remote sensing

MDPI

Article

Improved Faster RCNN Based on Feature Amplification and Oversampling Data Augmentation for Oriented Vehicle Detection in Aerial Images †

Nan Mo and Li Yan *

School of Geodesy and Geomatics, Wuhan University, 129 Luoyu Road, Wuhan 430079, China; nmo@whu.edu.cn

* Correspondence: lyan@sgg.whu.edu.cn

† This paper is an extended version of our conference paper that will be published in Nan Mo, Li Yan. Oriented Vehicle Detection in High-resolution Remote Sensing Images based on Feature Amplification and Category Balance by Oversampling Data Augmentation. In Proceedings of XXIV International Society for Photogrammetry and remote sensing (ISPRS 2020), Nice, France, 14–20 June 2020.

Received: 27 June 2020; Accepted: 7 August 2020; Published: 9 August 2020

Abstract: Vehicles in aerial images are generally with small sizes and unbalanced number of samples, which leads to the poor performances of the existing vehicle detection algorithms. Therefore, an oriented vehicle detection framework based on improved Faster RCNN is proposed for aerial images. First of all, we propose an oversampling and stitching data augmentation method to decrease the negative effect of category imbalance in the training dataset and construct a new dataset with balanced number of samples. Then considering that the pooling operation may loss the discriminative ability of features for small objects, we propose to amplify the feature map so that detailed information hidden in the last feature map can be enriched. Finally, we design a joint training loss function including center loss for both horizontal and oriented bounding boxes, and reduce the impact of small inter-class diversity on vehicle detection. The proposed framework is evaluated on the VEDAI dataset that consists of 9 vehicle categories. The experimental results show that the proposed framework outperforms previous approaches with a mean average precision of 60.4% and 60.1% in detecting horizontal and oriented bounding boxes respectively, which is about 8% better than Faster RCNN.

Keywords: oriented vehicle detection; oversampling data augmentation; feature amplification; center loss; aerial image

1. Introduction

The development of high-resolution remote sensing images makes vehicle detection possible, which is important for autonomous driving and traffic monitoring [1–3]. In some more specific tasks, types and orientations of vehicles are required so that traffic conditions can be better scheduled [4]. Therefore, the oriented vehicle detection of multiple types is significant. The transportations such as car, tractor, vans, plane, pick-up and so on are the common vehicles existing in aerial images and they are studied in this paper. The oriented vehicle detection is a more challenging task compared with multi-class object detection problems since the difference between different vehicle categories is small and vehicles are usually small objects [5].

Two commonly used definitions of small objects are as follows. One definition is the relative size. An object will be regarded as a small object if the size of an object is below 10% of the original images. The other definition is that an object will be regarded as small objects if their sizes are below 32×32 pixels. In high-resolution remote sensing images, vehicles usually occupy a small area below 10% of the image size or smaller than 32×32 pixels. We take the Vehicle Detection in Aerial

Imagery (VEDAI) [6] dataset as an example to prove this. The size of each image is 1024 × 1024 pixels. The VEDAI dataset contains 9 vehicle categories. The statistical results shown in Figure 1 demonstrate that most types of vehicles in this dataset are small objects.

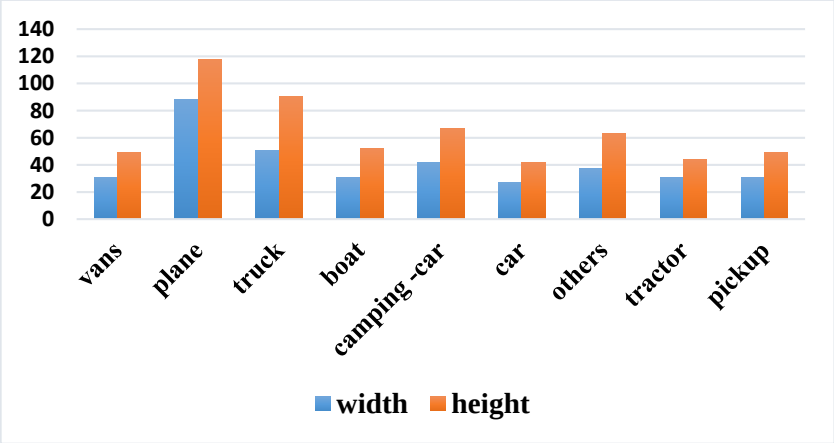


Figure 1. Average length and width of different vehicles in VEDAI dataset.

Traditional vehicle detection methods usually include four steps: (1) Data preprocessing such as improving the image quality and increasing the contrast between vehicles and their backgrounds. (2) Determination of potential positions of vehicles by calculating the contrast between different parts of images. (3) Segmentation is performed to accurately extract potential location of vehicles from the background. (4) Vehicles are finally recognized by extracted features from potential regions.

Recent vehicle detection methods are completely different from traditional methods since they try to decrease the influence of intermediate decisions on detection results obtained by machine learning methods. They are made up of handcrafted features-based methods and deep learning-based methods according to the different types of extracted features used [7]. Before 2012, handcrafted features-based approaches were the mainstream algorithms for vehicle detection. However, handcrafted features including Viola Jones Detectors [8], Bag of Words (BOW) [9], Deformable Parts Model (DPM) [10] and Histogram of Oriented Gradients (HOG) [11], cannot represent vehicles well because they lack the semantic information which is important for recognizing vehicles.

The development of vehicle detection approaches has been promoted since deep learning architectures appeared in 2012. Existing deep learning-based vehicle detection approaches can be divided into one-stage vehicle detection approaches such as Single Shot Multi-Box Detector (SSD) [12], You Only Look Once (YOLO) [13], YOLOv2 [14], YOLOv3 [15], YOLOv4 [16] and two-stage vehicle detection methods such as Region CNN (RCNN) [17], Spatial Pyramid Pooling Network (SPP-Net) [18], Fast RCNN [19] and Faster RCNN [20], according to the different detection processes employed. Compared with one-stage approaches the two-stage methods can achieve higher precision ratio with speeds that can meet real-time requirements. Therefore, this paper mainly investigates two-stage deep learning-based approaches.

When two-stage deep learning-based methods are applied to small vehicle detection, the following limitations may exist:

1. The quantity imbalance between diverse types of vehicles in the training dataset caused by the random frequency and spatial distribution of vehicles will have a negative influence on training the network. The Convolutional Neural Network (CNN) models tends to focus on vehicle categories with a larger number of samples, which may have a negative influence on detecting vehicle categories with a smaller number of samples.
2. The features of small objects are less detailed than those of large or medium objects, which increases the difficulty of detecting vehicles. Features extracted by CNN contain more semantic information

but the pooling operation in the CNN reduces the detailed information hidden in deep features, which decrease the discriminative ability of features in distinguishing different vehicles.

3. There may exist high inter-class similarity in vehicle detection as shown in Figure 2. Moreover, factors such as vehicle type, environmental background, lighting condition and shooting angle may lead to the increased diversity of vehicle representations, which may increase the difficulty of determining vehicle types.
4. The aerial images acquired by airborne sensors are taken overhead, which may lead to random directions of vehicles. The traditional horizontal bounding box (HBB) can only roughly predict the position of objects. For oriented objects such as vehicles, oriented bounding box (OBB) should be used to describe the position of objects more accurately.

Figure 2. Examples of different vehicle types in VEDAI dataset.

Considering the above problems and the research status of addressing each problem shown in Section 2. This paper proposes an oriented vehicle detection framework based on improved Faster RCNN for aerial images. The major contributions can be summarized as follows in this paper.

- Different from the basic data augmentation methods, we propose a data augmentation strategy by oversampling and stitching to reduce the negative impact of category imbalance and construct a dataset with balanced number of samples.
- The pooling operation in CNN may reduce the discriminative ability of features in distinguishing small objects. We perform feature amplification by bilinear interpolation in the last feature map to increase the capability of features with more simple operations.
- Considering the small inter-class diversity in different types of vehicles, center loss is introduced to the loss function in order to increase the model's ability to distinguish different vehicle types.
- Considering the random direction of vehicles, the oriented bounding boxes and horizontal bounding boxes are jointly trained in the same framework so as to more accurately determine the position of the vehicles.

The remainder of this paper can be highlighted as follows. Section 2 gives a brief introduction to the related work. Oriented vehicle detection based on feature amplification and oversampling data augmentation in aerial images is presented in Section 3. Section 4 describes the implementation details and dataset description along with vehicle detection results and ablation studies. Section 5 discusses and analyzes the experimental results in Section 4. Section 6 concludes the experimental conclusions with future directions.

2. Related Work

2.1. Class Imbalance Problem

Class imbalance of objects in the training dataset is a common problem in the object detection. There are usually two types of category imbalance: foreground-background and foreground-foreground imbalance [21]. Foreground-foreground imbalance is studied in this paper since it negatively affects the multi-class object detection.

Numerous studies have been done to address the foreground-foreground imbalance problem in the computer vision field. Ouyang et al. [22] proposed to fine-tune the distribution of under-represented categories by clustering similar categories to address the class imbalance. Oksuz et al. [23] proposed a foreground balanced sampling method, which decreases the imbalance between distributions of different objects within each batch by assigning a probability to each true bounding box. Wang et al. [24] proposed a sample exchange strategy to generate new samples and decrease the imbalance by exchanging the same type of objects in different natural images.

The above methods are mainly aimed at addressing the class imbalance problem in natural imagery. In the remote sensing field, few researchers have considered the imbalance between types of training samples, and existing data augmentation methods are usually aimed at enhancing the generalization ability of the model. The imbalanced class distribution makes the network training favor the vehicle categories with a larger number, which leads to unideal results for other categories. Therefore, this paper designs an oversampling and stitching data augmentation method for aerial images so that the numbers of different vehicles can be balanced for training CNN.

2.2. Represention of Small Objects

Due to low resolution, blurred images, less information and more noise, small object detection has been a difficult problem in object detection. Some researchers have carried out some works in order to improve the performance in representing small objects. These methods mainly consist of improving the resolution of images containing small objects and enriching the detail information of the feature maps describing the small objects.

In terms of improving image resolution, Ji et al. [25] fused an object detection network with an image super-resolution reconstruction network in order to increase the resolution of the original images. Bharat Singh et al. [26] proposed to establish multi-scale pyramids for training images by resizing training images. Moktari et al. [27] proposed a joint super-resolution and vehicle detection network that tries to generate high-resolution images of vehicles from low-resolution aerial images. By increasing the image resolution, the discriminative ability of the features can be enhanced.

In enriching the detail information of feature maps, Hilal et al. [28] proposed deconvolving feature maps continuously in order to increase the ability of shallow features to distinguish diverse objects. Mandal at el. [29] proposed using AVDNet to enlarge feature maps for vehicles by introducing ConvRes modules to difference scale layers. Lin et al. [30] proposed a layer-by-layer prediction using feature pyramids to detect multi-scale objects, which predicts the output of the feature map of each layer of the CNN and finally selects the optimal detection results.

The above methods based on feature pyramids or image pyramids increase the computational cost and set high requirements for computer graphics cards. In addition, complex deep learning architectures may not achieve the desired results in detecting diverse vehicles. Different from existing complex architectures, in this paper we perform a simple but effective bilinear interpolation to amplify the feature map and enrich detail information while maintaining the deep semantic information, which may increase the discrimination of features in representing vehicles.

2.3. The Discriminative Ability of Features

Vehicle is the typical small-scale object and the difference between diverse types of vehicles is relatively little. Therefore, it is more difficult to distinguish the specific category of each vehicle.

In terms of increasing the discriminative ability of features in distinguishing diverse objects, Li et al. [31] propose to use contextual features for increased discrimination of features. Contextual information related to the objects is proved to be helpful to improve the ability of the features [32]. Deng et al. [33] use different feature layers to extract candidate regions in the region proposal network (RPN) for increased recall ratio of multi-class objects. Deep and shallow features are also concatenated to increase the precision ratio of objects in the classification stage.

However, the above methods mainly address multi-category geographic object detection of remote sensing images rather than vehicle detection tasks. Compared with multi-class object detection, vehicle detection demonstrates higher inter-class similarity, which may increase the possibility of misclassification. In this paper, center loss [34] that can control intra-class differences is introduced to improve the capability of features to distinguish different vehicles.

2.4. Oriented Object Detection

The remote sensing images are acquired by overhead sensors and vehicles are a kind of moving object with random spatial distribution and arbitrary orientation. Traditional horizontal bounding boxes can only roughly describe the position of vehicles. Recently, some researchers have studied object detection algorithms with oriented bounding boxes. Ma et al. [35] proposed an oriented text detection algorithm to detect inclined text in the field of text detection. Yang et al. [36] proposed a multi-oriented ship detection algorithm of remote sensing images. Ding et al. [37] proposed an oriented multi-class object detection method in aerial images.

Few studies have been done to achieve both horizontal and oriented detection results in a CNN to get more accurate position of the vehicles. Therefore, this paper designs a joint training loss function for horizontal and oriented bounding boxes to regress the vehicle position and direction.

3. Materials and Methods

3.1. Overall Architecture

In this part, we introduce the proposed oriented vehicle detection algorithm for aerial images based on feature amplification and oversampling based data augmentation. This paper takes the Faster RCNN as the research basis and makes improvements on it. Figure 3 depicts the overall structure of the algorithm. The basic feature extractor in the proposed framework is the Resnet101 [38]. The proposed framework mainly consists of three parts, (1) Oversampling and stitching data augmentation, (2) Enlarging feature maps and (3) A joint training loss function combined with center loss for horizontal and oriented bounding boxes. Each step can be illustrated as follows.

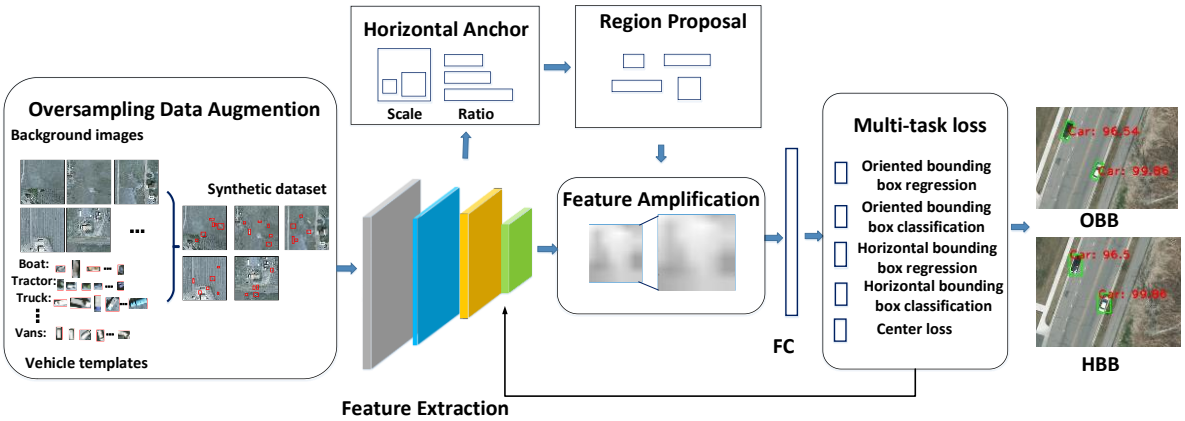


Figure 3. The flowchart of oriented vehicle detection based on feature amplification and oversampling data augmentation in aerial images.

First of all, we perform oversampling and stitching data augmentation on the training dataset by increasing the frequency of vehicles with fewer number of training data to synthesize a new dataset.

In the stage of RPN, we set up multi-scale and multi-shape horizontal anchors and select positive and negative samples for training a RPN network, by calculating the overlap between anchors and ground truths.

In the stage of classification, we amplify the feature map for increased ability of feature maps to represent vehicles. Considering the orientation of vehicles, we propose a multi-task loss function, which jointly trains oriented and horizontal bounding boxes, and introduces the center loss to decrease within-class difference.

3.2. Data Augmentation for Foreground-Foreground Imbalance Problem by Oversampling and Stitching

Motivation of data augmentation by oversampling and stitching. The proposed data augmentation method is aimed to address the foreground-foreground category imbalance problem. It is a common problem in vehicle detection since the frequency and location of different types of vehicles in aerial images are random. When there exists large quantity variance between diverse vehicles, objects may be over-presented or under-represented in the training process.

Two factors contribute to the foreground-foreground category imbalance, namely the imbalanced category distribution in a dataset and that within a batch of samples. We have counted the number of 9 types of vehicles in the VEDAI dataset. Table 1 describes the statistical results according to a descending order of vehicle number.

Table 1. Number of each vehicle category in VEDAI dataset.

Class	Car	Pick-Up	Camping Car	Truck	Other	Tractor	Boat	Vans	Plane
Number	1340	950	390	300	200	190	170	100	47

The above statistical results show that there exists a serious foreground-foreground category imbalance in the VEDAI dataset, which will negatively affect the detection results of vehicles with a small number of samples. In addition, vehicles occupy less image areas, and the vehicles with lower frequency usually have fewer matched anchors, which may increase the difficulty to learn useful information from the network.

Considering that each image contains only a small number of vehicles and backgrounds are with a large image area, this paper designs a data augmentation method based on oversampling and stitching in order to decrease the impact of foreground-foreground imbalance on the training process. The central idea of the proposed method shown in Figure 4 can be illustrated as follows.

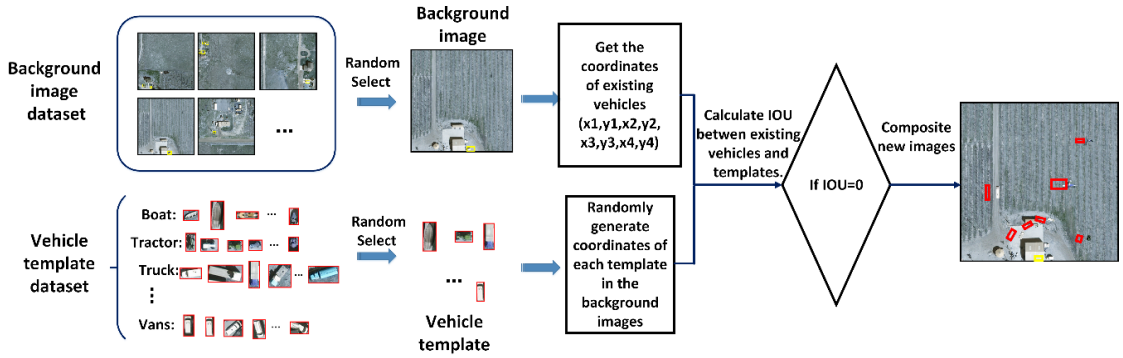


Figure 4. Schematic of oversampling and stitching data augmentation for foreground-foreground class imbalance problem.

Step 1: Augment the original training images by rotating them with angles of 90°, 180°, 270° to generate the rotation dataset ensuring the diversity of object direction.

Step 2: Segment each vehicle from the rotation dataset in Step 1 according to the type and location of vehicles in order to establish the vehicle template dataset. Meanwhile, considering that vehicles in each image occupy only a small area, the images in the rotation dataset with less 10 vehicles are selected as the background image dataset.

Step 3: Count the number of vehicles in each category in the rotation dataset. We take the most numerous type as an expansion benchmark. In order to keep a balance between the quantities of vehicles, the number of vehicles in each category to be augmented should be calculated.

Step 4: For each type of vehicles, certain number of images from the background dataset and a random vehicle from the template dataset are used to synthetize the new training images. We try to make each synthetized image include all types of vehicles to reduce the imbalanced distribution of the samples within a training batch.

Step 5: Considering the random location of vehicles in the geographic space, randomly generate the position of the vehicles in the background images. In order to avoid repetition, we calculate whether there is an overlap between positions of newly generated vehicles and those of original vehicles in the image. When the overlap is 0, image synthesis is performed. The gray values of generated vehicles replace those of original pixels in the background image.

Step 6: Repeat Steps 4 and 5 until the number of vehicles from different categories in the training dataset is balanced.

3.3. Amplification of Deep Features for Small Objects

Motivation of deep feature amplification. The pooling operations can decrease the number of deep neural network parameters but may lose the details of feature maps for small objects. Feature amplification can enlarge the deep feature map and restore the detailed information of the feature map. We use bilinear interpolation in the last feature map to increase the capability of features in representing small objects with more simple operations.

Resnet101 is the backbone for feature extraction in this paper. Figure 5 shows the structure of Resnet101 including four pooling operations. If a vehicle with the size of 32×32 pixels undergoes 4 pooling operations, the corresponding feature map size is 2×2 pixels. However, feature map of 2×2 pixels cannot fully describe the information of a vehicle. The differences between appearances of vehicles from different types are relatively small. The detailed information of the feature map plays a very important role in distinguishing vehicles. Therefore, we propose to perform amplification operation to the feature maps and increase the discriminative ability of features for vehicles.

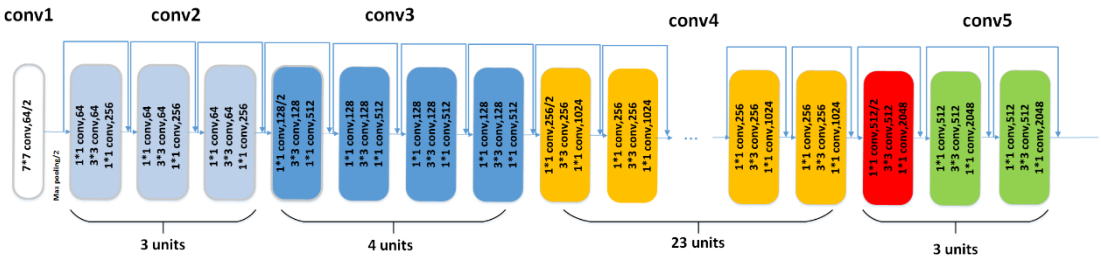


Figure 5. Structure of the Resnet101.

There are usually two main methods for upsampling feature map, interpolation and deconvolution. However, deconvolution usually produces checkboard artifacts, which is not conducive to the detailed description of features. Therefore, we adopt interpolation to enlarge the feature image. Here, we use bilinear interpolation to amplify the last feature map. The bilinear interpolation can be illustrated as follows.

Assume that the original feature map size is w ∗ h pixels and the enlarged feature map size is W ×H pixels. It is known that each pixel value in the original feature maps and enlarged feature maps is $f(x, y)$ and $f(X, Y)$ respectively. If you want to get the pixel value of the point $f(X, Y)$, you need to get

pixel values corresponding to the original feature map $f(x, y)$ according to the ratio of enlargement. As shown in Equation (1), if the calculated position is not an integer, you need to interpolate the pixels $f(\frac{w}{W} * X, \frac{h}{H} * Y)$ in the original image and assign them to the enlarged pixels $f(X, Y)$.

$$f(x, y) = f(\frac{w}{W} * X, \frac{h}{H} * Y) \tag{1}$$

As shown in Figure 6, the central idea of bilinear interpolation is to get the final pixel values to be interpolated by four adjacent points $f(i, j), f(i + 1, j), f(i, j + 1), f(i + 1, j + 1)$ next to the central pixel for linear interpolation in the vertical and horizontal directions. Suppose that the float coordinates of the pixel to be interpolated are $(i + u, j + v)$, where i, j are the integer part of the coordinate, and u, v are the decimal part of the coordinate whose range is [0, 1). Then the pixel value to be interpolated $f(i + u, j + v)$ can be determined by the corresponding values of the four surrounding pixels (i, j), $(i + 1, j)$, $(i, j + 1)$, $(i + 1, j + 1)$. The pixel value of the point to be interpolated is shown in Equation (2). Where $f(i, j)$ represents the pixel values of the location (i, j) in the original image.

$$f(i + u, j + v) = (1 - u)(1 - v)f(i, j) + (1 - u)vf(i, j + 1) + u(1 - v)f(i + 1, j) + uvf(i + 1, j + 1) \tag{2}$$

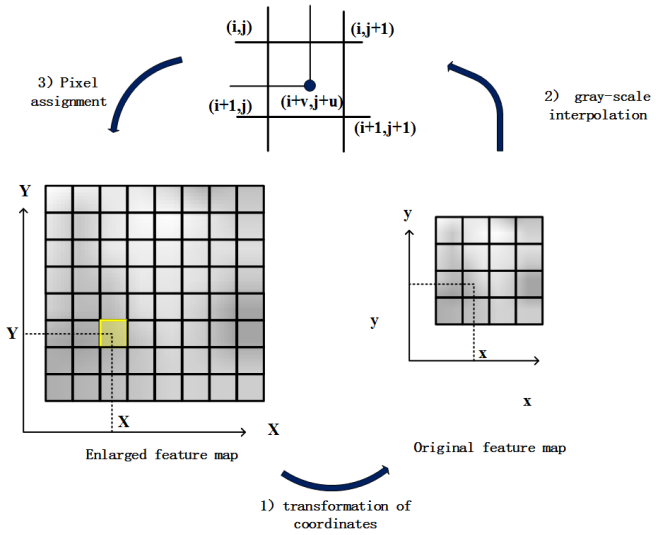


Figure 6. Flow diagram of bilinear interpolation to enlarge feature map.

3.4. Multi-Task Loss Function for Joint Horizontal and Oriented Bounding Boxes

Motivation of multi-task loss function. Multi-task loss function is aimed to detect horizontal and oriented vehicles simultaneously by combining the loss of horizontal bounding boxes with that of oriented bounding boxes. In addition, vehicle detection is a difficult problem due to diversity in object representation and small difference between different vehicles. Therefore, we also introduce center loss to the multi-task loss function to improve the discriminative ability of the features.

Traditional object detection methods usually use horizontal bounding boxes $(x_{\min}, y_{\min}, x_{\max}, y_{\max})$ to describe the position of the objects. However, vehicles on aerial images are usually with arbitrary orientation. For a oriented vehicle, we can describe the position more accurately by describing the coordinates of its four corners $(x_1, y_1, x_2, y_2, x_3, y_3, x_4, y_4)$.

As shown in Figure 7, when detecting an object that contains direction information, two types of anchors are usually used [39], namely horizontal anchors and oriented anchors. The horizontal anchor contains more contextual information around objects than oriented anchor in the ROI pooling, which can assist the object recognition. Therefore, the horizontal anchor rather than oriented anchor is adopted in the article.

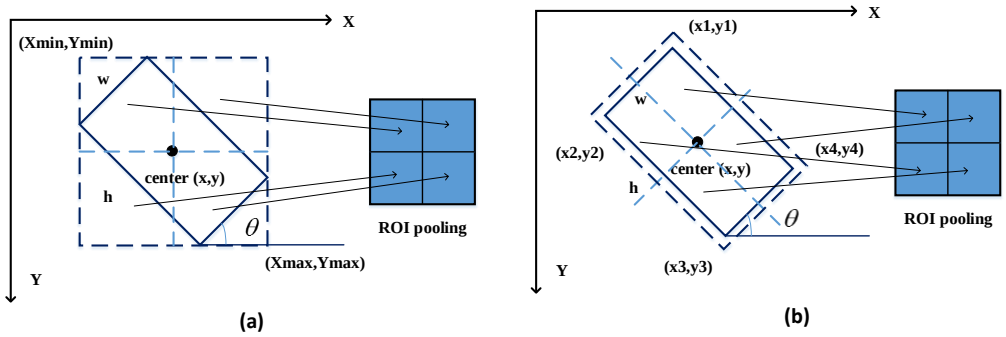


Figure 7. Two types of anchor setting methods of oriented objects. (**a**) Horizontal anchor. (**b**) Rotated anchor.

As shown in Equation (3), the proposed loss function consists of 5 parts, namely the cross-entropy loss of oriented objects $L_{cls}^{R}(P_R, P_R^*)$ in Equation (4), the cross-entropy loss of horizontal objects $L_{cls}^{H}(P_H, P_H^*)$ in Equation (5), the location loss function of the oriented objects $\sum_{i\in\{x,y,w,h,\theta\}} L_{reg}^{R}(t_i, t_i^*)$ and the horizontal objects $\sum_{i\in\{x,y,w,h\}} L_{reg}^{H}(t_i, t_i^*)$ in Equation (6) and center loss function $L_{Centerloss}$ in Equation (11). $\lambda_1, \lambda_2, \lambda_3$ are the balance parameters.

$$\begin{gathered} L\left(P_H, P_H^*, P_R, P_R^*, t, t^*\right) = L_{cls}^{H}(P_H, P_H^*) + L_{cls}^{R}(P_R, P_R^*) \\ +\lambda_1 \sum_{i\in\{x,y,w,h\}} L_{reg}^{H}(t_i, t_i^*) + \lambda_2 \sum_{i\in\{x,y,w,h,\theta\}} L_{reg}^{R}(t_i, t_i^*) + \lambda_3 L_{Centerloss} \end{gathered} \tag{3}$$

$$L_{cls}^{R}(P_R, P_R^*) = -\log(P_R) \tag{4}$$

$$L_{cls}^{H}(P_H, P_H^*) = -\log(P_H) \tag{5}$$

In terms of classification, P_H and P_R are the probabilities that predicted horizontal and oriented bounding boxes belong to each category respectively; P_H^* and P_R^* are true categories of the horizontal and oriented bounding boxes respectively.

In the process of location regression, we convert four corners to the (x, y, w, h, θ) in order to describe the position of oriented vehicles, where (x, y) represents the coordinates of vehicle center, (w, h) represents the width and height of vehicles and θ is the degrees from the horizontal perspective. For the horizontal bounding boxes, t^* represents the offset vector between true bounding boxes and positive anchors composed of x, y, w, h. For the oriented bounding boxes, t^* consists of x, y, w, h, θ. t represents the corresponding predicted coordinate vector. In Equations (7) and (8), x^*, x_a, x represent the true bounding box, anchor, and predicted box respectively. y, w, h, θ can be represented in a way similar to x.

$$L_{reg} = S_{L1}(t - t^*) = \begin{cases} if|t - t^*| < 1, 0.5 \cdot (t - t^*)^2 \\ otherwise, |t - t^*| - 0.5 \end{cases} \tag{6}$$

$$\begin{gathered} t_x^* = (x^* - x_a)/w_a, t_y^* = (y^* - y_a)/h_a \\ t_w^* = \log(w^*/w_a), t_h^* = \log(h^*/h_a), t_\theta^* = \theta^* - \theta_a \end{gathered} \tag{7}$$

$$\begin{gathered} t_x = (x - x_a)/w_a, t_y = (y - y_a)/h_a \\ t_w = \log(w/w_a), t_h = \log(h/h_a), t_\theta = \theta - \theta_a \end{gathered} \tag{8}$$

We introduce center loss as shown in Equation (9) to decrease within-class differences existing in features, and increase the ability of features in distinguishing diverse vehicles.

$$L_{Centerloss} = \frac{1}{2} \sum_{n=1}^{m} \left\| x_n - c_{y_n} \right\|_2^2 \tag{9}$$

where m is the batch size in the object classification stage, c_{y_n} is the feature center of the y-th category and x_n is the features before the fully connected layer.

4. Experimental Results and Setup

4.1. Description of Datasets and Implementation Details

4.1.1. Dataset Description

The experimental dataset in this paper is Vehicle Detection in Aerial Imagery (VEDAI). The VEDAI dataset contains aerial images with the size of 1024 × 1024 pixels extracted from the publicly available Utah AGRC database. The ground sampling distance (GSD) of the original image is 12.5 cm/pixel (cmpp). Each image consists of various types of small vehicles, backgrounds and objects that may lead to confusion. The VEDAI dataset contains 9 types of vehicles, namely van, tractor, pick-up, car, camping car, boat, plane and other vehicles. The research content of this paper is mainly for vehicle detection of nine categories. On average, there are 5.5 vehicles on each image, accounting for approximately 0.7% of the entire image. We split the dataset into two groups by randomly selecting 50% of the images for training while the left 50% for testing. The experiments have been repeated two times to reduce the measurement error.

The number of vehicles in the original training dataset for each category is shown in Table 2. The number of different vehicles in the training set is extremely imbalanced. Images before and after the proposed data augmentation method are shown in Figure 8. The red circle indicates the newly generated objects. The original images before data augmentation usually contain only a few types of vehicles. After performing the proposed data augmentation method, each image contains 9 different types of vehicles, and the vehicle position is randomly generated. The proposed method can help to increase the frequency of the categories with a smaller number of samples and increase the background diversity of vehicles to a certain extent.

Table 2. Vehicle number statistics in training dataset of VEDAI.

Class	Car	Pick-Up	Camping Car	Truck	Other	Tractor	Boat	Vans	Plane
Group1	700	449	209	156	103	85	90	48	21
Group2	740	501	181	144	97	85	80	52	26

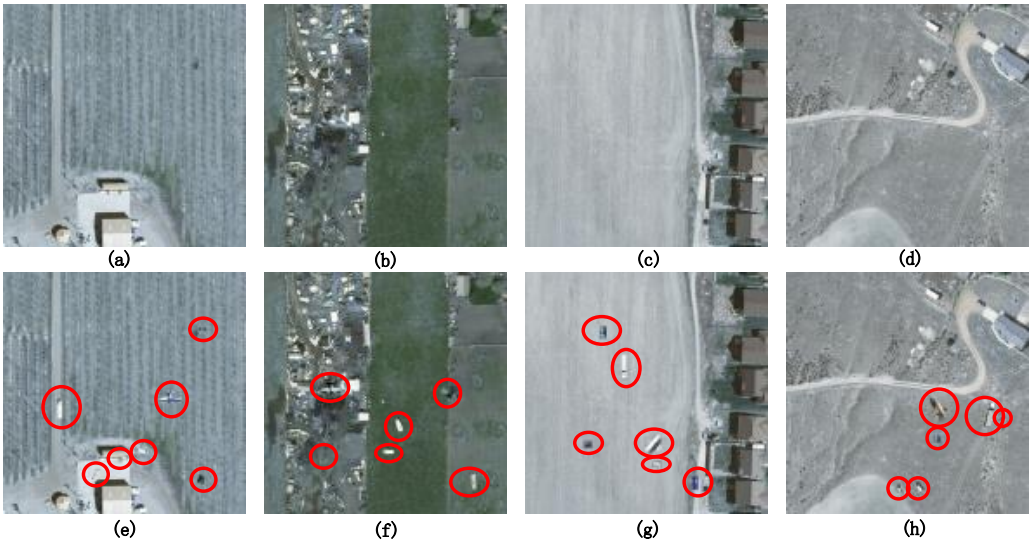


Figure 8. Comparison of images before and after augmentation by the oversampling and stitching method. (**a**–**d**) are original images from VEDAI dataset, and (**e**–**h**) are images synthesized by proposed method corresponding to (**a**–**e**).

4.1.2. Experimental Setup

Resnet101 which is pre-trained on the ImageNet dataset [40] is used for extracting features in this paper. The method in this paper is implemented in the TensorFlow framework with the Ubuntu 16.04 system. The computer hardware is GTX1080ti GPU with 11GB memory. The mini-batch size of the RPN stage and classification stage in this paper is 256 and 512. The initial learning rate of first 30,000 iterations is 0.003 while the learning rate of subsequent 70,000 epochs is 0.00003. The maximum iterations is set to 100,000. The momentum is 0.9 and the weight decay is 0.0001.

In the stage of RPN, we set the horizontal anchors with various shape and scale parameters, and set the thresholds of anchors and true objects to select positive and negative samples for training RPN networks. In this article, the anchor scale parameter is set to (8, 16, 32, 64, 128), and the shape parameter is set to (1, 1/2, 2/1, 1/3, 3/1, 1/4, 4/1, 1/5, 5/1, 1/6, 6/1, 1/7, 7/1). This article considers anchors with an IoU overlap below 0.3 as negative samples while that above 0.7 as positive samples. In training the RPN network, when the overlap between the ground truth and the anchors meets the above conditions, the anchors can be used to train the RPN network.

We have carried out comparative experiments with the baseline approaches that can detect oriented vehicles to demonstrate the effectiveness of the proposed algorithm. Among them, baseline methods include Faster RCNN, Feature Pyramid Network (FPN) and Dense Feature Pyramid Network (DFPN).We adopt Rotated anchors (RA) and Horizontal Anchors (HA) in FPN and DFPN, respectively.

4.1.3. Evaluation Metric

Mean average precision (mAP) is a comprehensive quality evaluation metric used in this paper, representing the mean value of average precision in each type of vehicle. The higher mAP is, the better vehicle detection performance is. Equation (10) shows how the average precision in each vehicle type is calculated.

$$AP = \sum_{n} (R_n - R_{n-1}) P_n \tag{10}$$

where R_n and P_n represent the recall ratio and precision ratio when n-th threshold is set. The precision ratio and recall ratio can be defined as Equations (11) and (12).

$$Precision = \frac{TP}{TP + FP} \tag{11}$$

$$Recall = \frac{TP}{TP + FN} \tag{12}$$

where FP and TP are the amount of wrongly and accurately detected vehicles. FN represents the amount of undetected vehicles. If the Intersection over Union (IoU) [40] between true locations and predicted locations computed by Equation (13) is above 0.5, the bounding box will be considered as TP. Otherwise, it will be FP.

$$IoU = \frac{area(predicted) \cap area(truth)}{area(predicted) \cup area(truth)} \tag{13}$$

4.2. Detection Results and Compared with Baseline Methods

Figure 9 shows the examples of the detection results by the proposed method. The proposed method can well detect diverse types of vehicles. While determining the specific categories of the vehicles, the proposed method can simultaneously obtain good detection results of the horizontal and the oriented bounding boxes.

Figure 9. Examples of the detection results by the proposed method. The odd rows are the detection results of the horizontal bounding boxes (HBB). The even rows are the detection results of the oriented bounding boxes (OBB).

Table 3 shows the horizontal and oriented detection results of baseline approaches. All baseline approaches are performed on the merged dataset after rotation augmentation and oversampling augmentation proposed in this paper. The detection results of the proposed framework are better than those of Faster RCNN algorithm, DFPN algorithm and FPN algorithm in both horizontal and oriented bounding boxes. Rotated anchors contain less context information compared with horizontal anchors,

so the detection accuracy of the FPN and DFPN with rotated anchors are obviously lower than that of the methods with horizontal anchors. The proposed framework with horizontal anchors is improved on the Faster RCNN method. The enlarged feature maps can increase the discriminative ability of features by restoring the detailed information of feature maps. Center loss can relatively increase the gap between features of different vehicle types by reducing the intra-class diversity existing in features belonging to the same vehicle type, which may lead to decreased misclassification of similar vehicle types. Therefore, the overall detection accuracy of the proposed approach is higher than that of Faster RCNN algorithm.

Table 3. Horizontal and oriented detection results of comparison approaches.

Method		Car	Pick-Up	Camping Car	Truck	Other	Tractor	Boat	Vans	Plane	mAP
Faster RCNN-HA	H	74.9	69.8	56.4	58.9	41.9	51.2	35.7	30.7	87.5	56.3
	O	70.4	66.3	**54.2**	50.2	38.8	50.1	36.5	29.5	87.9	53.8
FPN-RA	H	52.8	54.5	48.4	43.8	47.1	45.1	45.1	57.8	41.7	48.4
	O	54.7	56.1	51.4	43.8	38.4	46.4	48.3	57.8	41.7	48.7
DFPN-RA	H	56.6	53.0	43.7	31.3	41.3	51.8	53.8	55.8	61.1	49.8
	O	56.6	54.0	45.0	32.1	37.6	51.2	**57.1**	55.8	61.1	50.1
FPN-HA	H	69.7	63.9	**58.2**	48.7	**52.8**	**57.2**	53.8	**67.7**	52.1	58.2
	O	51.9	51.1	48.3	31.2	42.6	52.0	45.4	**71.1**	56.3	49.5
DFPN-HA	H	67.9	65.1	54.0	52.4	51.0	**57.2**	**56.7**	63.7	55.3	58.1
	O	54.1	52.2	44.6	33.4	41.7	42.3	46.1	51.9	60.6	47.4
Proposed method	H	**76.2**	**72.3**	50.9	**62.4**	41.9	52.9	52.3	38.6	**95.1**	**60.7**
	O	**76.5**	**71.9**	51.0	**61.2**	**44.7**	**52.7**	52.1	36.8	**95.1**	**60.4**

FPN method establishes a feature pyramid to detect multi-scale objects. The FPN method with horizontal anchors achieves the highest detection accuracy in the camping car, tractor and vans since it selects features suitable for detecting a certain type of vehicles from multilayer feature map pyramids. However, the detection results of other categories are unsatisfactory, especially for the airplane. That is because airplane is the category with the smallest number of samples in the original dataset, which limits the diversity of airplane samples. FPN increases the number of training parameters to build the feature pyramid and requires a larger number of samples, which may decrease the vehicle detection performance.

Compared with FPN, DFPN can make full use of multi-layer features to build a tighter feature pyramid between different feature layers, which makes the network better adapt to multi-scale objects. Therefore, the detection accuracy of DFPN with the rotated anchors is slightly higher than that of FPN. The detection accuracy of DFPN with the horizontal anchors are comparable with that of FPN with the horizontal anchors.

The same problem exists in the DFPN method, dense feature pyramid may increase number of network training parameters, which may lead to unideal detection accuracy of most vehicle categories. The horizontal detection results of FPN and DFPN are better than that of Faster RCNN because of multi-layer features. However, there is a big gap between results of detecting horizontal bounding boxes and those of detecting oriented bounding boxes in both DFPN and FPN methods with horizontal anchors. That is because it is difficult for both DFPN and FPN methods to regress the direction of oriented objects. But FPN can better obtain the direction information compared with DFPN. The oriented results of DFPN with horizontal anchors is about 2% mAP less than that of FPN.

In the proposed method, a multi-task loss function that introduces center loss is used to constrain the detection results of both the oriented bounding box and the horizontal bounding box. The accuracy difference between the horizontal bounding box and the oriented bounding box by the proposed method is small, and the overall accuracy of the proposed method is better than other methods.

4.3. Comparasion of the Different Data Augmentaion Methods

We compare different data augmentation approaches on the detection results in this section. We adopt three kinds of datasets for experiments. The first one is the dataset after rotation augmentation.

The second is the dataset after the proposed oversampling and stitching augmentation, and the third is combination of the previous two datasets. Three types of datasets are denoted as R, O and M respectively. To better prove the role of the proposed oversampling and stitching data augmentation algorithm, the proposed vehicle detection method in this paper along with the original Faster RCNN method are both verified on these three types of datasets.

The detection results of horizontal and oriented bounding box in the Faster RCNN algorithm for three training datasets are shown in Tables 4 and 5, respectively. The bold indicates higher accuracies of average precision and recall ratio in the R and O datasets. The underlined indicates the optimal average precision and recall ratios in the R, O and M datasets.

Table 4. Detection accuracy of Faster RCNN algorithm for HBB in three different types of datasets.

Class		Car	Pick-Up	Camping Car	Truck	Other	Tractor	Boat	Vans	Plane	Mean
R	Recall	**85.1**	**75.7**	**88.8**	68.2	64.0	**76.2**	65.4	52.8	85.2	73.5
	AP	71.8	**68.9**	**52.1**	50.1	**45.6**	40.3	27.5	**36.7**	80.1	52.6
O	Recall	84.3	**75.7**	87.8	**77.5**	**68.0**	**76.2**	**75.3**	**54.7**	**92.6**	**76.9**
	AP	**72.9**	67.1	48.2	**54.3**	44.9	**44.1**	**41.7**	32.5	**89.6**	**55.0**
M	Recall	84.5	79.5	92.0	76.8	65.0	80.0	67.9	58.5	92.6	77.4
	AP	74.9	69.8	56.4	58.9	41.9	51.2	35.7	30.7	87.5	56.3

Table 5. Detection accuracy of Faster RCNN algorithm for OBB in three different types of datasets.

Class		Car	Pick-Up	Camping Car	Truck	Other	Tractor	Boat	Vans	Plane	Mean
R	Recall	**82.2**	70.0	**86.2**	61.6	62.0	76.2	63.0	52.8	85.1	71.0
	AP	68.0	**65.5**	**47.9**	42.0	43.3	41.4	26.1	**37.2**	80.4	50.2
O	Recall	81.9	**74.0**	85.1	**75.5**	**66.0**	**76.2**	**69.1**	**58.5**	**92.6**	**75.4**
	AP	**69.8**	64.7	44.3	**53.3**	**43.9**	**44.2**	**37.5**	31.7	**89.7**	**53.2**
M	Recall	81.5	76.9	90.4	70.9	59.0	77.1	66.7	58.5	92.6	74.8
	AP	70.4	66.3	54.2	50.2	38.8	50.1	36.5	29.5	87.9	53.8

Among them, the mAP of the merged dataset reached the highest in horizontal and oriented bounding boxes, with 56.3% and 53.8%, respectively, which is better than R and O datasets. The mAP of O dataset reached 55% and 53.2% in the horizontal and oriented bounding box, which is better than the 52.6% mAP and 50.2% mAP for R dataset.

As can be seen in Tables 4 and 5, the rotation augmentation method does not address the uneven vehicle number distribution in the dataset. The training network still favors vehicle categories with a large number of samples. After the proposed augmentation of the oversampling and stitching, the frequency of vehicle categories with fewer training samples in the dataset is increased. Therefore, the network trained by the dataset after oversampling and stitching data augmentation can better distinguish diverse types of vehicles.

As shown in Tables 4 and 5, after performing the proposed data augmentation method, the recall ratios of the Pickup, Truck, Other, Tractor, boat, Vans, and Plane categories have been improved, and the corresponding mAP has also been improved. The Car, Pick-up, and Camping car are the categories with more training samples in the rotated training dataset. As the number of vehicles in other categories has been increased after the proposed data augmentation method, the tendency of network to categories with a larger number of training samples has been reduced. Therefore, the recall ratio and average precision of these categories are decreased slightly and that of remaining categories are increased after proposed data augmentation method.

Although the dataset based on oversampling and stitching augmentation method can increase accuracy of vehicle detection to some degree. This type of synthesis method increases the complexity of the background objects. Moreover, a certain gap exists between the synthetic and real imagery. In order to decrease the negative influence of background diversity on object detection, we combine O dataset with R dataset to form a new training dataset M. According to the results of the merged

dataset, we find that the results of the merged dataset are better than those of the previous two datasets. The detection accuracies of car, pickup, camping car, tractor and so on have been further improved compared to those of the oversampling and stitching data augmentation method. That is because the merged dataset improves the network's ability to adapt to the complex background of the vehicles.

Considering the shortcomings of feature representations for vehicles and limited discrimination ability of features for different categories in Faster RCNN, we propose an oriented vehicle detection method based on feature map amplification. We also use the above three datasets for training, and we can get similar conclusions as the Faster RCNN method. As shown in Tables 6 and 7, the proposed oversampling and stitching data augmentation method improves the recall ratios of vehicles with a small number of training samples such as Truck, Tractor, Boat, Vans, and Plane, and effectively improves the corresponding average detection accuracy. The merged datasets for training the proposed network can further improve the average vehicle detection accuracy.

Table 6. Detection accuracy of proposed algorithm for HBB in three different types of datasets.

Class		Car	Pick-Up	Camping Car	Truck	Other	Tractor	Boat	Vans	Plane	Mean
R	Recall	85.8	**78.7**	**90.4**	74.8	**73.0**	**79.0**	67.9	58.5	88.9	77.4
	AP	**75.3**	**70.5**	**52.9**	53.5	**50.8**	49.1	**43.0**	38.1	86.4	57.7
O	Recall	**86.3**	77.7	88.3	**80.1**	66.0	75.2	**69.1**	**62.3**	**96.3**	**77.9**
	AP	74.0	68.5	52.4	**62.2**	44.0	**51.9**	38.4	**40.5**	**93.7**	**58.3**
M	Recall	84.9	79.9	90.4	81.0	64.0	83.8	74.1	62.3	96.3	79.5
	AP	76.2	72.3	50.9	62.4	45.5	52.9	52.3	38.6	95.1	60.7

Table 7. Detection accuracy of proposed algorithm for OBB in three different types of datasets.

Class		Car	Pick-Up	Camping Car	Truck	Other	Tractor	Boat	Vans	Plane	Mean
R	Recall	85.7	**78.3**	**90.4**	73.5	**72.0**	**79.0**	66.7	58.5	88.9	77.0
	AP	**75.2**	**70.2**	**51.7**	52.3	**50.3**	49.2	**39.5**	37.7	84.7	56.8
O	Recall	**86.1**	77.3	88.3	**78.8**	66.0	76.2	**69.1**	**60.4**	**96.3**	**77.6**
	AP	73.7	68.0	50.6	**60.3**	43.0	**52.5**	38.4	**39.3**	**93.9**	**57.7**
M	Recall	84.9	79.5	89.9	79.5	66.0	83.8	74.1	62.3	96.3	79.6
	AP	76.5	71.9	51.0	61.2	44.7	52.7	51.2	38.6	95.1	60.4

4.4. Ablation Study

We performed ablation studies so as to demonstrate the role of each contribution in this paper. Tables 8 and 9 show the mean accuracy of detecting horizontal and oriented bounding boxes from two groups respectively. The bold represents the optimal recall or precision ratio in each column. We take the original Faster RCNN method as the basic method. In order to ensure fair comparison, all parameters in different methods are consistent with those in Section 4.1.2.

Table 8. The mean of ablation experimental results of HBB from two groups.

Class		Car	Pick-Up	Camping Car	Truck	Other	Tractor	Boat	Vans	Plane	mAP
Rotation Dataset	Recall	84.5	80.2	91.3	72.6	66.5	75.2	69.4	60.8	78.3	52.5
	AP	70.6	68.2	59.9	51.8	41.1	39.0	28.2	40.2	73.8	
Rotation Dataset + amplification	Recall	85.2	80.3	**91.9**	73.9	65.4	77.4	73.1	**64.6**	75.7	55.6
	AP	72.4	70.2	**57.4**	54.0	41.3	46.5	40.5	43.9	74.6	
Merge Dataset + amplification	Recall	**85.8**	79.3	91.8	78.6	**72.9**	80.5	76.2	65.7	92.9	58.6
	AP	**76.7**	69.8	55.3	59.6	**41.2**	**48.4**	42.0	46.7	87.9	
Proposed method	Recall	85.3	**82.4**	91.2	**81.1**	67.9	**81.3**	**79.9**	64.5	**93.4**	**60.4**
	AP	75.7	**72.0**	56.4	**61.7**	**41.9**	48.0	**52.8**	**46.4**	**88.8**	

According to the vehicle detection results, the bilinear interpolation feature amplification method improves the ability of the features to distinguish different vehicle categories by keeping more detailed information hidden in CNN features. The horizontal and oriented vehicle detection results are increase by 3% mAP compared to the original Faster RCNN framework. We replace the rotation dataset with

the merge dataset generated in this paper. The accuracy of horizontal and oriented vehicle detection are improved by 3% mAP and 5% mAP respectively. The center loss function improves the detection results by 2% mAP since it can reduce the intra-class feature differences of the different categories. The proposed framework achieve a mean average precision of 60.4% and 60.1% in detecting horizontal and oriented bounding boxes respectively. All the improvements in this article have achieved accuracy improvements of approximately 8% mAP totally in detecting horizontal and oriented bounding boxes, respectively.

Table 9. The mean of ablation experimental results of OBB from two groups.

Class		Car	Pick-Up	Camping Car	Truck	Other	Tractor	Boat	Vans	Plane	mAP
Rotation Dataset	Recall	83.0	77.0	90.3	69.0	65.0	75.8	68.2	58.7	78.3	50.9
	AP	68.2	66.3	57.6	46.9	38.6	39.5	27.2	40.1	74.0	
Rotation Dataset + amplification	Recall	83.8	78.5	88.7	69.0	63.9	76.0	74.2	64.6	75.7	53.2
	AP	70.4	67.3	53.1	47.3	38.0	44.2	40.3	43.9	74.5	
Merge Dataset + amplification	Recall	85.0	79.1	**91.2**	78.9	**73.4**	80.5	75.1	**67.7**	90.5	58.2
	AP	**76.4**	69.6	55.2	59.2	**40.9**	**48.3**	41.4	**46.8**	86.3	
Proposed method	Recall	**85.2**	**82.1**	90.7	**80.8**	68.5	**80.7**	**78.7**	64.5	**93.4**	**60.1**
	AP	75.6	**71.7**	**57.0**	**60.3**	40.4	48.0	**52.4**	46.4	**88.8**	

In order to verify the role of each component in the multi-task loss, this paper conducts ablation study on the loss function as shown in Table 10. The experimental results show that center loss can improve the detection accuracy of both the oriented bounding box and the horizontal bounding box at the same time.

Table 10. Ablation Experimental Results of Multi-task loss.

Cls_R	Reg_R	Cls_H	Reg_H	Center_Loss	mAP(%)	
					H	O
√	√	√	√	√	60.7	60.4
√	√	√	√		58.4	58.0
		√	√	√	62.4	-
		√	√		62.2	-
√	√			√	-	56.1
√	√				-	54.4

Meanwhile, we separately train the loss function of oriented bounding box and the horizontal bounding box. When only using the loss function of the horizontal bounding box, the horizontal results has a higher detection accuracy, and center loss has a little effect on the detection accuracy. However, the horizontal bounding box cannot describe the direction information of the vehicles, and the position is relatively rough.

When only using the oriented bounding box loss function, the oriented detection accuracy is poor, and center loss can improve the detection accuracy by about 2% mAP. The joint training can improve the detection accuracy of the oriented bounding box. The joint training loss function that incorporates center loss can simultaneously obtain ideal detection results for both horizontal and oriented bounding boxes.

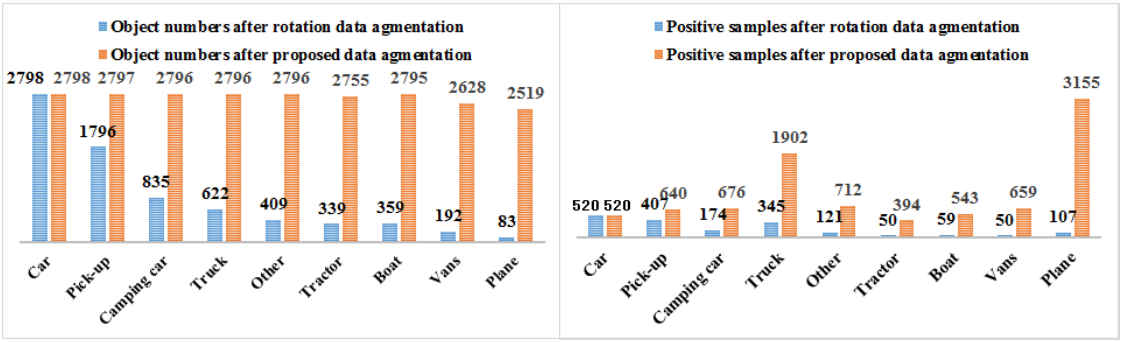
5. Discussion

5.1. Analysis of Object Number and Positive Samples

The small size of the vehicles, the center deviation caused by pooling operation, and the anchor parameters make detection of small objects difficult. The category imbalance between vehicles in the

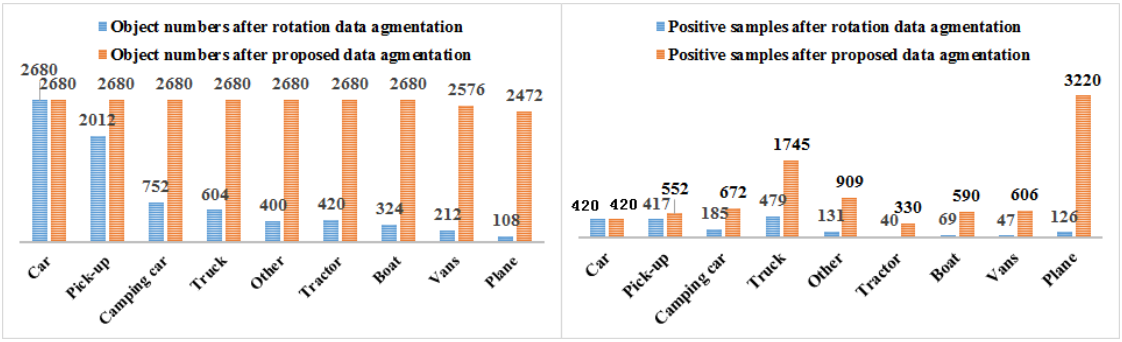
dataset will further have a negative impact on training the network. In this section, we discuss the proposed data augmentation methods on the number of objects and positive samples.

In order to increase the angle diversity of trained objects, rotation augmentation is performed. As shown in Figure 10a, the number of vehicles of Group 1 varies from 83 to 2798. And as shown in Figure 11a, the number of vehicles of Group 2 varies from 108 to 2680. The rotation augmentation cannot solve the category imbalance, but increase the difference between the numbers of different classes.



(a) (b)

Figure 10. Comparison between rotation augmentation and proposed method of Group 1. (**a**) Object numbers (**b**) Positive samples (IOU > 0.7).



(a) (b)

Figure 11. Comparison between rotation augmentation and proposed method of Group 2. (**a**) Object numbers (**b**) Positive samples (IOU > 0.7).

The proposed the oversampling and stitching data augmentation method can make the number of each type of vehicle relatively equal. We take the most numerous type of the car as an expansion benchmark and increase the frequency of remaining 8 vehicle categories in different background images. As shown in Figure 10a, the number of vehicles after proposed method ranges from 2519–2798 in Group 1. Similarly, as shown in Figure 11a, the number of vehicles ranges from 2472–2680 in Group 2. We find that the vehicle number has been balanced after the proposed data augmentation method.

In the region proposal stage, the horizontal anchors are used to selective the samples for the network. For small vehicles that are difficult to invest in network training, increasing the number of samples is conducive to train the model. We separately count the number of positive samples (IoU > 0.7) extracted from R and O dataset on the last feature map extracted by resnet101 so as to prove the role of the proposed data augmentation method. The statistical results are shown in Figures 10b and 11b. The proposed data augmentation approach can effectively increase the amount of positive samples used for training RPN network and increase the diversity of training samples. Therefore, the proposed data augmentation method is an effective method to improve the effective sample number and the ability of the network to detect small objects.

5.2. Analysis of the Feature Amplification Parameters

In order to prove that vehicle detection results can be enhanced by feature map magnification operation, we use the merged dataset to discuss the magnification operation of deep feature maps and analyze the impact of different amplification parameters on vehicle detection. Two interpolation methods including bilinear interpolation and nearest neighbor (NN) interpolation are for comparison. H and O are respectively denoted as the horizontal and oriented detection results.

The bold represents the optimal average precision of horizontal or oriented bounding boxes in each line of Table 11. The detection accuracy of horizontal and oriented vehicles by the proposed method without feature amplification are 58.9% mAP and 58.3% mAP, respectively. The experimental results show that the enlarged feature map by the bilinear interpolation can improve the detection accuracy to a certain extent.

Table 11. Detection accuracy of different parameters in the feature amplification.

Category	no		Bilinear-1.5		Bilinear-2		Bilinear-2.5		Bilinear-3.0		NN-2.0	
	H	**O**	**H**	**O**	**H**	**O**	**H**	**O**	**H**	**O**	**H**	**O**
vans	29.6	29.2	36.9	35.5	38.6	38.6	38.8	39.1	**43.9**	**43.5**	32.8	31.4
plane	95.3	95.3	**97.9**	**97.4**	95.1	95.1	96.5	96.5	95.4	95.4	96.0	95.9
truck	**68.0**	**65.4**	63.3	62.6	62.4	61.2	61.8	60.6	61.9	60.7	57.7	55.9
boat	40.7	41.3	46.6	46.0	52.3	**52.1**	**53.0**	51.0	46.5	45.7	43.3	42.6
camping-car	48.9	48..0	45.6	45.8	50.9	51.0	49.1	49.3	**52.5**	**51.5**	49.0	48.2
car	76.0	75.3	75.5	75.3	76.2	**76.5**	74.4	74.3	**76.3**	76.2	71.2	70.7
others	**49.7**	**49.3**	46.5	46.0	45.5	44.7	42.8	42.1	46.1	45.7	44.8	44.2
Tractor	**54.0**	**53.4**	49.1	49.8	52.9	52.7	50.5	50.2	50.6	50.9	51.6	51.7
pickup	67.8	67.6	69.5	69.0	72.3	**71.9**	71.8	71.5	**72.5**	71.8	68.8	68.5
mean	58.9	58.3	59.0	58.6	**60.7**	**60.4**	59.9	59.4	60.6	60.2	57.2	56.6

When bilinear interpolation is used to enlarge the feature map, we investigate the influence of feature maps with different magnifications on detection results. Among the comparison experiments with amplification multiples of 1.5, 2.0, 2.5, and 3.0, feature amplification by bilinear interpolation with multiples of 2.0 increases the most detection accuracy, with about 2% mAP for both horizontal and oriented vehicles.

At the same time, we use the nearest neighbor interpolation method of 2.0 to perform amplification experiments on feature maps. The nearest neighbor amplification method shows lower accuracy than of the method without feature map amplification since it will cause a jagged effect on the enlarged feature map, which is not beneficial to representing truck, car, others, and tractor.

6. Conclusions

Vehicle detection is difficult for remote sensing images because of the limited size and class imbalance. To enhance the vehicle detection results, we propose an oriented vehicle detection method for aerial images consisting of three indispensable parts, namely oversampling and stitching data augmentation method, amplifying the feature map and a joint training loss function for horizontal and oriented bounding boxes with the center loss. Three parts are aimed to address the problem of foreground-foreground category imbalance, the reduced discriminative ability of features caused by pooling operation and small inter-class diversity between types of oriented vehicles respectively. The experiments on the VEDAI dataset can draw the following conclusions.

(1) The proposed framework outperforms most of previous vehicle detection approaches. The method proposed in this paper can effectively detect oriented vehicles with a 8% higher mAP than the original Faster RCNN approach.

(2) The proposed oversampling and stitching data augmentation method is an effective way to address class imbalance. The datasets combining oversampling and stitching data augmentation

with rotation augmentation can improve about 3% mAP since they can increase the number of effective samples in the network and reduce the imbalance between vehicle categories to a certain extent.

(3) The amplified feature map makes the network better distinguish different categories of vehicles by about 3% mAP.

(4) The multi-task loss function can get the horizontal and oriented detection results simultaneously and the center loss can improve the accuracy since it can reduce the intra-class diversity of the vehicle categories to a certain extent.

Although the mAP can be improved by the proposed method compared with the Faster RCNN method, the overall precision ratio is still low. That is mainly because a large number of background objects are wrongly detected as foreground objects. In the future, we will study how to increase the ability of features to distinguish small objects.

Author Contributions: Conceptualization, N.M.; methodology, N.M.; writing—original draft preparation, N.M.; writing—review and editing, N.M.; supervision, L.Y.; funding acquisition, L.Y. 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 Program of China under grant no. 2017YFC0803801.

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

References

1. Sun, Z.; Bebis, G.; Miller, R. On-road vehicle detection: A review. *IEEE Trans. Pattern Anal. Mach. Intell.* **2006**, *28*, 694–711. [CrossRef]
2. Fan, Q.F.; Brown, L.; Smith, J. A closer look at Faster R-CNN for vehicle detection. *IEEE Intell. Veh. Symp. (IV)* **2016**, 124–129. [CrossRef]
3. Dai, X. Hybridnet: A fast vehicle detection system for autonomous driving. *Signal Process Image Commun.* **2019**, *70*, 79–88. [CrossRef]
4. Gleason, J.; Nefian, A.V.; Bouyssounousse, X.; Fong, T.; Bebis, G. Vehicle detection from aerial imagery. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), Shanghai, China, 9–13 May 2011; pp. 2065–2070.
5. Fang, J.; Zhou, Y.; Yu, Y.; Du, S. Fine-grained vehicle model recognition using a coarse-to-fine convolutional neural network architecture. *IEEE Trans. Intell. Transp. Syst.* **2016**, *18*, 1782–1792. [CrossRef]
6. Razakarivony, S.; Jurie, F. Vehicle detection in aerial imagery: A small target detection benchmark. *J. Vis. Commun. Image Represent* **2016**, *34*, 187–203. [CrossRef]
7. Cheng, G.; Han, J. A survey on object detection in optical remote sensing images. *ISPRS J. Photogram. Remote Sens.* **2016**, *117*, 11–28. [CrossRef]
8. Viola, P.; Jones, M. Rapid object detection using a boosted cascade of simple features. In Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Kauai, HI, USA, 8–14 December 2001; pp. 511–518. [CrossRef]
9. Csurka, G.; Dance, C.; Fan, L.; Willamowski, J.; Bray, C. Visual categorization with bags of keypoints. In Proceedings of the Workshop on Statistical Learning in Computer Vision, ECCV, Prague, Czech Republic, 10–16 May 2004; pp. 1–2.
10. Felzenszwalb, P.; McAllester, D.; Ramanan, D. A discriminatively trained, multiscale, deformable part model. In Proceedings of the 2008 IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, AK, USA, 23–28 June 2008; pp. 1–8.
11. Dalal, N.; Triggs, B. Histograms of oriented gradients for human detection. In Proceedings of the 2005 IEEE Computer Vision and Pattern Recognition, San Diego, CA, USA, 20–25 June 2005; pp. 886–893.
12. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; Berg, A.C. SSD: Single shot multibox detector. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 8–16 October 2016; pp. 21–37.

13. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
14. Redmon, J.; Farhadi, A. YOLO9000: Better, faster, stronger. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7263–7271.
15. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
16. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. YOLOv4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934.
17. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
18. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial pyramid pooling in deep convolutional networks for visual29recognition. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; pp. 346–361.
19. Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 13–16 December 2015; pp. 1440–1448.
20. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 7–12 December 2015; pp. 91–99.
21. Oksuz, K.; Cam, B.C.; Kalkan, S.; Akbas, E. Imbalance problems in object detection: A review. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**. [CrossRef]
22. Ouyang, W.; Wang, X.; Zhang, C.; Yang, X. Factors in finetuning deep model for object detection with long-tail distribution. In Proceedings of the IEEE Conference on Computer Vision & Pattern Recognition, Las Vegas, NV, USA, 26–30 June 2016; pp. 864–873.
23. Oksuz, K.; Cam, B.C.; Akbas, E.; Kalkan, S. Generating positive bounding boxes for balanced training of object detectors. In Proceedings of the IEEE Winter Conference on Applications of Computer Vision, Snowmass village, CO, USA, 1°C5 March 2020; pp. 894–903.
24. Wang, H.; Wang, Q.; Yang, F.; Zhang, W.; Zuo, W. Data augmentation for object detection via progressive and selective instance-switching. *arXiv* **2019**, arXiv:1906.00358.
25. Ji, H.; Gao, Z.; Mei, T.; Ramesh, B. Vehicle detection in remote sensing images leveraging on simultaneous super-resolution. *IEEE Geosci. Remote Sens. Lett.* **2020**, *17*, 676–680. [CrossRef]
26. Singh, B.; Davis, L.S. An analysis of scale invariance in object detection snip. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 3578–3587.
27. Mostofa, M.; Ferdous, S.N.; Riggan, B.S.; Nasrabadi, N.M. Joint-Srvdnet: Joint super resolution and vehicle detection network. *IEEE Access* **2020**, *8*, 82306–82319. [CrossRef]
28. Tayara, H.; Soo, K.G.; Chong, K.T. Vehicle detection and counting in high-resolution aerial images using convolutional regression neural network. *IEEE Access* **2018**, *6*, 2220–2230. [CrossRef]
29. 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]
30. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
31. Li, K.; Cheng, G.; Bu, S.; You, X. Rotation-insensitive and context-augmented object detection in remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 2337–2348. [CrossRef]
32. Mo, N.; Yan, L.; Zhu, R.; Xie, H. Class-specific anchor based and context-guided multi-class object detection in high resolution remote sensing imagery with a convolutional neural network. *Remote Sens.* **2019**, *11*, 272. [CrossRef]
33. Deng, Z.; Sun, H.; Zhao, J.; Lei, L.; Zou, H.; Zhou, S. Multi-scale object detection in remote sensing imagery with convolutional neural networks. *ISPRS J. Photogram. Remote Sens.* **2017**, *145*, 3–22. [CrossRef]
34. Wen, Y.; Zhang, K.; Li, Z.; Qiao, Y. A discriminative feature learning approach for deep face recognition. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 499–515.

35. Ma, J.; Shao, W.; Ye, H.; Wang, L.; Wang, H.; Zheng, Y.; Xue, X. Arbitrary-oriented scene text detection via rotation proposals. *IEEE Trans. Multimedia* **2018**, *20*, 3111–3122. [CrossRef]
36. Yang, X.; Sun, H.; Fu, K.; Yang, J.; Sun, X.; Yan, M.; Guo, Z. Automatic ship detection in remote sensing images from google earth of complex scenes based on multiscale rotation dense feature pyramid networks. *Remote Sens.* **2018**, *10*, 132. [CrossRef]
37. Ding, J.; Xue, N.; Long, Y.; Xia, G.S.; Lu, Q. Learning roi transformer for detecting oriented objects in aerial images. *arXiv* **2018**, arXiv:1812.00155.
38. He, K.; Zhang, X.; Ren, S. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision & Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
39. Yang, X.; Sun, H.; Sun, X.; Yan, M.; Guo, Z.; Fu, K. Position detection and direction prediction for arbitrary-oriented ships via multitask rotation region convolutional neural network. *IEEE Access* **2018**, *6*, 50839–50849. [CrossRef]
40. Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; Fei-Fei, L. ImageNet: A large-scale hierarchical image database. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Miami, FL, USA, 20–25 June 2009; pp. 248–255.