ORIENTED VEHICLE DETECTION IN HIGH-RESOLUTION REMOTE SENSING IMAGES BASED ON FEATURE AMPLIFICATION AND CATEGORY BALANCE BY OVERSAMPLING DATA AUGMENTATION.

Nan Mo , Li Yan *

School of geodesy and geomatics, Wuhan University, China-nmo@whu.edu.cn, lyan@sgg.whu.edu.cn

Commission III, WG III/1

KEY WORDS: Oriented Vehicle Detection, Oversampling Data Augmentation, Feature Amplification, Center Loss

ABSTRACT:

Vehicles usually lack detailed information and are difficult to be trained on the high-resolution remote sensing images because of small size. In addition, vehicles contain multiple fine-grained categories that are slightly different, randomly located and oriented. Therefore, it is difficult to locate and identify these fine categories of vehicles. Considering the above problems in high-resolution remote sensing images, this paper proposes an oriented vehicle detection approach. First of all, we propose an oversampling and stitching method to augment the training dataset by increasing the frequency of objects with fewer training samples in order to balance the number of objects in each fine-grained vehicle category. Then considering the effect of the pooling operations on representing small objects, we propose to improve the resolution of feature maps so that detailed information hidden in feature maps can be enriched and they can better distinguish the fine-grained vehicle categories. Finally, we design a joint training loss function for horizontal and oriented bounding boxes with center loss, to decrease the impact of small between-class diversity on vehicle detection. Experimental verification is performed on the VEDAI dataset consisting of 9 fine-grained vehicle categories so as to evaluate the proposed framework. The experimental results show that the proposed framework performs better than most of competitive approaches in terms of a mean average precision of 60.7% and 60.4% in detecting horizontal and oriented bounding boxes respectively.

1. INTRODUCTION

Small object detection is attracting increasing attention due to the advancement of high-resolution remote sensing images. Real-time detection of vehicles is important in the tasks of autonomous driving and traffic monitoring recently (Sun et al., 2006). In some more specific tasks, we sometimes need to determine their types and orientations. The traffic conditions can be better scheduled and analysed by providing more accurate information about vehicles. Therefore, the oriented fine-grained detection of vehicles is of great research significance. However, the oriented fine-grained vehicle detection is a more challenging task compared with other multi-class object detection problems since it contains small objects.

Traditional remote sensing image vehicle detection methods usually include the following four steps: 1) Data preprocessing. It includes operations such as improving the image quality, improving the contrast between vehicles and backgrounds, and clustering. 2) Determination of the potential position of vehicles. For example, the contrast between different parts of images can be calculated to determine the potential position of vehicles. 3) Segmentation. It is performed to accurately extract potential vehicles from the background. 4) Recognition. Features are extracted from potential objects and the category of vehicles is finally determined by extracted features.

Recent remote sensing image vehicle detection methods are completely different from traditional methods since they try to decrease the influence of intermediate decisions on the detection results obtained by machine learning methods. The machine learning methods constitute deep learning based approaches and shallow features based approaches based on different types of extracted features (Cheng et al., 2016). Before 2012, shallow features based approaches are the mainstream algorithms for object detection. However, shallow features including Viola Jones Detectors (Viola et al., 2001), Deformable Parts Model (DPM) (Felzenszwalb et al., 2008) and Histogram of Oriented Gradients (HOG) (Dalal et al., 2005) usually deliver poor performance in representing vehicles because they lack semantic information which are important for recognizing objects.

Convolutional Neural Networks (CNN) can automatically learn semantic features and perform well in representing objects. The development of vehicle detection approaches has been promoted since deep learning architectures emerged in 2012. Vehicle detection approaches based on deep learning architectures consist of one-stage vehicle detection approaches such as Single Shot Multi-Box Detector (SSD) (Liu et al., 2016), You Only Look Once (YOLO) (Redmon et al., 2016), YOLOv2 (Redmon et al., 2017)，YOLOv3 (Redmon et al., 2018) and two-stage vehicle detection methods such as Region CNN (RCNN) (Van et al., 2011), SPPNet (He et al., 2015), Fast RCNN (Girshick et al., 2015), and Faster RCNN (Ren et al., 2015) according to different detection processes. Compared with two-stage vehicle detection approaches, one-stage approaches are with lower precision ratio and faster detection speed. The two-stage methods can achieve high precision ratio with a speed that can meet real-time requirements. Therefore,

* Corresponding author

this paper mainly investigates two-stage vehicle detection approaches.

Existing two-stage vehicle detection methods may suffer from class imbalance, the reduced discriminative ability caused by pooling operation, the arbitrary orientation of objects. The research status of these three problems is as follows.

In remote sensing images, vehicles are a kind of moving objects and usually distributed in geographic space at different frequencies. Therefore, there are usually uneven numbers of different fine-grained vehicle categories in the dataset. The imbalanced class distribution makes the network training favour the vehicle categories with a larger number, which makes it difficult to obtain ideal detection results for fine-grained vehicles. Ouyang et al.(2016) propose to fine-tune the distribution of under-represented categories by clustering similar categories to address the class imbalance problem. Oksuz et al. (2020) proposes an online foreground balanced sampling method, which decreases the imbalance between distributions of different objects within each batch by assigning probability to each true bounding box. Wang et al.(2019) proposed a sample exchange strategy to generate new samples and decrease the imbalance by exchanging the same type of objects in diverse images. However, the above methods are mainly aimed at the object imbalance problem in the natural imagery. Unlike natural imagery, remote sensing imageries are with a larger coverage area and more complex backgrounds. Therefore, the above class balance methods may deliver unsatisfactory performance when applied to the field of remote sensing.

Some researchers have carried out some works in order to improve the performance in detecting small objects. These methods mainly start from two aspects, one is to improve the resolution of the training images including small objects, and the other is to improve the detail information of the feature maps describing the small objects. In terms of improving image resolution, Ji et al.(2019) fuse the object detection network with the image super-resolution reconstruction network to increase the image size. Singh et al.(2018) propose to establish multi-scale pyramids for training images by resizing them. By increasing the image resolution, the discriminative ability of the features from different vehicles and the performance of locating and detecting vehicles can be optimized theoretically. In improving the resolution of the feature map, Tayara et al (2017). propose to perform deconvlution on the feature maps continuously to improve the discrimination of shallow features and retain the detailed information of feature maps. AVDNet (Mandal et al, 2019) is proposed to keep the detail information by increasing the spatial resolution of feature maps for vehicles and introducing ConvRes modules to difference scale layers. Lin et al. (2017) propose a layer-by-layer prediction using feature pyramids (FPN) to detect multi-scale objects. The advantage of this method is that the multi-scale feature map is an inherent transition module in the CNN, 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 amount and computational cost of training data and set high requirements for computers and graphics cards. Since the computers may not meet the requirements of the hardwares, these methods are not commonly used in practical applications. In addition, the above methods may lose the deep semantic information hidden in features while improving the feature resolution, and the discriminative ability to distinguish different vehicles is still limited.

The remote sensing image acquired by the sensor taken overhead. Therefore, the direction of the vehicle on the image is arbitrary. Traditional horizontal bounding boxes can only roughly describe the position of vehicles. The directions of vehicles can help to accurately locate the position of vehicles. Therefore, it is necessary to study oriented vehicle detection algorithms. In text detection, Ma et al.(2018) propose an oriented text detection algorithm to detect inclined text. Yang et al.(2018) propose a multi-oriented ship detection algorithm of remote sensing images. Ding et al. (2018) propose an oriented multi-class object detection method in aerial images. Oriented object detection methods are more and more important in the field of object detection, especially for fine-grained vehicle detection. It is necessary to accurately determine the vehicle's direction information.

An oriented vehicle detection framework based on feature amplification and oversampling data augmentation is proposed for high-resolution remote sensing images so as to address the problems mentioned above. This paper takes the two-stage object detection algorithm Faster RCNN as the research basis and makes improvements on this framework. Considering the distribution characteristics of vehicles in remote sensing images, we design an oversampling and stitching data augmentation method for remote sensing images so that the numbers of different vehicles in images are balanced for training model and the negative influence of category imbalance on training dataset can be reduced to some extent. Considering both the discriminative ability of features and computer hardware requirements, we explore semantic information hidden in deep feature maps and performs magnification operation for deep feature maps. By performing bilinear interpolation on feature maps, the detailed information can be enriched while maintaining the deep semantic information, which may improve the discrimination of the features in representing vehicles. We also designs a joint training loss function for horizontal and oriented bounding boxes with the center loss (Wen et al., 2016). The proposed method regresses the vehicle position and direction by setting the horizontal anchors, jointly training the horizontal bounding boxes and orientated bounding boxes. The method can simultaneously acquire the horizontal and oriented detection results to get more accurate position of the vehicles.

2. METHODOLOGY

2.1 The overall architecture

Figure 1 shows the overall architecture of the proposed approach, where the Resnet101(He et al., 2016) is used for extracting feature maps. The proposed approach three parts, 1) oversampling based data augmentation, 2) improving the size of the feature maps and 3) a joint training loss function for horizontal and oriented bounding boxes combined with center loss. The details of three steps are as follows.

First of all, we perform oversampling data augmentation on the training dataset. The frequency and location of the vehicles in the remote sensing images are random. The number of fine-grained vehicles in the training dataset is usually uneven. We perform stitching and oversampling augmentation on the training data by increasing the frequency of vehicles with fewer number of training data to synthesize a new dataset.

In the stage of region proposal network (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 truth. The oversampling augmentation method can improve the diversity and number of positive samples in the RPN stage.

In the stage of classification network, we perform the feature map magnification, to enhance the ability of feature maps to represent vehicles by increasing the size of deep feature maps. 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 for minimized within-class difference. The loss function can increase the ability of deep features to distinguish fine-grained vehicles, and obtain the accurate positions of the vehicles.

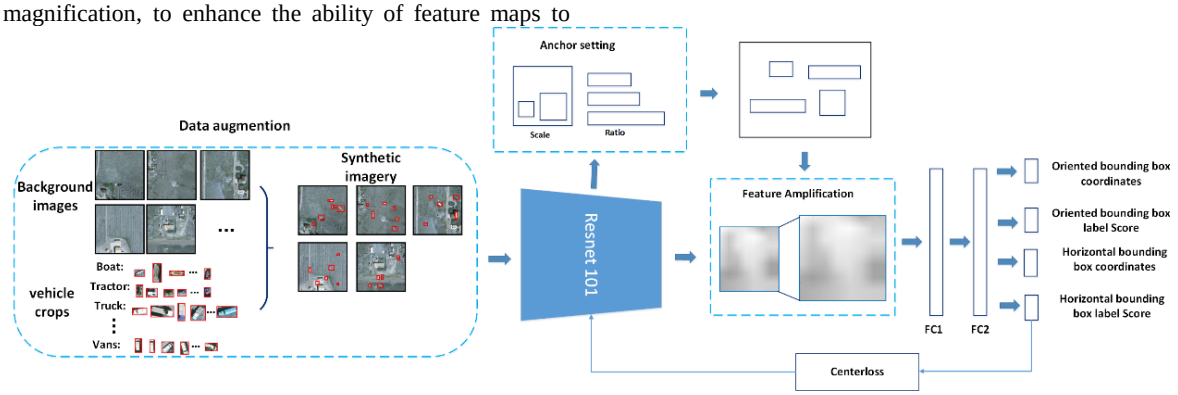

Figure 1. The architecture of the proposed framework.

2.2 Data augmentation by oversampling and stitching

We propose an oversampling and stitching method to augment the training dataset by increasing the frequency of objects with fewer training samples in order to keep a balance between the number of objects in each fine-grained vehicle category as shown in Figure 2. We segment the vehicles in the dataset according to their coordinates. By increasing the frequency of these vehicles in different background images, the number of vehicles in each category reaches a balanced state. Considering the random location of vehicles in the geographic space and reduced impact of batch size on foreground category imbalance, the rules that each image may contain all types of objects are applied to data augmentation of each image. Meanwhile, no overlap between augmented objects and existing objects requires to be ensured.

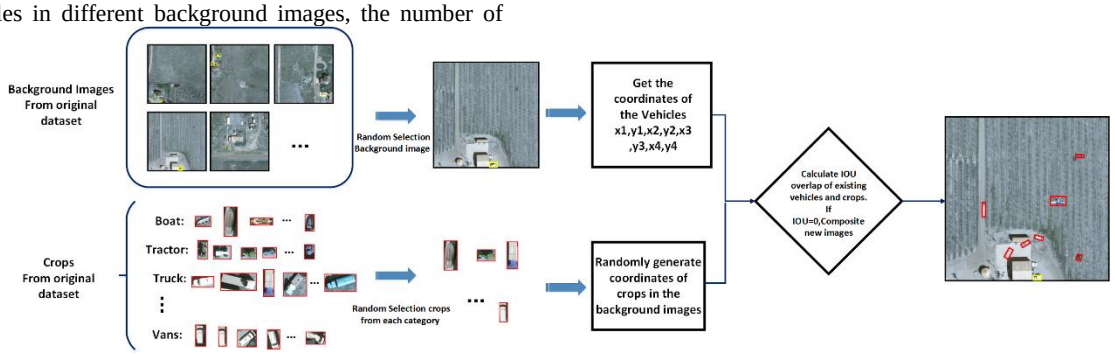

Figure 2. Schematic of oversampling and stitching data augmentation.

2.3 Magnification operation of deep feature maps

Considering the effect of the pooling operations on representing small objects in convolutional neural network, we propose to improve the resolution of feature maps so that they can better distinguish the fine-grained vehicle categories and detailed information hidden in feature maps can be enriched. There are usually two main methods for upsampling the feature map, one is interpolation, and the other is deconvolution. However, enlarging the image by deconvolution usually produces checkboard artifacts, which is not conducive to the detailed description of features. Therefore, we employ interpolation to enlarge the feature maps. Here, we use bilinear interpolation to improve the size of deep feature maps. Figure 3 shows details of bilinear interpolation to enlarge the feature map can be illustrated.

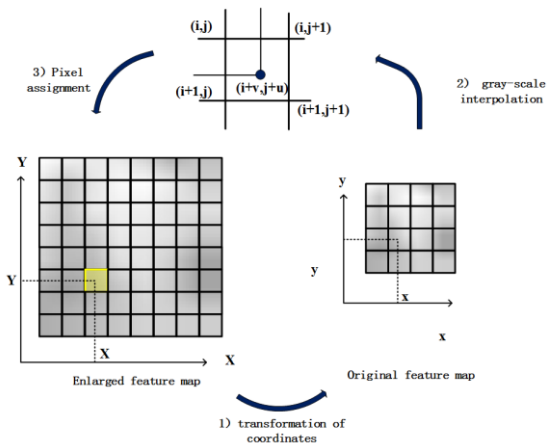

Figure 3. Flow chart of bilinear interpolation to enlarge feature map.

2.4 Multi-task loss function for joint horizontal and oriented bounding boxes

We design a joint training loss function for horizontal and oriented bounding boxes with the center loss, in order to decrease the impact of within-class diversity on vehicle detection. The proposed method in this paper can detect horizontal and rotational objects simultaneously by combining the loss function of horizontal bounding boxes with that of rotational bounding boxes. As shown in Eq. (1) and (2), the joint training loss function used in this paper consists of 5 parts, namely the cross-entropy loss of rotational objects $L_{cls}^{R}(P_R, P_R^*)$, the location loss function of the rotational objects $\sum_{i\in\{x,y,w,h\}} L_{reg}(v_i, v_i^*)$, the cross-entropy loss of horizontal objects $L_{cls}^{H}(P_H, P_H^*)$, the location loss function of the horizontal objects $\sum_{i\in\{x,y,w,h\}} L_{reg}(v_i, v_i^*)$ and center loss function $L_{Centerloss}$.

$$\begin{aligned} L(P_H, P_H^*, P_R, P_R^*, u, u^*, v, v^*) &= L_{cls}^{H}(P_H, P_H^*) + L_{cls}^{R}(P_R, P_R^*) \\ &+ \lambda_1 \sum_{i\in\{x,y,w,h\}} L_{reg}(v_i, v_i^*) + \lambda_2 \sum_{i\in\{x,y,w,h,\theta\}} L_{reg}(u_i, u_i^*) \\ &+ \lambda_3 L_{Centerloss} \end{aligned} \quad (1)$$

$$L_{centerloss} = \frac{1}{2}\sum_{n=1}^{m} || x_n - c_{y_n} ||_2^2 \quad (2)$$

Where m is the mini-batch size, c_{y_n} represents the feature center of the y_n -th category, and x_n is the feature before the fully-connected layer.

3. EXPERIMENTAL RESULTS AND ANALYZES

3.1 Description of Experimental Data

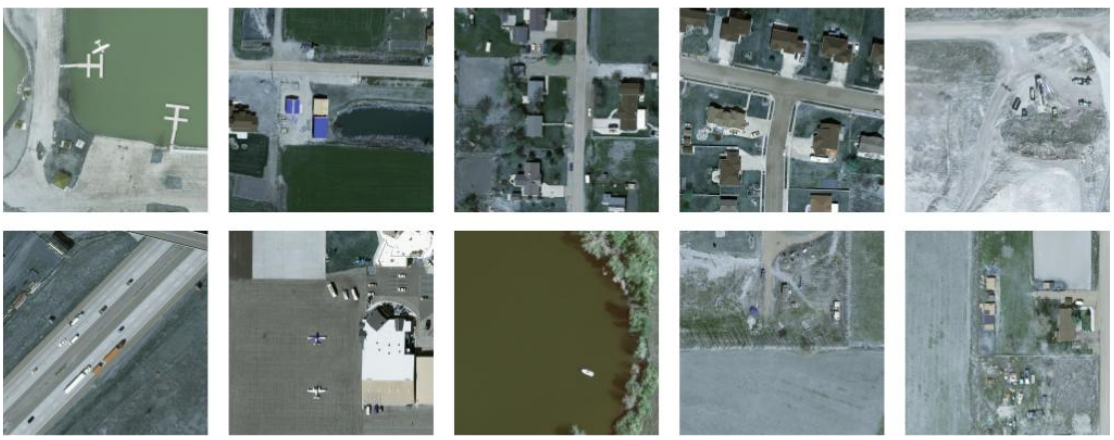

Figure 4. Samples in the VEDAI dataset.

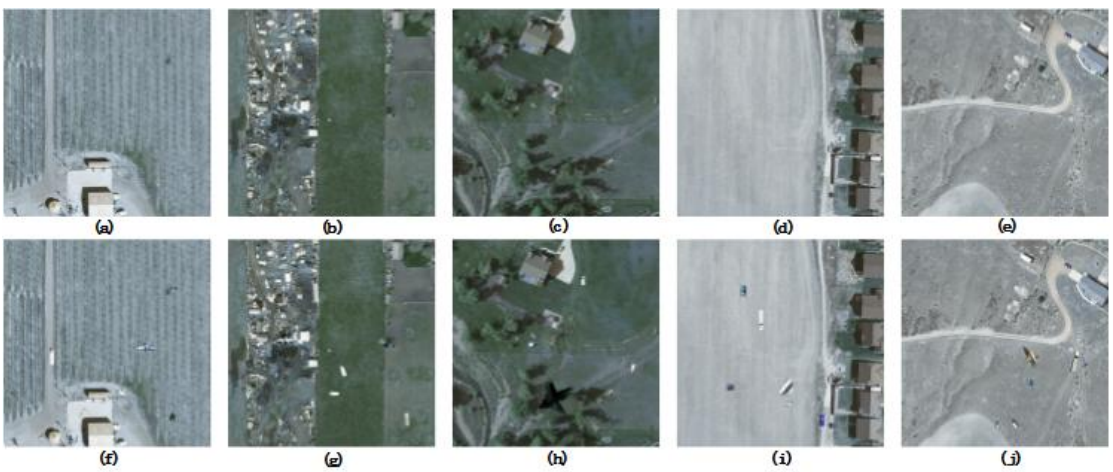

Figure 5. Comparison of images before and after synthesis by the oversampling and stitching method. (a)-(e) are original images from VEDAI data, and (f)-(j) are corresponding images synthesized by proposed method.

Figure 4 shows the experimental dataset named Vehicle Detection in Aerial Imagery (VEDAI) (Razakarivony et al., 2016) that is used in this paper. Images in the VEDAI dataset are with a size of 1024x1024. The ground sampling distance (GSD) of the original image is 12.5 centimetres. The image consists of four bands including red, green, blue and near infrared. Since it is enough to produce satisfactory performance with only the visible wavelengths for fine-grained vehicle detection. We uses the visible light channels of images for vehicle detection in this paper. The VEDAI dataset contains 9 types of fine-grained vehicles including van, tractor, pick-up, car, camping-car, boat, plane and other vehicles. The average number of vehicles in each image is 5.5, accounting for approximately 0.7% of the entire image. The research content of this paper is mainly for vehicle detection of nine fine-grained

categories. We randomly select 50% of images to be training samples while the other 50% for testing.

In this paper, the backbone for extracting features is Resnet101 which is pre-trained on the ImageNet dataset. TensorFlow framework with the Ubuntu 16.04 system is used for implementing the proposed method. The GPU for accelerating computation is GTX1080ti with 12GB display memory. The sizes of mini-batch in the RPN stage and classification stage are 256 and 512 respectively. The learning rate of first 30000 iterations is set to 0.003 while the learning rate of subsequent 70,000 epochs is 0.00003. The proposed framework will not stop until 100,000 iterations are reached. The momentum and weight decay are set to 0.9 and 0.0001. 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 above 0.7 as positive samples and those with an IOU overlap below 0.3 will be considered as negative samples.

The images augmented by the proposed method are shown in the Figure 5. The original image before data augmentation usually contains only a few vehicle types. After data augmentation, each image contains at least 9 different types of vehicles, and the vehicle position is randomly generated by the proposed algorithm. No overlap between vehicles are ensured in order to increase the frequency of the categories with a smaller number in the dataset. The proposed synthesis method also increases the background diversity of the vehicles to a certain extent.

3.2 Evaluation Metric

In the paper, the commonly used evaluation metric mean average precision (mAP) that represents the average of average precision in each type of vehicle. The higher mAP is, the better object detection performance is. The average precision in each type can be calculated as equation (3).

$$AP = \sum_{n} (R_n - R_{n-1}) P_n \tag{3}$$

Where R_n and P_n represent the recall ratio and precision ratio when n-th threshold is set. The recall ratio and precision ratio can be defined as equation (4) and (5).

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

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) between a bounding box and its true locations is above 0.5, the bounding box will be considered as TP. Otherwise, it will be FP.

3.3 Experimental Results

3.3.1 Analysis of the data augmentation approaches

The proposed vehicle detection framework in this paper are verified to prove the superiority of the data augmentation by oversampling and stitching method. We adopt three different datasets for experiments. The first dataset is after performing rotation augmentation with the angles of 90 °, 180 °, 270 ° denoted as R, the second is the dataset after the proposed oversampling and stitching augmentation in this paper denoted as O, and the third is combination of the previous two datasets denoted as M.

As shown in Table 1 and 2, 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. The oversampling and stitching data augmentation method improves the recall ratios of the vehicles with a small number of training samples such as Truck, Tractor, Boat, Vans, and Plane, and effectively improves the corresponding average detection accuracy. 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. 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 merged datasets for training the proposed network can further improve the average vehicle detection accuracy since the effective samples to train the network are further increased.

Class		Car	Pickup	Camping car	Truck	Other	Tractor	Boat	Vans	Plane	mean
R	Recall	0.858	0.787	**0.904**	0.748	**0.73**	**0.79**	0.679	0.585	0.889	0.774
	ap	0.753	**0.705**	**0.529**	0.535	**0.508**	0.491	0.43	0.381	0.864	0.577
O	Recall	**0.863**	0.777	0.883	**0.801**	0.66	0.752	**0.691**	**0.623**	**0.963**	**0.779**
	ap	0.74	0.685	0.524	**0.622**	0.44	**0.519**	0.384	**0.405**	**0.937**	**0.583**
M	Recall	0.849	0.799	0.904	0.801	0.64	0.838	0.741	0.623	0.963	0.795
	ap	0.762	0.723	0.509	0.624	0.455	0.529	0.523	0.386	0.951	0.607

Table 1. Detection accuracy of proposed algorithm for horizontal bounding boxes in three different types of datasets.

Class		Car	Pickup	Camping car	Truck	Other	Tractor	Boat	Vans	Plane	mean
R	Recall	0.857	0.783	**0.904**	0.735	**0.72**	**0.79**	0.667	0.585	0.889	0.77
	ap	0.752	**0.702**	**0.517**	0.523	**0.503**	0.492	**0.395**	0.377	0.847	0.568
O	Recall	**0.861**	0.773	0.883	**0.788**	0.66	0.762	**0.691**	**0.604**	**0.963**	**0.776**
	ap	0.737	0.68	0.506	**0.603**	0.43	**0.525**	0.384	**0.393**	**0.939**	**0.577**
M	Recall	0.849	0.795	0.899	0.795	0.66	0.838	0.741	0.623	0.963	0.796
	ap	0.765	0.719	0.51	0.612	0.447	0.527	0.512	0.386	0.951	0.604

Table 2. Detection accuracy of proposed algorithm for oriented bounding boxes in three different types of datasets.

3.3.2 Analysis of the feature map magnification operation

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 analyse the impact of different upsampling parameters on vehicle detection. Simultaneously two interpolation methods including bilinear interpolation and nearest neighbour interpolation are used for comparison. H and O are respectively denoted as the horizontal and oriented detection results.

The bold in Table 3 represents the optimal average precision of horizontal or oriented bounding boxes in each line. The mAP of the proposed framework without magnification operation for horizontal and oriented vehicles are 58.9% and 58.3%, respectively. The detection accuracy obtained by bilinear upsampling for the feature map is higher than the detection results without feature amplification. The experimental results show that the bilinear interpolation enlarged feature map can improve the detection accuracy to a certain extent.

Among them, in the comparison experiments with upsampling multiples of 1.5, 2.0, 2.5, and 3.0, bilinear interpolation by 2.0 upsampling improves the most average detection accuracy, with about 2%. At the same time, we use the nearest neighbor interpolation method of 2.0 to perform upsampling experiments on feature maps. The nearest neighbor upsampling method shows lower accuracy than that of without feature map amplification. The nearest-neighbor interpolation method will cause a Jagged effect on the enlarged feature map, which is not beneficial to the feature representations of truck, car, others, and tractor. Therefore, detection accuracy is lower than the detection results without feature amplification.

category	No interpolation		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	0.296	0.292	0.369	0.355	0.386	0.386	0.388	0.391	**0.439**	**0.435**	0.328	0.314
plane	0.953	0.953	**0.979**	**0.974**	0.951	0.951	0.965	0.965	0.954	0.954	0.96	0.959
truck	**0.68**	**0.654**	0.633	0.626	0.624	0.612	0.618	0.606	0.619	0.607	0.577	0.559
boat	0.407	0.413	0.466	0.46	**0.523**	**0.521**	0.53	0.51	0.465	0.457	0.433	0.426
camping -car	0.489	0.48	0.456	0.458	0.509	0.51	0.491	0.493	**0.525**	**0.515**	0.49	0.482
car	0.76	0.753	0.755	0.753	0.762	**0.765**	0.744	0.743	**0.763**	0.762	0.712	0.707
others	**0.497**	**0.493**	0.465	0.46	0.455	0.447	0.428	0.421	0.461	0.457	0.448	0.442
Tractor	**0.54**	**0.534**	0.491	0.498	0.529	0.527	0.505	0.502	0.506	0.509	0.516	0.517
pickup	0.678	0.676	0.695	0.69	**0.723**	**0.719**	0.718	0.715	0.725	0.718	0.688	0.685
mean	0.589	0.583	0.590	0.586	**0.607**	**0.604**	0.599	0.594	0.606	0.602	0.572	0.566

Table 3. Detection accuracy of different parameters in the magnification operation.

3.3.3 Compared with existing approaches

We have carried out comparative experiments with the existing excellent methods to demonstrate the superiority of the proposed vehicle detection framework. Faster RCNN are respectively performed on the rotation augmentation and on the merged dataset after rotation augmentation and oversampling augmentation. All the other comparison methods are only performed on the merged dataset. The detection results of the proposed method in this paper are better than those of Faster RCNN algorithm and FPN algorithm in both horizontal and oriented bounding boxes as can be seen in Table 4.

Although the FPN method selects features suitable for detecting a certain type of vehicles from the multilayer pyramid feature maps, with the highest detection accuracy in the tractor and vans. However, the detection results in other categories are not good, which is especially unsatisfactory for the airplane. Airplane is the category with the smallest number of samples. FPN builds a feature pyramid and increases the amount of training parameters, requiring numerous samples. So it is difficult to achieve the ideal detection results.

The method in this paper is improved on the Faster RCNN algorithm. The enlarged feature maps may improve the ability to distinguish fine-grained vehicles by restoring the detailed information of feature maps. Center loss may relatively enhance the gap between features of different vehicle types in the feature space 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 mAP of the Faster RCNN algorithm is lower than that of the proposed framework.

The results of the Faster RCNN of the rotation dataset and merged dataset proof that the merged datasets can further improve the average vehicle detection accuracy. Compared with the original Faster RCNN algorithm, the proposed method with the merged dataset in this paper can achieve an improvement of about 10% mAP.

Class		Car	Pickup	Camping car	Truck	Other	Tractor	Boat	Vans	Plane	mean
Faster RCNN-R	H	0.718	0.689	0.521	0.501	**0.456**	0.403	0.275	0.367	0.801	0.526
	O	0.68	0.655	0.479	0.42	**0.433**	0.414	0.261	0.372	0.804	0.502
Faster RCNN-M	H	0.749	0.698	**0.564**	0.589	0.419	0.512	0.357	0.307	0.875	0.563
	O	0.704	0.663	**0.542**	0.502	0.388	0.501	0.365	0.295	0.879	0.538
FPN-M	H	0.594	0.596	0.508	0.435	0.376	**0.594**	**0.527**	**0.665**	0.396	0.521
	O	0.621	0.599	0.502	0.43	0.365	**0.566**	0.510	**0.665**	0.396	0.517
Proposed method-M	H	**0.762**	**0.723**	0.509	**0.624**	0.419	0.529	0.523	0.386	**0.951**	**0.607**
	O	**0.765**	**0.719**	0.51	**0.612**	0.447	0.527	**0.521**	0.368	**0.951**	**0.604**

Table 4. Vehicle detection comparison experiments.

4. CONCLSION

The experimental results show that the oversampling and stitching data augmentation method can improve the imbalance of vehicle categories in the dataset to a certain extent. The combined datasets of the oversampling and stitching augmentation and rotation augmentation can improve about 3% mAP. The feature maps by magnification operation method proposed in this paper can increase the ability to classify the fine-grained vehicle categories for the network by restoring the detailed information of the feature map without increasing the amount of calculation. Considering the random direction of the vehicle, the combined horizontal and oriented bounding box proposed in this paper with center loss can simultaneously obtain the horizontal and oriented detection results. The center loss constraint can reduce the intra-class diversity of the fine-grained categories to a certain extent and improve the accuracy of vehicle detection. The proposed method achieves the mAPs of 60.7% and 60.4% in horizontal and oriented bounding boxes, respectively, which outperforms other competitive vehicle detection approaches. Compared with the original Faster RCNN algorithm, the proposed method in this paper can achieve an improvement of about 10% mAP.

ACKNOWLEDGEMENTS

We would like to thank anonymous editors and reviewers for their kind suggestions. This study is supported by The National Key Research and Development Program of China under grant no. 2017YFC0803802.

REFERECES

Cheng G, Han J. A survey on object detection in optical remote sensing images. ISPRS Journal of Photogrammetry and Remote Sensing, 2016, 117: 11-28.

Dalal N, Triggs B. Histograms of oriented gradients for human detection. 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05). IEEE, 2005, 1: 886-893.

Ding J, Xue N, Long Y, et al. Learning roi transformer for detecting oriented objects in aerial images. arXiv preprint arXiv:1812.00155, 2018.

Felzenszwalb P, McAllester D, Ramanan D. A discriminatively trained, multiscale, deformable part model. 2008 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2008: 1-8.

Girshick R. Fast r-cnn. Proceedings of the IEEE international conference on computer vision. 2015: 1440-1448.

He K, Zhang X, Ren S, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE transactions on pattern analysis and machine intelligence, 2015, 37(9): 1904-1916.

He K, Zhang X, Ren S, et al. Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

Ji H, Gao Z, Mei T, et al. Vehicle Detection in Remote Sensing Images Leveraging on Simultaneous Super-Resolution. IEEE Geoscience and Remote Sensing Letters, 2019.

Lin T Y, Dollár P, Girshick R, et al. Feature pyramid networks for object detection. Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 2117-2125.

Liu W, Anguelov D, Erhan D, et al. Ssd: Single shot multibox detector. European conference on computer vision. Springer, Cham, 2016: 21-37.

Ma J, Shao W, Ye H, et al. Arbitrary-oriented scene text detection via rotation proposals. IEEE Transactions on Multimedia, 2018, 20(11): 3111-3122.

Mandal M, Shah M, Meena P, et al. AVDNet: A Small-Sized Vehicle Detection Network for Aerial Visual Data. IEEE Geoscience and Remote Sensing Letters, 2019.

Oksuz K, Cam B C, Akbas E, et al. Generating positive bounding boxes for balanced training of object detectors. The IEEE Winter Conference on Applications of Computer Vision. 2020: 894-903.

Ouyang W , Wang X , Zhang C , et al. Factors in Finetuning Deep Model for Object Detection with Long-Tail Distribution.IEEE Conference on Computer Vision & Pattern Recognition. IEEE, 2016.

Razakarivony S, Jurie F. Vehicle detection in aerial imagery: A small target detection benchmark. Journal of Visual Communication and Image Representation, 2016, 34: 187-203.

Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection. Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 779-788.

Redmon J, Farhadi A. YOLO9000: better, faster, stronger. Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 7263-7271.

Redmon J, Farhadi A. Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767, 2018.

Ren S, He K, Girshick R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems. 2015: 91-99.

Singh B, Davis L S. An analysis of scale invariance in object detection snip. Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 3578-3587.

Sun Z, Bebis G, Miller R. On-road vehicle detection: A review. IEEE transactions on pattern analysis and machine intelligence, 2006, 28(5): 694-711.

Tayara H, Soo K G, Chong K T. Vehicle detection and counting in high-resolution aerial images using convolutional regression neural network. IEEE Access, 2017, 6: 2220-2230.

Van de Sande K E A, Uijlings J R R, Gevers T, et al. Segmentation as selective search for object recognition. 2011 International Conference on Computer Vision. IEEE, 2011: 1879-1886.

Viola P, Jones M. Rapid object detection using a boosted cascade of simple features. Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001. IEEE, 2001, 1: I-I.

Wang H, Wang Q, Yang F, et al. Data augmentation for object detection via progressive and selective instance-switching. arXiv preprint arXiv:1906.00358, 2019.

Wen Y, Zhang K, Li Z, et al. A discriminative feature learning approach for deep face recognition. European conference on computer vision. Springer, Cham, 2016: 499-515.

Yang X, Sun H, Fu K, et al. Automatic ship detection in remote sensing images from google earth of complex scenes based on multiscale rotation dense feature pyramid networks. Remote Sensing, 2018, 10(1): 132.