSD: Single shot multibox detector. In Proceedings of the Proceedings of the 14th European Conference on Computer Vision, 2016, pp. 21–37.
18. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollar, P. Focal Loss for Dense Object Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2020**, *42*, 318–327.

19. 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 Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 8231–8240.
20. Cheng, G.; Wang, J.; Li, K.; Xie, X.; Lang, C.; Yao, Y.; Han, J. Anchor-free Oriented Proposal Generator for Object Detection, 2021, [arXiv:cs.CV/2110.01931].
21. Xie, X.; Cheng, G.; Wang, J.; Yao, X.; Han, J. Oriented R-CNN for Object Detection. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 3520–3529.
22. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
23. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2017**, *39*, 1137–1149.
24. 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 Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 6450–6458.
25. Zhong, Z.; Lin, Z.Q.; Bidart, R.; Hu, X.; Daya, I.B.; Li, Z.; Zheng, W.S.; Li, J.; Wong, A. Squeeze-and-attention networks for semantic segmentation. In Proceedings of the Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2020, pp. 13062–13071.
26. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. CBAM: Convolutional block attention module. In Proceedings of the Computer Vision - ECCV 2018 - 15th European Conference, 2018, pp. 3–19.
27. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollar, P. Arbitrary-Oriented Object Detection with Circular Smooth Label. *Yang, Xue and Yan, Junchi* **2020**, *12353*, 677–694.
28. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7132–7141.
29. Cheng, G.; Han, J. A survey on object detection in optical remote sensing images. *ISPRS Journal of Photogrammetry and Remote Sensing* **2016**, *117*, 11–28.
30. Cheng, G.; Han, J.; Zhou, P.; Guo, L. Multi-class geospatial object detection and geographic image classification based on collection of part detectors. *ISPRS Journal of Photogrammetry and Remote Sensing* **2014**, *98*, 119–132.
31. Pang, J.; Li, C.; Shi, J.; Xu, Z.; Feng, H. R2-CNN: Fast Tiny Object Detection in Large-Scale Remote Sensing Images. *IEEE Transactions on Geoscience and Remote Sensing* **2019**, *57*, 5512 – 5524.
32. Ding, J.; Xue, N.; Long, Y.; Xia, G.S.; Lu, Q. Learning roi transformer for oriented object detection in aerial images. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 2849–2858.