International Journal of Geo-Information

MDPI

Article

Vehicle Detection in Very-High-Resolution Remote Sensing Images Based on an Anchor-Free Detection Model with a More Precise Foveal Area

Xungen Li [1,2], Feifei Men [1], Shuaishuai Lv [1,2,*], Xiao Jiang [3], Mian Pan [1], Qi Ma [1] and Haibin Yu [1]

[1] School of Electronics and Information Engineering, Hangzhou Dianzi University, Hangzhou 310018, China; lixg@hdu.edu.cn (X.L.); menff@hdu.edu.cn (F.M.); ai@hdu.edu.cn (M.P.); maqi@hdu.edu.cn (Q.M.); shoreyhb@hdu.edu.cn (H.Y.)
[2] Pujiang Microelectronics and Intelligent Manufacturing Research Institute, Hangzhou Dianzi University, Jinhua 322200, China
[3] School of Mechanical Engineering, Hangzhou Dianzi University, Hangzhou 310018, China; jx@hdu.edu.cn
* Correspondence: lvshuai@hdu.edu.cn

Abstract: Vehicle detection in aerial images is a challenging task. The complexity of the background information and the redundancy of the detection area are the main obstacles that limit the successful operation of vehicle detection based on anchors in very-high-resolution (VHR) remote sensing images. In this paper, an anchor-free target detection method is proposed to solve the problems above. First, a multi-attention feature pyramid network (MA-FPN) was designed to address the influence of noise and background information on vehicle target detection by fusing attention information in the feature pyramid network (FPN) structure. Second, a more precise foveal area (MPFA) is proposed to provide better ground truth for the anchor-free method by determining a more accurate positive sample selection area. The proposed anchor-free model with MA-FPN and MPFA can predict vehicles accurately and quickly in VHR remote sensing images through direct regression and predict the pixels in the feature map. A detailed evaluation based on remote sensing image (RSI) and vehicle detection in aerial imagery (VEDAI) data sets for vehicle detection shows that our detection method performs well, the network is simple, and the detection is fast.

Keywords: vehicle detection; remote sensing image; convolutional neural network; anchor-free

Citation: Li, X.; Men, F.; Lv, S.; Jiang, X.; Pan, M.; Ma, Q.; Yu, H. Vehicle Detection in Very-High-Resolution Remote Sensing Images Based on an Anchor-Free Detection Model with a More Precise Foveal Area. *ISPRS Int. J. Geo-Inf.* **2021**, *10*, 549. https://doi.org/10.3390/ijgi10080549

Academic Editors: Wolfgang Kainz and Sébastien Lefèvre

Received: 18 March 2021
Accepted: 10 August 2021
Published: 14 August 2021

1. Introduction

Using computer technology to complete tasks such as the classification [1], detection [2], and segmentation of remote sensing images has always been a hot topic in the field of image research [3,4]. Among these tasks, vehicle detection in remote sensing images plays an important role in urban vehicle supervision [5–7], defense, traffic planning, safety-assisted driving, etc. [8–10]. For such tasks, accurate and fast detection methods are essential. Compared with other detection targets in remote sensing images (such as buildings [11], ships [12], and airplanes [13]), the task of vehicle detection is more difficult because there is complex background interference information, an uneven distribution of vehicle targets, and fewer target pixels. The detection of vehicle targets in remote sensing images has been developed for many years and has achieved promising research results [14–16].

Remote sensing image vehicle detection aims to detect all instances of vehicles in remote sensing images. In early methods, researchers usually directly designed and extracted vehicle features manually and then classified them to achieve vehicle detection. The key idea is to extract vehicle features and use classical machine learning methods for classification. Commonly, the features used for recognition include scale-invariant feature transform (SIFT) features [17], integration channel features [18], and histogram

of oriented gradient (HOG) features [19], etc. The methods used for classification are SVM [20], AdaBoost [21], intersection kernel support vectors (IKSVM) [22], etc.

However, traditional target detection methods pay more attention to the completion of remote sensing image vehicle detection tasks, and it is difficult to balance accuracy and speed. Compared with the rapid development of deep learning technology, there is a big gap in the effectiveness and accuracy of detection. Network models based on deep learning methods can extract richer features and map complex non-linear relationships. Thanks to the improvement of hardware technology and massive data, two types of target detection network model have been continuously formed and optimized: two-stage networks (such as faster RCNN [23] and cascade RCNN [24]) and single-stage networks (such as YOLOv3 [25] and SSD [26]).

For vehicle detection in remote sensing images, scholars have proposed various improvements based on general detection models and achieved good results. The study [27] presented an accurate vehicle proposal network (AVPN) based on a hyper feature map. The network first combines multi-layer features to detect small target vehicles in remote sensing images more accurately. Then, a coupled R-CNN method is constructed on this basis, which combines AVPN and a vehicle attribute learning network and extracts the vehicle's location and attributes at the same time to obtain the final detection result. Lichao et al. [28] used the principle of residual learning (ResNet) to propose a unified multi-task learning network that can perform vehicle region segmentation and semantic boundary detection at the same time. In order to eliminate the influence of the background information in the horizontally labeled target box on the detection results, some researchers have tried to introduce a rotating rectangular box in text detection. Researchers began to introduce the rotatable target box in the text inspection field into the current task to solve this problem. For example, Li et al. [29] integrated the rotatable region proposal network into the real-valued flexibly connected neural network (RFCN) [30] to add angle information to the horizontal rectangular box in the RFCN and then complete the image and video detection of vehicles in any direction. The studies [31,32] investigated the rotatable rectangular box proposal by introducing angle information into the anchor definition and achieving the target detection of vehicles in any direction. Liu et al. [33] proposed a novel algorithm to generate orientation proposals that can correctly encapsulate vehicle objects as rotating rectangles with direction.

Current methods focus on adopting anchor-based methods. However, the use of anchors has led to a variety of shortcomings. (a) The anchor requires the artificial setting of a large number of parameters, which increases the computational burden of the network and reduces the detection speed. (b) The discrete anchor scale setting will cause some vehicle targets to fail to match the anchor, resulting in missing targets. (c) The pre-defined anchor shape may not meet the requirements of the target in the data set. In order to ensure the recall rate, too many negative samples are introduced, which will limit the detection accuracy.

Apart from the above-mentioned defects of the anchor, vehicle detection based on remote sensing images still has the following difficulties compared with general target detection: (a) the vehicle targets in remote sensing images usually occupy only a small area (as shown in Figure 1a, the length and width of the vehicle in the VEDAI data set are concentrated at 5–40 pixels—i.e., the upper left corner of the figure), which will cause a serious loss of spatial information in the transmission of the image in the deep neural network and affect the detection results; (b) since the vehicle targets are naturally rectangular and have monotone colors in remote sensing images, vehicle-like targets are common in complex scenes, such as those including trees, lawns, roofs, etc., as shown in Figure 1b, which will cause false detections; (c) the direction of the vehicle target is arbitrary in the remote sensing image, which will result in a large number of redundant areas in the horizontally labeled target and introduce too many negative samples for the training of the anchor-free model, as shown in Figure 1c.

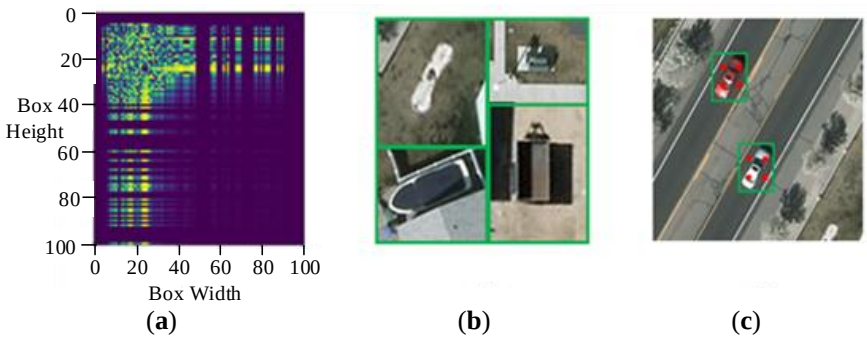


Figure 1. Picture information in the VEDAI data set: (**a**) VEDAI image Scale distribution, (**b**) similar vehicles, (**c**) FCOS ground truth.

In this paper, to solve the problems mentioned above, we propose an anchor-free network that supports end-to-end training to complete the rapid and accurate detection of vehicle targets in remote sensing images. The main contributions of this article are as follows:

(1) The multi-attention feature pyramid network (MA-FPN) module is designed to filter remote sensing images and noise information. This method has incorporated a variety of attendance information into the feature pyramid network (FPN) [34], including channel attention (CA) and spatial attention (SA) mechanisms. At the same time, it can also increase the feature representation of the vehicle.

(2) The more precise foveal area (MPFA) method is proposed to avoid redundant information in the target box in vehicle detection. We propose this new method to distinguish between positive and negative sample pixels. By determining the ground truth with a uniform distribution and reasonable sampling, high-quality training samples with a proper distribution that will eliminate redundant information can be provided for network training.

(3) We use a single-stage anchor-free detection model to detect vehicle targets in remote sensing images. Our method eliminates the negative impact of anchor detection and completes the end-to-end design of vehicle detection in remote sensing images compared with the existing anchor model. Moreover, our model has the advantages of precision and speed. Thus, it is easier to apply in practical engineering applications, such as in small multi-axis aircraft with limited computing ability.

The rest of this study is organized as follows. Related work is discussed in Section 2. In Section 3, the details of the proposed method are presented. In Section 4, experiments on VHR remote sensing image data sets are carried out to verify the proposed framework's effectiveness. The results and conclusion are provided in Section 5.

2. Related Work

First, we briefly introduce anchor-free vehicle detection models and attention mechanisms for enhancing the feature extraction of vehicle targets in remote sensing images.

2.1. Anchor-Free Models

As shown in Figure 2, RetinaNet [35] provides a good network structure and focal loss. On the basis of RetinaNet, two anchor-free target detection networks, fully convolutional one-stage object detection (FCOS) [36] and FoveaBox [37], were derived using the FCN [38] network to predict the classification and position of the pixels in the feature map to replace the anchor of RetinaNet. By predicting the dense points on the input images, FCOS and FoveaBox eliminate the setting of anchor hyperparameters, reduce the computational burden of the network, and increase the strength of the network's fitting of vehicle targets. The core difference between FCOS and FoveaBox is that the processing methods for the low-quality pixels are different in the input ground truth minimum enclosed rectangle. FoveaBox limits the network artificially to learn the pixels that deviate from the center of

the object in the ground truth minimum enclosed rectangle, while FCOS filters out the low-quality edge detection box automatically by adding a sub-network.

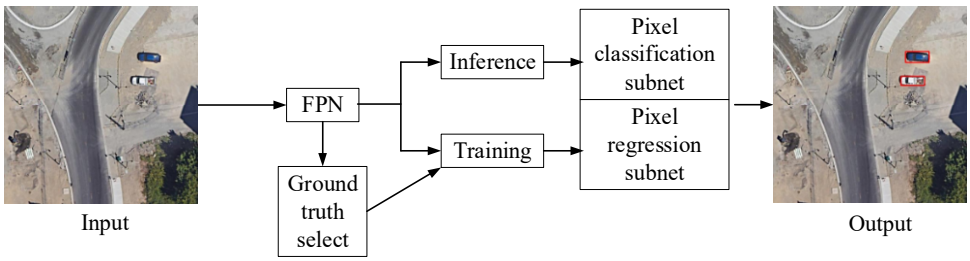


Figure 2. RetinaNet network structure.

In the training process, FCOS uses each position in the label box as a training sample (as shown in Figure 3a). To solve the problem of the low-quality prediction boxes generated at positions far from the target center, Lin et al. added a branch, called "centerness", which is defined as follows:

$$centerness^* = \sqrt{\frac{\min(l^*, r^*)}{\max(l^*, r^*)} \times \frac{\min(t^*, b^*)}{\max(t^*, b^*)}}, \tag{1}$$

where t^*, b^*, l^*, and r^* represent the distance of the current pixel from the top, bottom, left, and right of the box.

(a) **(b)**

Figure 3. Sampling method. (**a**) FCOS sampling, (**b**) FoveaBox sampling.

In the training processes of FoveaBox, the impact of pixels far away from the target on the detection result is considered. In the selection of the training samples, the parameter σ_1 is used to shrink the real box position to the center of the target, while the points in the shrinking area are used for training. (x_1, y_1, x_2, y_2) represent the upper left and lower right corners of the ground truth of the target. The positive sample area $(x_1'', y_1'', x_2'', y_2'')$ is determined as follows:

$$\begin{cases} x_1'' = c_x' - 0.5(x_2' - x_1')\sigma_1 \\ x_2'' = c_x' + 0.5(x_2' - x_1')\sigma_1 \\ y_1'' = c_y' - 0.5(y_2' - y_1')\sigma_1 \\ y_2'' = c_y' + 0.5(y_2' - y_1')\sigma_1 \end{cases}, \tag{2}$$

where (x_1', y_1', x_2', y_2') is the mapping of the ground truth of the object on the corresponding feature map, and (c_x', c_y') are the center coordinates of the rectangular box represented by (x_1', y_1', x_2', y_2'). In this paper, we chose a novel positive sample selection method to avoid the shortcomings caused by the direct use of FCOS and FoveaBox in remote sensing image vehicle target detection.

2.2. Attention Mechanism

The attention mechanism is widely used in computer vision to extract the main information and filter out interference information [39–42]. It can make the network sensitive to the region of interest by assigning different weights to different features at the same level. Park et al. [41] inferred an attention map along two separate pathways, channel

and spatial, and proposed a bottleneck attention module (BAM). The proposed method could be trainable in an end-to-end manner jointly with any feed-forward models. Woo and Park et al. [42] also proposed a convolutional block attention module (CBAM). The proposed CBAM is a lightweight and general module, and it is also end-to-end trainable along with base CNNs.

Although a deeper network can obtain better semantic information about vehicle targets and is more conducive to the determination of vehicle categories when extracting vehicle targets in remote sensing images, the loss of vehicle target spatial information is serious, and is not conducive to vehicle bounding box regression. Due to the special perspective of the remote sensing image, the vehicle detection task is disturbed by factors such as complex background information, inconsistent scale distribution, and noise. As shown in Figure 4, after the input image is resized to 1333 × 800, the scale decreases exponentially, which means that the learnable offset during pixel regression will be very small. The underlying feature map is not conducive to the classification of pixels due to the low amount of semantic information.

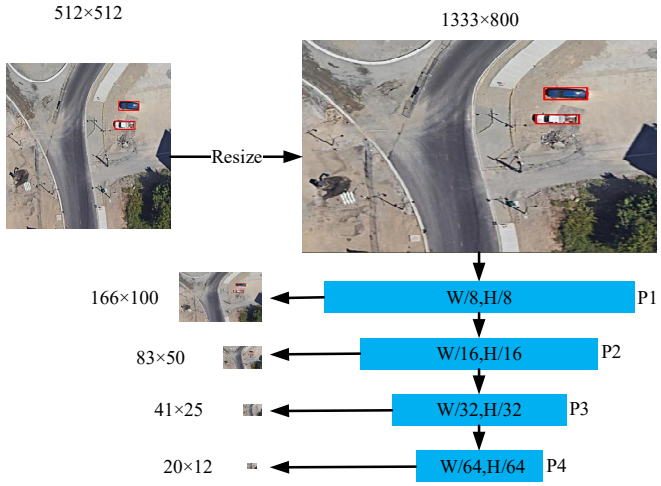


Figure 4. Scale change of feature pyramid network (FPN) after image resize.

The attention mechanism makes the network sensitive to the region of interest in the task by assigning different weights to different features at the same level. Through the channel attention mechanism, different channels in the feature map are given different weights. After training and screening, it can be ensured that the underlying feature map can still successfully represent the semantic information of the vehicle. Through the spatial attention information, the value weight of the non-vehicle target area becomes lower. In the subsequent processing, these uninteresting areas will be filtered out, thereby reducing background information and noise interference in the remote sensing image.

3. Methods

We propose a VHR remote sensing image detection method based on the anchor-free detection model. The model consists of three parts, as shown in Figure 5. (1) MA-FPN: FPN structure that integrates multiple attention mechanisms; (2) MPFA: a more accurate method for calculating pit regions; (3) prediction subnet: classification and minimum enclosed rectangle regression subnets are used to obtain the classification information of the pixels in the feature map and the offset of the current position pixels from four directions—i.e., up, down, left, and right.

The process of vehicle detection can be described as follows. First, the image is inputted into MA-FPN to generate multiple feature maps with different scales containing attention information. Second, the vehicle instance information in the rectangular box in the image will also be inputted during network training. Combined with the corresponding feature map scale, to determine a more accurate foveal area, the pixels in this area will

be regarded as positive samples. Finally, the feature map obtained by MA-FPN will be inputted first into the classification branch and then into the box regression branch to determine the category and regression information of the pixels in the feature map.

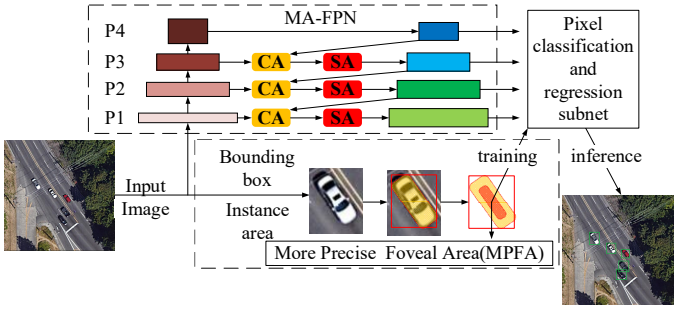


Figure 5. Schematic diagram of the overall network.

3.1. Feature Pyramid That Combines Multiple Attention Information (MA-FPN)

After the image is input into the neural network layer through operations such as pooling, the low-level feature map will slowly become a high-level feature map with a reduced scale and increased channel count. The proposed FPN method proves the effectiveness of multi-level feature map information fusion. We chose to add this mechanism to the feature fusion process from the top to the bottom of the feature pyramid. The specific fusion method is as follows.

The neural network performs feature extraction on the input image to obtain feature maps of different scales. First, channel attention information is extracted from the feature maps of two adjacent feature maps with a small scale, and they are added to adjacent feature maps with larger scales. Then, spatial attention information is extracted from the obtained feature map and the result we need is obtained.

In this paper, we use ResNet [43] as the backbone of the feature extraction and select the last layer of each residual block to form a bottom-up feedforward network. After the image passes through the network, a multiple feature map with a decreasing scale and increasing channel number will be generated. For an input image with a size of $W \times H \times 1$, the scale change of the output feature map {p1, p2, p3, p4} is $\{W/8 \times H/8, W/16 \times H/16, W/32 \times H/32, W/64 \times H/64\}$ and the number of channels changes to {256, 512, 1024, 2048}. After obtaining multiple feature maps, we perform a convolution operation on each feature map to unify the channel numbers to facilitate subsequent attention fusion between the feature maps. The channel unification formula can be expressed as:

$$P_i = \mathrm{Conv}_{3\times3}(C_i, 256, 1, 1) \tag{3}$$

where Conv(·) represents the convolution operation in the neural network, the 3×3 subscript of the convolution operation is the size of the convolution kernel used in the convolution operation, C_i is the number of input feature map channels, 256 means the output feature map channel number, and the last two parameters in the brackets represent stride and padding, respectively.

Then, in the process of horizontal linking and the top-down transmission of the feature map in the upper left area, the channel attention operation and the spatial attention operation are successively carried out. Each feature map fusion always has two adjacent high-level features and low-level features participating in the procedure. As shown in Figure 6, maximum channel pooling and average pooling are performed on low-level features. Then, the sigmoid function (Equation (4)) is used to activate the merged block to obtain the feature map B_c with rich channel attention information. Finally, a 1×1 convolution is input to reduce the dimensionality to obtain a block with 256 channels and a scale of 1.

This block will be multiplied by a low-level feature map of a channel to obtain a low-level feature map containing channel attention.

$$s(z) = \frac{1}{1 + e^{-z}} \tag{4}$$

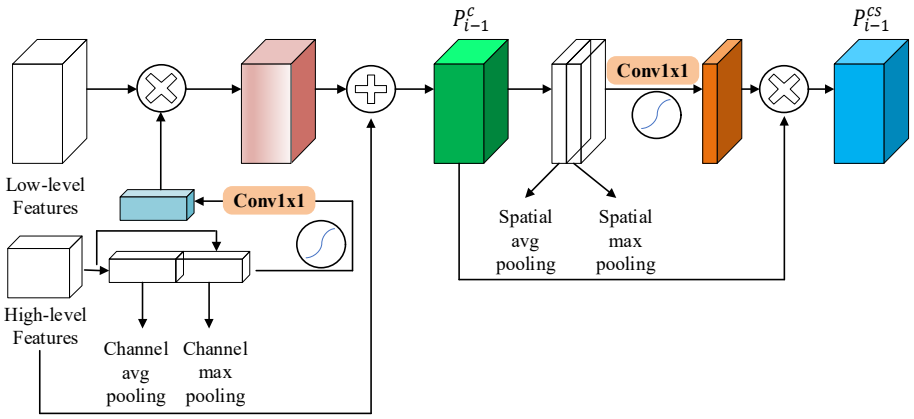


Figure 6. Schematic diagram of the attention mechanism fusion.

$$B_c = \text{sigmoid}\left(\text{cat}\left(C_{\text{maxpool}}(P_i), C_{\text{avgpool}}(P_i)\right)\right) \tag{5}$$

Finally, through the residual connection, the feature map containing the channel attention and the high-level feature map after double-down sampling are added to obtain P_{i-1}^{c}. The process can be expressed in the following form:

$$P_{i-1}^{c} = \text{Conv}_{1\times1}(B_c) \times P_{i-1} + \text{Unsample}(P_i) \tag{6}$$

where P_{i-1}^{c} represents a feature map incorporating channel attention information c, P_{i-1} is the next layer network of P_i, and $\text{cat}(\cdot)$ represents the concatenate operation of the feature map. After merging the two blocks, the sigmoid function is activated. $C_{\text{maxpool}}(\cdot)$ represents the maximum pooling of the feature map in the channel direction, $C_{\text{avgpool}}(\cdot)$ represents the average pooling of the channel, and $\text{Unsample}(\cdot)$ double upsamples the feature map.

$$B_{cs} = \text{sigmoid}\left(\text{cat}\left(S_{\text{maxpool}}\left(P_{i-1}^{c}\right), S_{\text{avgpool}}\left(P_{i-1}^{c}\right)\right)\right) \tag{7}$$

The feature map P_{i-1}^{c} obtained by the channel attention information extraction will be successively subjected to maximum spatial pooling $S_{\text{maxpool}}(\cdot)$ and average spatial pooling $S_{\text{avgpool}}\left(P_{i-1}^{c}\right)$. After concatenating and convolution, the sigmoid will be used to activate the obtained feature map to obtain a feature map B_{cs} with rich channel attention information and spatial attention information. Finally, the result is multiplied by P_{i-1}^{c} again, and the final feature map $P_{\text{i-1}}^{\text{cs}}$ is obtained, which will form a feature pyramid together with other feature maps rich in attention information obtained by the same operation and used for subsequent network classification and regression.

$$P_{i-1}^{cs} = \text{Conv}_{1\times1}(B_{cs}) \times P_{i-1}^{c} \tag{8}$$

Through a series of feature fusion operations, the current structure is dedicated to integrating attention information into the feature information extracted from remote sensing images. The output feature map can express vehicle information more abundantly without significantly increasing the computational complexity of the network, which facilitates the subsequent network classification and the regression of the pixels in the feature map.

3.2. More Precise Foveal Area (MPFA)

This section introduces detected vehicle instance segmentation information to remove redundant information in the positive sample selection stage to make the target semantic information clearer. We also propose a new way to distinguish positive and negative samples by determining a more accurate concave point area in the vehicle target box and using the pixel coordinates as a positive training sample, which provides a new ground truth selection method that is anchor-free. The specific implementation steps of the MPFA method are shown in Table 1 and Figure 7.

Table 1. The operation flow of accurate area point sampling.

The Input Is the Ground Truth Minimum Enclosed Rectangle Scaled by the Corresponding Feature Map and the Instance Segmentation Information of Vehicle 1.
1. For the marked rectangle, the pixel of the point in the vehicle segmentation area is set to 1; otherwise, it is 0.
2. After setting the network hyperparameters, assign the current vehicle target box to a feature map of a specific scale.
3. Zoom the input image to the corresponding ratio of the feature map, and mark the corresponding input vehicle with a rectangular box, and the binary image M will also be zoomed.
4. Combine the segmentation information of the vehicle to determine the binary map M in the minimum enclosed rectangle.
5. Calculate the weight $w_{i,j}$ of all pixels in the binary image M, and $P_{i,j}$ is the value at the current coordinate i,j. $w_{i,j} = \sum\sum P_{i,j}$, where $i \in \{i, i-1, i+1\}, j \in \{j, j-1, j+1\}$.
6. Recording threshold thr $\leq \max(w_{i,j})$.
7. Traverse the binary graph M, the list stores all sampling points, list $\leq w_{i,j} \geq$ thr.

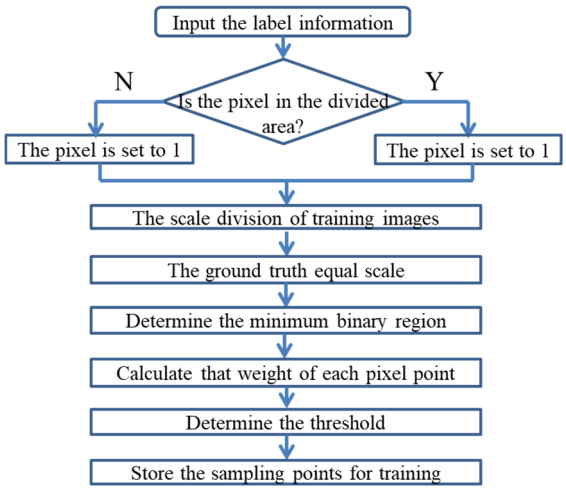


Figure 7. The acquisition process of accurate area point sampling.

In the network training stage, in addition to inputting the horizontal rectangular frame marked in the original remote sensing image, we also select the instance segmentation information for the vehicle in the current marked rectangular box. For the data set designed by us, the points around the vehicle target in the VHR remote sensing image have been marked in the preliminary marking process, and the area composed of these points forms the segmentation information of the vehicle. In the public data set VEDAI, the four corners of the vehicle are provided to form the smallest rectangular box surrounding the vehicle; therefore, it can simply be considered that the smallest rectangular box has the same meaning as the instance segmentation information of the vehicle. In addition, it is very simple to obtain the horizontal rectangular box of the vehicle from the minimum target area surrounding the vehicle target—i.e., the combination of the extreme values of the horizontal and vertical coordinates of the surrounding area points. The supervision

information will be sent to the network together for subsequent learning after data labeling and preprocessing operations. The remote sensing images will undergo a series of random flip, resize pooling, and attention information fusion operations, and finally form a set of feature maps with an increasing scale and constant channel number—i.e., the top-down feature pyramid in the middle part of Figure 5. The important point in the feature pyramid is that the prediction of the input supervision information is hierarchical—i.e., the feature map that should be used is determined according to the size of the input target box. The larger target selects the higher-level features, and the smaller target selects the lower-level features. This is caused by changes in the semantic information and spatial information of the feature map of the input image in the feature pyramid. In the VHR remote sensing image target detection, because the area of the vehicle targets is relatively clustered, we know that most of the vehicle targets selected the lowest-level feature map during the training and testing process—i.e., the output feature map corresponding to the P1 feature map in Figure 5.

It can be seen from the ResNet scaling of the picture that the lowest-level feature map is eight times larger than the original input image. In the network training stage, the supervision information following the input VHR remote sensing image will also be scaled proportionally. At this time, the point in the input vehicle target box aggregates the 8×8 area information in the input remote sensing image. To show whether the pixels in the rectangular box belong to the segmentation information of the vehicle, we binarize the values of the pixels in the rectangular box, which indicate that 0 in the rectangular box represents the redundant information mentioned above, and 1 is the pixel in the rectangular box that belongs to vehicle information. After completing the processing of the input information, we calculate the new pixel value at each position in the rectangular box on the binary image. These values express the abundance of vehicle information around the current pixel and indicate whether the point should be selected as the weight of supervision information during training. In the two-dimensional binary matrix of the target, the pixel points are traversed from top left to bottom right, the surrounding eight direction values are accumulated for each point, and, finally, the accumulated result is added to its value. For the pixels on the border of the rectangular box, the missing positions are filled with 0.

In the process of accumulating the pixel value, the record of the maximum value of the pixel is updated with the number of traversals as the sampling point selection threshold. When the traversal accumulation is completed, we perform another traversal to select the final trainable pixel points and record all the points in the rectangular frame whose pixel values are not less than the threshold value. At this point, these selected pixels represent the supervision information in the rectangular box of the input VHR remote sensing image vehicle target that can be used for subsequent network learning.

After the above steps, we will completely obtain the vehicle target training points in an input remote sensing image. As can be seen in Figure 8, the pixels obtained by our sampling method completely filter out the influence of redundant information in the target box. The pixels obtained by our sampling method are close to the center of the object and away from the surrounding rectangular box, keeping a certain distance from it, which facilitates the return of subsequent pixels to the surrounding rectangular box.

3.3. The Prediction Network

As shown in Figure 5, the feature map of the obtained multi-attention mechanism will be sent to class subnets and box subnets. Both of these two sub-networks use the FCN structure for the dense prediction of pixels. According to the different tasks of the sub-network, the same feature map head is convolved four times (on the dotted line $\times$ 4 in Figure 9), and the $W \times H \times 1$ and $W \times H \times 4$ outputs are obtained by one dimensionality reduction. The 1 of the class subnets indicates whether it belongs to the vehicle class, and the 4 of the box subnets indicates the offset of the distance between the current pixel and the four sides of the box.

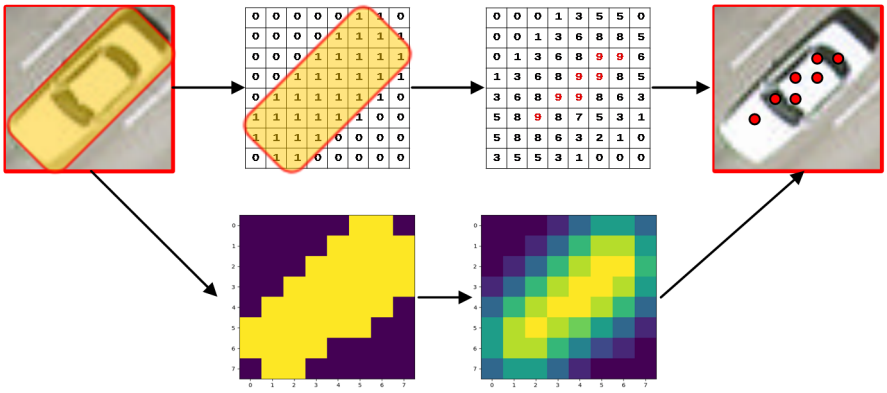


Figure 8. Example of accurate foveal area sampling.

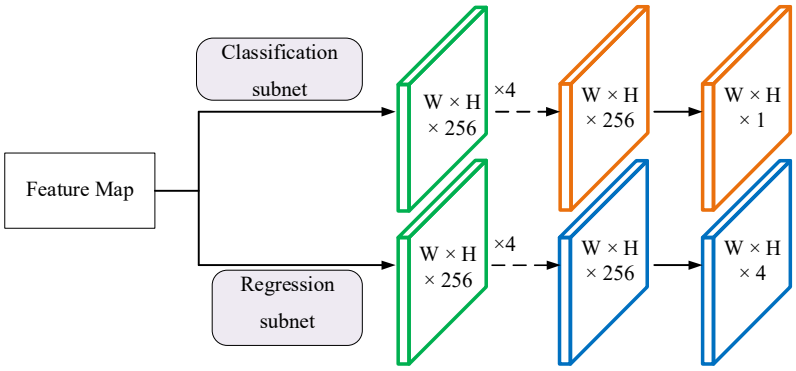


Figure 9. Prediction network. Classification subnet(top), Regression subnet(bottom).

It can be seen from Figure 9 that the prediction network first generates feature maps with sizes W × H × 1 and W × H × 4 with a value of 0. For the determined positive sample pixels, the label of the corresponding position in the W × H × 1 feature map will be set to 1, indicating that the current pixel position contains the vehicle target. At the same time, the offset of the distance between the pixel and the four sides of the ground truth is calculated and assigned to W × H × 4. The mapping relationship is as follows:

$$\begin{cases} t_{x_{\min}} = \log \frac{2^l(x+0.5)-x_{\min}}{z}, t_{y_{\min}} = \log \frac{2^l(y+0.5)-y_{\min}}{z} \\ t_{x_{\max}} = \log \frac{x_{\max}-2^l(x+0.5)}{z}, t_{y_{\max}} = \log \frac{y_{\max}-2^l(y+0.5)}{z} \end{cases}, \tag{9}$$

where x and y represent the coordinates of the positive sample pixels in the feature map; $x_{\min}, y_{\min}, x_{\max}, y_{\max}$ represent the values of ground truth on the feature map scale; l represents the reduction factor of the feature map relative to the input image; z is a normalization factor that can map the output space to space with a center of 1 to make the training stable. Finally, the log space function is used for regularization.

The loss function consists of two parts. The first part, $\boldsymbol{L}_{\text{reg}}$, represents the regression loss function of the pixels in the feature map, and the second part, $\boldsymbol{L}_{\text{cls}}$, is the category loss function of the pixels in the feature map. Using focal loss, it is convenient to reduce the weight of a large number of simple negative samples in training, and it can also be understood as a difficult sample mining method that provides a good solution to the classification problem of imbalance between positive and negative samples. The specific form of the total loss function can be represented as follows:

$$L = \frac{1}{N_{pos}} L_{cls} + \frac{\lambda}{N_{pos}} L_{reg}, \tag{10}$$

where N_{pos} represents all the positive targets in the ground truth, and λ is an adjustable parameter to balance two different loss functions. For the loss function L_{reg}, we use the common smooth L1 loss to minimize the objective function. It is defined as follows:

$$\text{Smooth L}_1 = \begin{cases} 0.5x^2, |x| < 1 \\ |x| - 0.5, x < 1 \text{ or } x > -1 \end{cases}, \quad (11)$$

$$\text{Focal Loss} = -\alpha(1 - \text{P}_\text{c})^\gamma \log(\text{P}_\text{c}), \quad (12)$$

where P_c represents the probability that the classifier predicts the car class, α and γ are adjustable parameters. Of these, α is used to balance the uneven proportion of positive and negative samples. α and γ are determined by the grid search method. Because the negative samples are easy to distinguish and to make the proportion of positive samples smaller than that of negative samples, α is searched in [0,0.5]. Focal loss becomes a cross-entropy loss function when $\gamma = 0$, and the increase in γ will slow down the rate of weight reduction in simple samples, so γ is searched in [0,10]. Throughout the experiments, $\alpha = 0.15$ and $\gamma = 2.5$ are selected. In the training stage, the selected positive samples are used to adjust the parameters of the network. In the prediction stage, the two sub-networks output the category information and regression information of the pixels and form the target box. The final VHR remote sensing image vehicle target detection result is obtained by setting the category confidence threshold and non-maximum suppression (NMS).

4. Experimental Results

In this section, we will first introduce the two data sets used in the experiments. Then, we will describe several sets of comparative experiments to explore the performance of the proposed framework. The ablation experiment will also be demonstrated to show the effectiveness of our proposed method. All the programs in our experiment were completed in PyTorch, developed by Facebook. The operating environment includes an Nvidia RTX 2080Ti graphics card with 12 GB memory and a multi-core (Intel Xeon E5-2667 v3) CPU.

4.1. Data Set and Settings

In this study, we conducted experiments on the data set RSI (designed by us) and the public data set VEDAI, which came from Utah AGRC, USA [43].

RSI includes 50 images from different environments (including cities, deserts, and ports) collected from the public source Google Maps, USA, which has a resolution of 4.92 in. × 4.92 in. per pixel. The image size is 5000× 5000 pixels, and the ground sample distance (GSD) is 12.5 cm/pixel. In order to enrich the vehicle remote sensing image data set, the geographic attributes of the image were considered in the image selection, and different areas, such as city center, city suburbs, desert, port, etc., were selected, respectively. In the RSI data set, the annotation of the vehicle target not only includes ($\boldsymbol{x}_{\min}$, $\boldsymbol{y}_{\min}$, $\boldsymbol{x}_{\max}$, $\boldsymbol{y}_{\max}$) but also includes the instance label of the vehicle in the target box, which consists of a series of point sets. We performed the random segmentation of these large images without overlap, selected images with a resolution between 512 × 512 and 1500 × 1500, and instantiated the vehicle targets in them. We obtained 500 images in total. We randomly selected 250 images for training, and the remaining 250 images were used for testing. At the same time, we used data enhancement operations on the training images, including rotation (90°, 180°, 270°), flip (up and down, left and right, main diagonal flip), and adding Gaussian noise. The final training data set totaled 2000 sheets.

The VEDAI is a data set for vehicle detection in the aerial image, provided as a tool to benchmark automatic target recognition algorithms in unconstrained environments, and collected from Utah, USA. The VEDAI contains color and infrared images. The color images have three 8-bit channels (R, G, B) while the infrared images have only one 8-bit channel. In addition, the VEDAI data set provides two different image resolutions: 512 × 512 and 1024 × 1024. In the VEDAI data set, eight-point coordinates of the vehicle target are provided, which form a minimum rectangular box. We use this smallest rectangular box as

the instantiation label of the vehicle. Compared with other public data sets, the average vehicle targets available for training in each picture in the VEDAI data set are few and small. Therefore, we chose VEDAI512 to verify the robustness of our proposed model in the case of a few targets and small target boxes, and its ground sampling distance was 25 cm/pixel. Additionally, after screening the vehicle targets, there were a total of 1066 different images in the VEDAI512 data set, and each image contained three visible color channels, corresponding to an image of one near-infrared channel.

However, the infrared image in the VEDAI data set can be regarded as a three channel image, and only the values of the three channels are the same. Hence, to ensure the number of the training and testing images, we uniformly processed the infrared image and the color image according to the color image and collected the values of the RGB three channels as training and test data. First, we randomly selected half the image set from the three visible color channels; then, we selected the corresponding near-infrared channel image set. The ratio of the training set to the test set was 1:1.

The backbone of all models involved in the paper was based on the ResNet50 pre-trained on ImageNet [44]. The batch size during training was set to two, and the initial learning rate was 0.01. We used stochastic gradient descent (SGD) [45] for 24 epochs, and weight decay and momentum were set as 0.001 and 0.9. The learning rate was reduced by a factor of 10 at the end of the last 20 K training steps.

4.2. Evaluation Metrics

The output of our model was the category information of the pixel in the feature map and the offset of the pixel in the top, bottom, left, and right directions. First, the detection box with a confidence level of less than 0.05 in the detection result was filtered out. Next, the non-maximum suppression method was used to eliminate some overlapping detection boxes, where the intersection over union (IoU) of NMS was set to 0.5.

To evaluate the performance of different models on the data set, similar to the literature [14], we also used average precision (AP), average recall (AR), precision–recall curve (PRC), F-measure (F1), and frames per second (FPS) as the basis for judging the results.

TP (true positive) means the positive samples are correctly identified as positive samples, representing the number of vehicles correctly detected; FNs (false negatives) are the positive samples predicted as negative samples, representing the quantity of undetected or missed vehicles; FP (false positive) represents the quantity of falsely detected vehicles. The precision rate reveals the proportion of the number of target boxes that are confirmed as vehicle target boxes in the predicted result target box in the remote sensing image. In contrast, the recall rate indicates the proportion of the predicted vehicle target boxes in all real target boxes of the remote sensing image. The larger the ratio, the better the model can be expressed. The definitions of accuracy and recall are as follows:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{13}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{14}$$

The F1 score takes into account both the precision and recall of the model. The F1 score can be regarded as a harmonic average of model precision and recall. It is defined as follows:

$$\text{F1-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{15}$$

4.3. Comparison with Other Target Detection Methods

To verify the effectiveness and comparative advantages of our proposed model, we selected some advanced detection models for comparative experiments, including two-stage models such as faster RCNN, cascade RCNN, and FPN, and single-stage models such as SSD512, RetinaNet, FCOS, and FoveaBox. To further illustrate the effectiveness of

the proposed method in remote sensing image vehicle detection, some results on vehicle detection, including VCSOP detector [46], convolutional neural networks (CNNs) [47], and Oriented_SSD512 (Oriented_SSD300) [31], are directly used to compare the advantages of the proposed method, because it is difficult for us to reproduce the experimental results due to lack of open-source code and data set.

The initial learning rate was 0.001 and maximum training steps was 10 k times; the learning rate for the previous 80,000 times remains unchanged and the learning rate of the latter 20,000 times is decreased to 0.0003, the weight is attenuated to 0.0001, and the momentum is 0.9. The initial values of the weights and bias are obtained by the randomization method.

4.3.1. VEDAI Data Set Experimental Results

Table 2 shows the experimental results for all the models for the public data set VEDAI. The precision and recall are the average values of the confidence level in the interval of 0:0.1:1. It can be seen that our detection method performs well in many contrasting models. The detection precision and recall rates are 90.1% and 96.0%, respectively, and the harmonic average F1 is increased by 1.9%, compared to the second-ranked RetinaNet detection method. Our detection method FPS also has certain advantages.

Table 2. Comparison of multiple detection methods in VEDAI.

	Precision (%)	Recall (%)	F1 (%)	FPS	Million Parameters
		Two-stage			
Faster R-CNN [23]	82.1	87.7	84.8	5.8	3.39
Faster R-CNN w FPN	88.9	91.6	90.2	17.4	4.14
Cascade R-CNN [24]	84.1	86.9	85.5	2.3	3.31
Cascade R-CNN w FPN	88.4	90.1	89.2	14.6	6.92
		Single-stage			
SSD512 [26]	77.2	91.5	83.7	22.9	2.44
RetinaNet [38]	87.8	94.7	91.1	18.6	3.63
FCOS [35]	86.9	93.0	89.8	24.7	3.21
FoveaBox [36]	86.5	90.4	88.4	22.7	5.69
Ours	90.1	96.0	93.0	21.5	3.80

Compared with faster RCNN in the two-stage experiments, it can be seen that the detection results for faster RCNN are significantly improved by adding an FPN. This can also be verified by two groups of experiments—cascade RCNN and cascade RCNN with FPN. The results show that the FPN can improve the detection performance in vehicle target detection in VHR remote sensing images. The low-level feature maps can be rich while maintaining sufficient spatial information by combining high- and low-level feature maps. The semantic information enhances the detection of vehicle targets in remote sensing images.

As an anchor-based detection network, SSD512 only performs a single target recognition and regression on the prior box. Although the detection is fast, the accuracy of the detection is reduced. The detection accuracy of 77.2% is much lower than the 82.1% accuracy of faster RCNN. In the experimental results for the single-stage target detection model, the proposal of RetinaNet makes it possible for the detection performance of a single-stage network to exceed the two-stage result. The detection result of RetinaNet is not only better than that of faster RCNN but also significantly higher than that of cascade RCNN. The above results also verify the excellent network structure of RetinaNet and the effectiveness of the focal loss.

However, as an anchor-free variant of the RetinaNet network, it can be seen from the results that FCOS and FoveaBox are inferior to RetinaNet in terms of precision, recall, and F1 score. It can be seen that the direct introduction of the two anchor-free networks to the VHR remote sensing image vehicle target detection task did not improve the results.

This is because the detection models based on FCOS and FoveaBox cannot eliminate the redundant information in the horizontally labeled target in the VHR remote sensing image well. Unlike the FCOS, although FoveaBox shrinks the training area to avoid the influence of low-quality samples around the center of the target on the training results, it is difficult to take into account the characteristics of the arbitrary direction of the target and the complex background information. Compared with the best-performing RetinaNet network in Table 2, we have incorporated an attention mechanism (MA-FPN) into the FPN structure to enhance the feature extraction of vehicle targets and reduce the influence of background information and noise in the VHR remote sensing images. The experimental results show that the proposed method has obvious advantages; the precision not only reaches 90.1%, but the proposed method also leads the other models in the table in terms of recall, reaching 96.0%.

In addition to detection accuracy, another important performance indicator of the target detection algorithm is speed and the space complexity of the detect model. In order to further illustrate the effectiveness of our model, we performed statistics on the parameters of each model and the FPS. Due to the addition of modules to improve detection accuracy, such as MA-FPN and MPFA, the parameters of the proposed model are 3.8 million. However, this is still lower than the number of parameters of the faster R-CNN with FPN, cascade R-CNN with FPN, and FoveaBox. It also can be seen that the performance improvement does not depend on the complex models with many more free parameters, but on an effective pipeline and beneficial architectural modifications.

For the superiority of the proposed method in VHR remote sensing image vehicle detection, we directly compared the results of the vehicle detection methods proposed in [31,46,47] on the VEDAI data set, including a vehicle center, scale, and orientation prediction-based detector (VCSOP detector) [46], convolutional neural networks (CNNs) [47], and Oriented_SSD512 (Oriented_SSD300) [31]; the results are shown in Table 3. Among these methods, the proposed outperforms the aforementioned methods in terms of recall, precision, and F1 score for vehicle detection in remote sensing images. In order to improve the detection performance, the VCSOP detector for orientation-aware vehicle detection contains four subtasks to directly predict high-level vehicle features, and the detection precision and recall rates are 86% and 94.62%, respectively. The proposed method introduces the MA-FPN module for background information filtering and vehicle target enhancement, and the MPFA method is designed to determine the training sample selection area of the vehicle target in the VHR remote sensing image. Hence, the detection precision and recall rates are the best, reaching 90.1% and 96%, respectively. The higher accuracy and recall in this data set further demonstrate the effectiveness and efficiency of the proposed method in the remote sensing image vehicle detection.

The PR curve is used to measure the performance of the models. First, the detection box with a confidence level of less than 0.05 in the detection result was filtered out. Next, the non-maximum suppression method was used to eliminate some overlapping detection boxes, where the intersection over union (IoU) of NMS was set to 0.5. Then, the confidence of the box with IOU greater than 0.5 was calculated. Finally, the precision and recall were calculated under different confidence which changes from 0:0.01:1.

Table 3. Results of different methods on the VEDAI512 data set.

Method	Precision (%)	Recall (%)	F1 (%)
VCSOP detector	86	94.62	90
CNNs	56	79	66
Oriented_SSD512	80.46	60.12	69
Oriented_SSD300	78.36	52.6	63
Ours	90.1	96	93

Figure 10 shows the P–R curves of various models, indicating the pros and cons of our model and other models. In Figure 10a, our model can still maintain the lowest decline in detection accuracy as the recall rate increases. This means that our model can detect vehicle targets as possible, the accuracy is higher, and the error rate is lower. On the contrary, in the other detection models, such as faster RCNN and SSD512, the curve decays faster, the detection accuracy rate also drops significantly, and the recall rate stops when it approaches 90%. AUC is the standard for judging the pros and cons of the two-class prediction model, which reflects the ranking quality predicted by the model, i.e., the ratio of positive examples in front of negative examples. The AUC value of our proposed model in the public data set VEDAI is 0.973. Compared with the comparative model, it shows the superiority of our model.

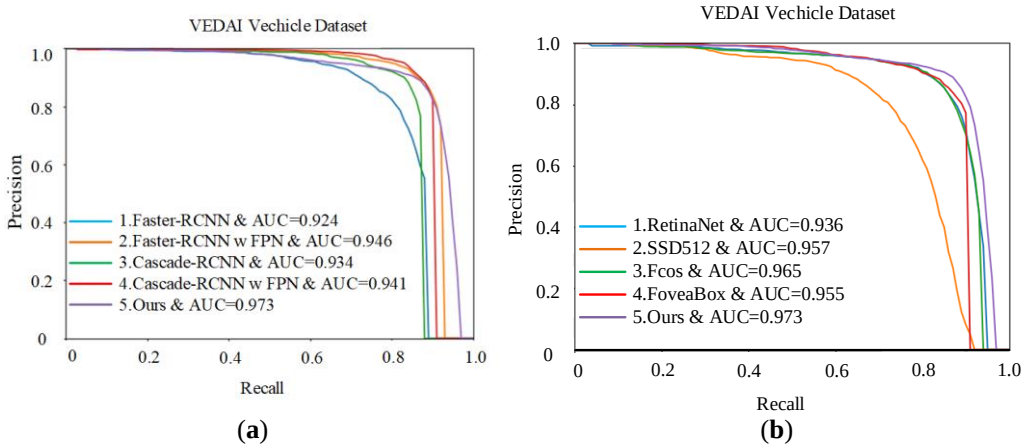


Figure 10. P–R curve of the VEDAI experiment: (**a**) two-stage, (**b**) single-stage.

The example images of the detection result are shown in Figure 11. Figure 11a shows the test picture, Figure 11b shows the faster RCNN detection result, and Figure 11c shows the SSD512 detection result. In the results, the green box means the detected vehicle, the red box means the missed vehicle, and the yellow box means the error detection. It can be seen that many false and missed detections in faster RCNN and SSD512 will lead to low precision and recall.

Figure 11. The detection results for faster RCNN and SSD512: (**a**) test image, (**b**) faster RCNN, (**c**) SSD512.

In order to show the detection results for this model for the VEDAI data set, the detection results are visualized. Figure 12 shows some test results for the proposed model. The

results show that this model has a good detection performance for both three-color images and monochrome images and can accurately detect the vehicle targets to be inspected in remote sensing images.

Figure 12. Detection results for VEDAI detection with the proposed method.

4.3.2. Experimental Results for the RSI Data Set

In the entire VEDAI512 data set, there are only 1.11 vehicle targets in each image. Thus, the task of vehicle detection is relatively simple for the existing detection model. Hence, as a supplement to the VEDAI data set, the RSI we designed focuses on rich background information and dense vehicle targets. There are 38.2 vehicle targets, on average, in each VHR remote sensing image, and the scenes are diverse, which increases the difficulty of vehicle detection. Table 4 shows the experimental results for all models in the RSI data set.

From the results, we can see that our detection model is still better than other comparable models. Our detection results are better in the RSI data set, and the accuracy rate, recall rate, and F1 indicators are improved by at least 5.2%, 8.2%, and 7.2%, respectively.

Table 4. Comparison of multiple detection methods in RSI.

	Precision (%)	Recall (%)	F1 (%)	FPS	Million Parameters
		Two-stage			
Faster R-CNN [23]	72.9	79.6	76.1	5.3	3.39
Faster R-CNN w FPN	73.2	78.0	75.5	16.8	4.14
Cascade R-CNN [24]	73.2	78.6	75.8	2.2	3.31
Cascade R-CNN w FPN	75.9	80.3	78.0	13.6	6.92
		Single-stage			
SSD512 [26]	71.3	85.7	77.8	23.1	2.44
RetinaNet [38]	78.1	86.4	82.0	18.7	3.63
FCOS [35]	80.3	87.1	83.6	24.2	3.21
FoveaBox [36]	80.7	86.8	83.6	22.3	5.69
Ours	85.9	95.3	90.4	21.4	3.80

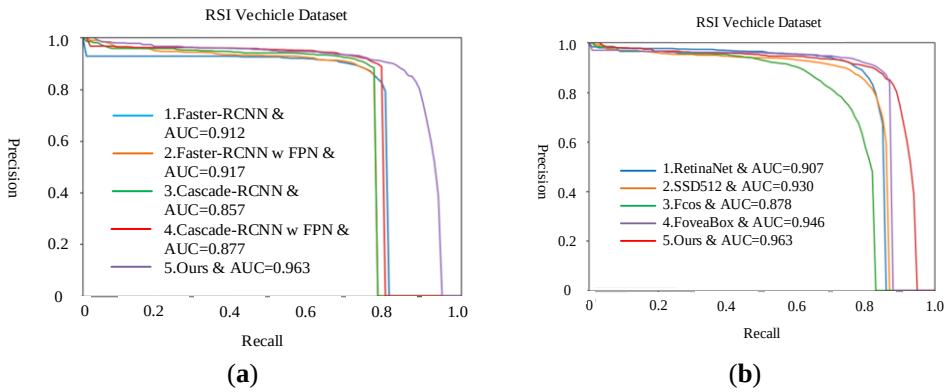
Table 4 shows the experimental results for the RSI data set. Compared with the detection results for the VEDAI data set, the detection performance has significantly decreased, which means that the detection difficulty of the RSI data set is higher than that of the VEDAI data set. However, the pros and cons of the performance between the models do not change with different data sets. From the experimental results, it can be seen that the detection results for the anchor-free models—i.e., FCOS and FoveaBox—surpass those of the anchor-based model, indicating that the anchor-free model has more advantages with respect to the RSI data set. This is because the remote sensing images of the RSI data

set contain more vehicle targets, which may even exceed the upper limit of the proposed area in the anchor-based model, and the vehicle targets also have complex backgrounds and rich colors. Additionally, the vehicle targets are interfered with by environmental factors such as shadow, occlusion, and reflection, which also increase the difficulty of detecting vehicle targets in the RSI data set. As a result, the discrete anchor point scale in the anchor-based method makes it difficult to match the vehicle target well, and it may introduce too many negative samples. The FCOS and FoveaBox based on the anchor-free model can detect vehicle targets as well as possible by predicting all the pixels in the feature map. At the same time, the focal loss compensates effectively for the imbalance between foreground and background categories and reduces the impact of negative samples on the detection performance. Therefore, the anchor-free method can still maintain a good detection performance in complex scenes.

Compared with FCOS and FoveaBox, as an anchor-free method, our model still has the best detection performance, reaching 85.9%, 95.3%, and 90.4% in terms of precision, recall, and F1 score, respectively. The reason why our model has better detection results is that, on the one hand, our model uses the MA-FPN to decrease the influence of image background information and noise in the feature map and enhance the semantic information of the vehicle target in the feature map. For a vehicle target in shadow or occlusion, the channel that represents the semantic information of the current pixel area is highlighted through the channel attention mechanism. However, in FCOS and FoveaBox, all channels in the pixel area of the current feature map are treated equally, making the normal vehicle target different from the vehicle target in terms of shadow or occlusion, which affects the learning and detection of the subsequent network. On the other hand, by using the MPFA, the training samples of our model do not contain background information in the VHR remote sensing images, which further reduces the influence of negative samples on the network.

In summary, the proposed model can still maintain good detection results for the RSI data set, which further demonstrates its performance and robustness. The overall performance is shown in both the VEDAI data set and the RSI data set. In order to illustrate the effectiveness of the MA-FPN and MPFA modules, corresponding ablation experiments are subsequently described in the next section.

From the P–R curves in Figure 13a,b, it can be seen that the PR curves of some models have a rapid decline when the recall rate is increased to a certain level; this is especially obvious for the two-stage network. This is because there are too many vehicle targets in the VHR remote sensing image. As the percentage of the total number of vehicles correctly detected by the network increases, the ratio of all the detected vehicles that are indeed vehicle targets decreases, and the detection accuracy is seriously attenuated. The results also show that the anchor-based model has an upper limit on the detection result due to the anchor.



(a) **(b)**

Figure 13. PR curve of the RSI experiment: (**a**) two-stage, (**b**) single-stage.

Figure 14a–g show the detection results for different anchor-based models on the RSI data set. The green box is the detection result, and the red box is the undetected vehicle. It

can be seen that our detection method has a high detection limit, especially for partially occluded vehicle targets, with which it still has a good detection performance. The AUC values also show the superiority of our model.

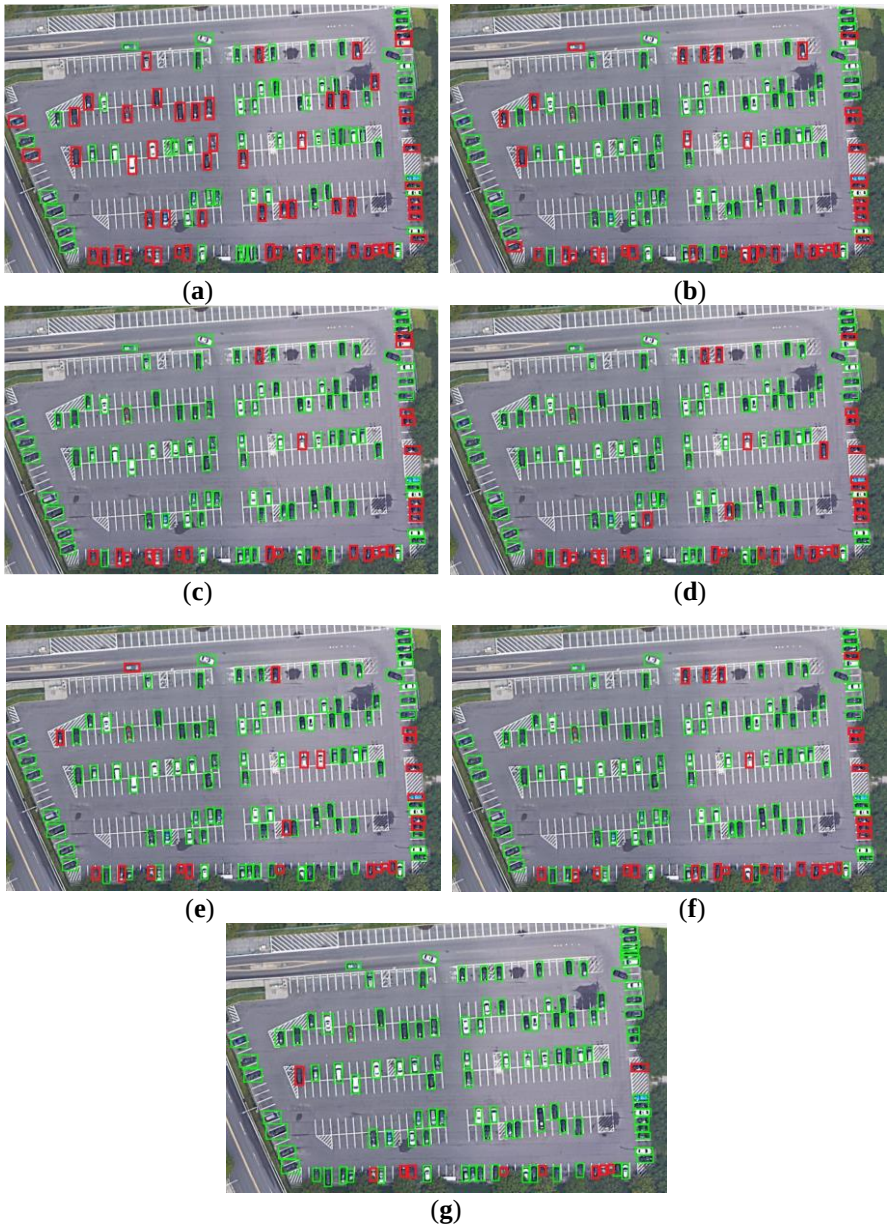


Figure 14. RSI detection results: (**a**) SSD512, (**b**) faster RCNN-FPN, (**c**) RetinaNet, (**d**) cascade RCNN-FPN, (**e**) Foveabox, (**f**) FCOS, (**g**) our model.

4.4. Ablation Experiment

In this section, we verify the effectiveness of the proposed method through ablation experiments.

4.4.1. Evaluation of MA-FPN

The image is forwarded through a multi-layer network, and the scale is doubled and reduced, causing the object's location information to be lost quickly, which is unfavorable for the minimum enclosed rectangle regression of the vehicle target. Therefore, we incorporate a variety of attention information into the FPN top-down process to improve the representation of vehicle targets in the feature map from the channel and space levels. The specific fusion experiment results are shown in Tables 5 and 6. It can be seen that the

addition of attention information improves the detection effect significantly. The accuracy rate is increased from 86.5% to 89.1%, and the recall rate is increased from 90.4% to 96.2% by fusing spatial and channel attention information. The F1 score is improved by 4.1%.

Table 5. Performance of the FPN that incorporates multiple attention mechanisms in RSI.

	Spatial Attention	Channel Attention	Precision (%)	Recall (%)	F1 (%)
			80.7	86.8	83.6
FoveaBox [36]		√	83.7	90.9	87.2
	√		84.3	90.8	87.4
MA-FPN	√	√	84.7	91.3	87.9

Table 6. Performance of the FPN that incorporates multiple attention mechanisms in VEDAI.

	Spatial Attention	Channel Attention	Precision (%)	Recall (%)	F1 (%)
			86.5	90.4	88.4
FoveaBox [36]		√	87.9	95.7	91.6
	√		87.2	95.7	91.3
MA-FPN	√	√	**88.5**	**96.0**	**92.1**

In summary, adding an attention mechanism to the top-down process of the feature pyramid is a simple and effective method. To preserve the spatial information of the vehicle target, the missing semantic information will be screened by the channel attention so that the semantic information that is more suitable for vehicle target detection can receive a higher weight. Additionally, the spatial attention makes the vehicle target area information clearer and reduces the interference of complex backgrounds and noise in remote sensing images in the detection results.

4.4.2. Evaluation of MPFA

The existing anchor-free method discards the manually specified anchor and chooses to classify and regress the target in the image at the pixel level. The original anchor-based method is based on the IOU of the proposal and ground truth to determine the positive and negative samples, but the optimal solution has not yet been found to determine the positive and negative samples in the anchor-free method. In this paper, by combining the instance information of the vehicle target to determine the true foveal area of the vehicle, the uncertainty of the redundant pixels in the ground truth is eliminated.

From the experimental results in Tables 7 and 8, we can see that, by re-determining more accurate positive sample pixels, the accuracy, recall, and F1 score are improved by 3.4%, 5.6%, and 4.4%, respectively.

Table 7. Performance of a more precise foveal area in RSI.

	Precision (%)	Recall (%)	F1 (%)
Foveabox [36]	86.5	90.4	88.4
MPFA	**89.9**	**96.0**	**92.8**

Table 8. Performance of a more precise foveal area in VEDAI.

	Precision (%)	Recall (%)	F1 (%)
Foveabox [36]	80.7	86.8	83.6
MPFA	**82.9**	**91.6**	**86.8**

In short, the positive sample pixels we determined all represent the vehicle itself and are evenly distributed. There will not be a situation in which the entire vehicle can be determined only through certain points. Due to the shrinkage of the vehicle instance area, the positive sample pixels we obtained can make the regression interval more relaxed when

performing minimum enclosed rectangle regression, which can eliminate the interference of redundant areas well.

5. Conclusions

In this paper, we have proposed a novel end-to-end anchor-free single-stage target detection model for detecting vehicle targets in remote sensing images. First, we used the anchor-free single-stage target detection framework to alleviate the contradiction between detection speed and detection accuracy. Second, to solve the contradiction between the insufficient semantic information of low-level feature maps and the low amount of spatial information in high-level feature maps, the spatial and channel attention mechanisms were integrated in the top-down process of FPN. For the task of vehicle detection in remote sensing images, we added the instance information of vehicles to determine a more accurate foveal area and chose a novel positive sample determination method in the anchor-free model. Compared with the existing methods, the area determined using the proposed method is closer to the central area and does not change with the shape and direction of the object. Finally, our experimental results show that our proposed model performs well in terms of detection accuracy and speed.

However, the proposed method cannot identify the type of vehicle and uses a rotating frame to detect vehicles. In future work, we need to enrich the types of data sets to achieve multi-class detection of vehicle targets. Additionally, we will directly detect rotating vehicle targets through a single-stage anchor-free model. At the same time, although the MA-FPN module has proved that the attention mechanism is simple to use in target detection, and it has improved obvious characteristics, the advantages of MA-FPN cannot be judged from the structure, and there is still some room for the actual use of attention information. The influence of the spatial resolution, height, angle, vehicle color, and other information of remote sensing images on the detection results is also the focus of our future research.

Author Contributions: Conceptualization, Xungen Li and Mian Pan; methodology, Mian Pan and Shuaishuai Lv; software, Feifei Men and Xiao Jiang; validation, Xungen Li; formal analysis, Mian Pan and Feifei Men; data curation, Feifei Men and Shuaishuai Lv; writing—original draft preparation, Feifei Men; writing—review and editing, Mian Pan, Shuaishuai Lv, and Haibin Yu; supervision, Xungen Li and Qi Ma. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Key Research and Development Project of China (Grants No. 2016YFC1400302), the National Natural Science Foundation of China (Grant No. 61501155, 61871164), National Defense Science and Technology Key Laboratory Fund (6142401200201), and Zhejiang Provincial Natural Science Foundation of China (No. LQ19E070003).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available from the author upon reasonable request.

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

References

1. Xia, G.S.; Hu, J.; Hu, F.; Shi, B.; Bai, X.; Zhong, Y.; Zhang, L.; Lu, X. AID: A Benchmark Data Set for Performance Evaluation of Aerial Scene Classification. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 3965–3981. [CrossRef]
2. 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 IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018.
3. Mesquita, D.B.; Dos Santos, R.F.; Macharet, D.G.; Campos, M.F.M.; Nascimento, E.R. Fully Convolutional Siamese Autoencoder for Change Detection in UAV Aerial Images. *IEEE Geosci. Remote Sens. Lett.* **2020**, *17*, 1455–1459. [CrossRef]
4. Gómez-Candón, D.; De Castro, A.I.; Lopez-Granados, F. Assessing the accuracy of mosaics from unmanned aerial vehicle (UAV) imagery for precision agriculture purposes in wheat. *Precis. Agric.* **2013**, *15*, 44–56. [CrossRef]
5. Javadi, S.; Dahl, M.; Pettersson, M.I. Change Detection in Aerial Images Using Three-Dimensional Feature Maps. *Remote Sens.* **2020**, *12*, 1404. [CrossRef]

6. Tang, Y.; Zhang, C.; Gu, R.; Li, P.; Yang, B. Vehicle detection and recognition for intelligent traffic surveillance system. *Multimed. Tools Appl.* **2017**, *76*, 5817–5832. [CrossRef]
7. Leitloff, J.; Rosenbaum, D.; Kurz, F.; Meynberg, O.; Reinartz, P. An Operational System for Estimating Road Traffic Information from Aerial Images. *Remote Sens.* **2014**, *6*, 11315–11341. [CrossRef]
8. Hu, X.; Xu, X.; Xiao, Y.; Chen, H.; He, S.; Qin, J.; Heng, P.-A. SINet: A Scale-Insensitive Convolutional Neural Network for Fast Vehicle Detection. *IEEE Trans. Intell. Transp. Syst.* **2019**, *20*, 1010–1019. [CrossRef]
9. Nicolas, A.; Bertrand, L.S.; Lefèvre, S. Segment-before-Detect: Vehicle Detection and Classification through Semantic Segmentation of Aerial Images. *Remote Sens.* **2017**, *9*, 368.
10. Yin, J.; Liu, L.; Li, H.; Liu, Q. The infrared moving object detection and security detection related algorithms based on W4 and frame difference. *Infrared Phys. Technol.* **2016**, *77*, 302–315. [CrossRef]
11. Ok, A.O.; Senaras, C.; Yuksel, B. Automated Detection of Arbitrarily Shaped Buildings in Complex Environments from Monocular VHR Optical Satellite Imagery. *IEEE Trans. Geosci. Remote Sens.* **2013**, *51*, 1701–1717. [CrossRef]
12. Wang, C.; Bi, F.; Zhang, W.; Chen, L. An Intensity-Space Domain CFAR Method for Ship Detection in HR SAR Images. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 529–533. [CrossRef]
13. Siyu, W.; Xin, G.; Hao, S.; Xinwei, Z.; Xian, S. An aircraft detection method based on convolutional neural networks in high-resolution SAR images. *Radars* **2017**, *6*, 195–203.
14. Jiandan, Z.; Tao, L.; Guangle, Y. Robust Vehicle Detection in Aerial Images Based on Cascaded Convolutional Neural Networks. *Sensors* **2017**, *17*, 2720.
15. Mandal, M.; Shah, M.; Meena, P.; Devi, S.; Vipparthi, S.K. AVDNet: A Small-Sized Vehicle Detection Network for Aerial Visual Data. *IEEE Geosci. Remote Sens. Lett.* **2020**, *17*, 494–498. [CrossRef]
16. Xiaofei, L.; Tao, Y.; Jing, L. Real-Time Ground Vehicle Detection in Aerial Infrared Imagery Based on Convolutional Neural Network. *Electronics* **2018**, *7*, 78.
17. Lowe, D.G. Object Recognition from Local Scale-Invariant Features. In Proceedings of the IEEE Transactions on Pattern Analysis and Machine Intelligence, Kerkyra, Greece, 20–27 September 1999; Volume 2, pp. 1150–1157.
18. Liu, K.; Mattyus, G. Fast multi-class vehicle detection on aerial images. *IEEE Geosci. Remote Sens. Lett.* **2015**, *12*, 1938–1942.
19. Dalal, N.; Triggs, B. Histograms of Oriented Gradients for Human Detection. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR, San Diego, CA, USA, 20–25 June 2005; Volume 1, pp. 886–893.
20. Moranduzzo, T.; Melgani, F. Automatic Car Counting Method for Unmanned Aerial Vehicle Images. *IEEE Trans. Geosci. Remote Sens.* **2013**, *52*, 1635–1647. [CrossRef]
21. Viola, P.; Jones, M. Rapid Object Detection Using a Boosted Cascade of Simple Features. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Kauai, HI, USA, 8–14 December 2001; pp. 511–518.
22. Shao, W.; Yang, W.; Liu, G. Car Detection from High-Resolution Aerial Imagery Using Multiple Features. In Proceedings of the IEEE International Geoscience and Remote Sensing Symposium, Munich, Germany, 22–27 July 2012.
23. Ren, S.; He, K.; Girshick, R. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef] [PubMed]
24. Cai, Z.; Vasconcelos, N. Cascade R-CNN: Delving Into High Quality Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018.
25. Redmon, J.; Farhadi, A. YOLOv3: An Incremental Improvement. *arXiv* **2018**, arXiv:1804.02767.
26. Liu, W.; Anguelov, D.; Erhan, D. SSD: Single Shot MultiBox Detector. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 8–16 October 2016.
27. Deng, Z.; Sun, H.; Zhou, S.; Zhao, J.; Zou, H. Toward Fast and Accurate Vehicle Detection in Aerial Images Using Coupled Region-Based Convolutional Neural Networks. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2017**, *10*, 3652–3664. [CrossRef]
28. Lichao, M.; Xiang, Z.X. Vehicle Instance Segmentation from Aerial Image and Video Using a Multi-Task Learning Residual Fully Convolutional Network. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 6699–6711.
29. Li, Q.; Mou, L.; Xu, Q.; Zhang, Y.; Zhu, X.X. R^3-Net: A Deep Network for Multioriented Vehicle Detection in Aerial Images and Videos. *arXiv* **2019**, arXiv:1808.05560. [CrossRef]
30. Dai, J.; Li, Y.; He, K.; Sun, J. R-FCN: Object Detection via Region Based Fully Convolutional Networks. In Proceedings of the 30th Conference on Neural Information Processing Systems, Barcelona, Spain, 5–10 December 2016; pp. 379–387.
31. Tang, T.; Zhou, S.; Deng, Z.; Lei, L.; Zou, H. Arbitrary-Oriented Vehicle Detection in Aerial Imagery with Single Convolutional Neural Networks. *Remote Sens.* **2017**, *9*, 1170. [CrossRef]
32. Chen, C.; Zhong, J.; Tan, Y. Multiple-Oriented and Small Object Detection with Convolutional Neural Networks for Aerial Image. *Remote Sens.* **2019**, *11*, 2176. [CrossRef]
33. Liu, C.; Ding, Y.; Zhu, M.; Xiu, J.; Li, M.; Li, Q. Vehicle Detection in Aerial Images Using a Fast Oriented Region Search and the Vector of Locally Aggregated Descriptors. *Sensors* **2019**, *19*, 3294. [CrossRef] [PubMed]
34. Lin, T.Y.; Dollár, P.; Girshick, R. Feature Pyramid Networks for Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017.
35. Tian, Z.; Shen, C.; Chen, H. FCOS: Fully Convolutional One-Stage Object Detection. In Proceedings of the International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019.

36. Lin, T.Y.; Goyal, P.; Girshick, R. Focal Loss for Dense Object Detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; Volume 99, pp. 2999–3007.
37. Long, J.; Shelhamer, E.; Darrell, T. Fully Convolutional Networks for Semantic Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; Volume 39, pp. 640–651.
38. Kong, T.; Sun, F.; Liu, H. FoveaBox: Beyond Anchor-based Object Detector. *IEEE Trans. Image Process.* **2019**, *29*, 7389–7398. [CrossRef]
39. Dumoulin, V.; Perez, E.; Schucher, N.; Strub, F.; Vries, H.D.; Courville, A.; Bengio, Y. Feature-wise transformations: A simple and surprisingly effective family of conditioning mechanisms. *Distill* **2018**. [CrossRef]
40. Park, J.; Woo, S.; Lee, J.Y.; Kweon, I.S. BAM: Bottleneck attention module. *arXiv* **2018**, arXiv:1807.06514.
41. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. CBAM: Convolutional block attention module. In Proceedings of the European conference on computer vision, Munich, Germany, 8–14 September 2018; pp. 3–19.
42. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2015**, arXiv:1409.15.
43. Razakarivony, S.; Jurie, F. Vehicle Detection in Aerial Imagery: A small target detection benchmark. *J. Vis. Commun. Image Represent.* **2015**, *34*, 187–203. [CrossRef]
44. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. ImageNet Large Scale Visual Recognition Challenge. *Int. J. Comput. Vis.* **2015**, *115*, 211–252. [CrossRef]
45. Bottou, L. Large-Scale Machine Learning with Stochastic Gradient Descent. In Proceedings of the COMPSTAT'2010, Paris, France, 22–27 August 2010; Lechevallier, Y., Saporta, G., Eds.; Springer: Berlin/Heidelberg, Germany, 2010. [CrossRef]
46. Shi, F.; Zhang, T.; Zhang, T. Orientation-Aware Vehicle Detection in Aerial Images via an Anchor-Free Object Detection Approach. *IEEE Trans. Geosci. Remote Sens.* **2021**, *69*, 5221–5233. [CrossRef]
47. Yu, Y.; Gu, T.; Guan, H.; Li, D.; Jin, S. Vehicle Detection From High-Resolution Remote Sensing Imagery Using Convolutional Capsule Networks. *IEEE Geosci. Remote Sens. Lett.* **2019**, *16*, 1894–1898. [CrossRef]