VEHICLE DETECTION WITH PARTIAL ANCHORS IN REMOTE SENSING IMAGES

Fuyan Ma[1]*, Bin Sun*[1]*, Shutao Li*[1] *and Jun Sun*[2]

College of Electrical and Information Engineering, Hunan University, Changsha, China[1]
Information Department, Fujitsu Reaserch and Development Center, Bejing, China[2]

ABSTRACT

Vehicle detection in remote sensing(RS) images has been an active topic with the development of computer vision in recent years. However, directly applying conventional horizontal anchor-based detection methods in oriented vehicle detection often acquires poor performance. Although rotated anchors have been used to tackle this problem, this design leads to heavy computational cost because of thousands of rotated anchors generated in each level feature map. In this paper, we propose to detect vehicles with partial anchors, which greatly accelerates detection process. The novel Partial Anchors based Detection Network(PADeN) filter out redundant anchors with semantic information. To boost the performance of PADeN, the centerness mask branch is added into the network. The results demonstrate that PADeN significantly outperforms previous approaches in vehicle detection and achieves the mAP of 76.9%.

***Index Terms*—** Vehicle detection, anchor-based detection, rotated anchors

1. INTRODUCTION

Vehicle detection in RS images[1, 2] has drawn more and more attention due to the need of intelligent transportation and earth observation. It aims at identifying the category of vehicles and accurately locating each vehicle in RS images. Though numerous efforts have been devoted to tackling this task, vehicle detection is still quite challenging because of the birdview perspective, the various scales and appearances of vehicles. Especially, detecting vehicles with arbitrary orientations also makes it an extremely difficult task, as directly applying horizontal object detection methods often introduces mismatches between the Region of Interests(RoIs) and vehicles, which drastically enlarges the searching space.

Anchor-based detection methods [3, 4] have been proved effective in open benchmarks. Take Rotation Region Proposal Networks(RRPN) [5] for instance, it first generates oriented region proposals from a dense of anchors with angles, and then refines their locations via rotated bounding box regression. However, anchor-based detection pipelines usually begin with a large set of densely distributed anchors, which leads to significant computational cost especially when the network employs intersection over union(IOU) between anchor boxes and ground-truth boxes in training objectives. Anchor-free detection methods [6, 7] predict bounding boxes by keypoints instead of anchors with predefined scales and aspect ratios. However, since only keypoints are used to predict bounding boxes, anchor-free detectors suffer from lower recall compared with anchor-based detectors.

In this paper, we propose an effective and more efficient end to end detection framework called PADeN for multi-oriented vehicles in RS images. The semantic features of vehicles are leveraged to filter out redundant anchors. Then the oriented region proposals are regressed from filtered anchors through RRPN. The performance of our method is verified by using a high-resolution RS dataset[8].

2. THE PROPOSED METHOD

In this section, we first formalize the oriented vehicle detection problem and then introduce the details of the proposed method.

2.1. Problem Definition

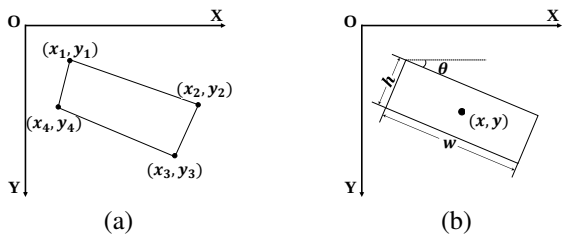


Fig. 1. Different ways to represent oriented bounding boxes.

The straightforward method to represent a rotated bounding box is using four points in clockwise $(x1, y1, x2, y2, x3, y3, x4, y4)$, as shown in Fig. 1(a). However, the four points strategy is not adopted in our work because what the four points

This work is supported by the National Natural Science Fund of China(No. 61520106001, No. 61801178), the Fund of Key Laboratory of Visual Perception and Artificial Intelligence of Hunan Province(No. 2018TP1013), and the Natural Science Fund of Hunan Province(2018JJ3071).

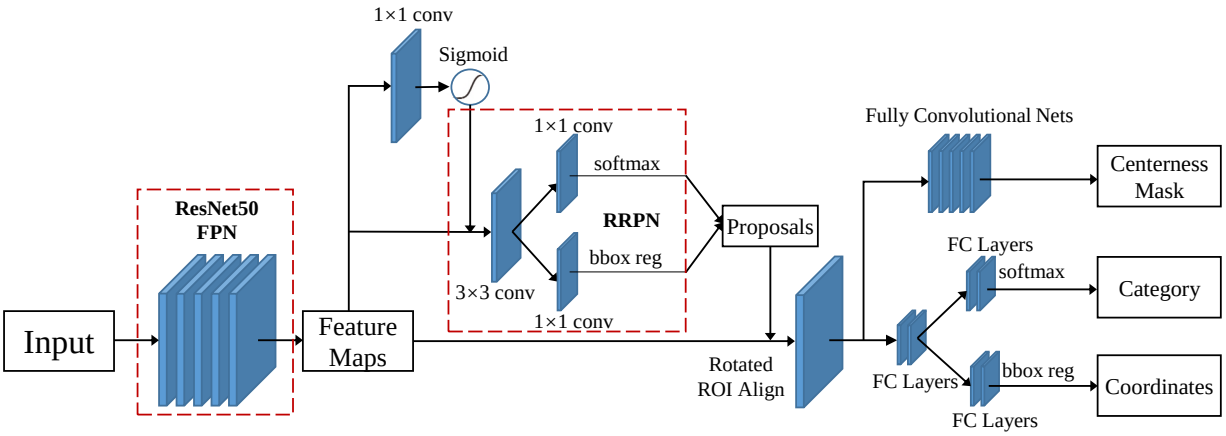


Fig. 2. The illustration of our proposed vehicle detection framework.

form is a quadrangle and is probably not a rectangle all the time. The vehicles in RS images can be closely enclosed by rectangles with orientation. 5 tuples (x, y, w, h, θ) are used to represent a rectangular bounding box as shown in Fig. 1(b). The coordinate (x, y) represents the geometric center of the bounding box. The width w is set as longer side of the bounding box and the height h as the shorter side. The angle θ is the orientation of longer side. We fix the range of θ to $[-\pi/2, \pi/2)$ by shifting the opposite direction if the angle θ is out of this range. Notice that the bounding box in the rest paper refers to a rectangular box represented with (x, y, w, h, θ).

2.2. Partial Anchors-based Detection Network

The architecture of the proposed Partial Anchors-based Detection Network(PADeN) is shown in Fig. 2. We adopt the popular two-stage oriented detection framework that consists of rotation region proposal and region classification. Partial Anchors-based Detection Network is based on RRPN. As shown in Fig. 2, we employ the convolutional layers of ResNet50 [9] and feature pyramid networks(FPN)[10] in the front of the framework. The centerness estimation branch is first designed to filter out redundant anchors. The RRPN is used to generate oriented region proposals, i.e., the oriented bounding boxes that enclose the vehicles respectively. And further R-CNN performs bounding box regression for the proposals to better fit the vehicle instances. Meanwhile, the centerness mask branch is used to guide the network estimate the existence probability of vehicles precisely.

Centerness Map Generation. Particularly, we only consider the case that the bounding box is a rectangle as we mention above. The center region of the rectangle on the centerness map is designed to be roughly a shrunk version of the original one. We denote an original bounding box $(x_o, y_o, w_o, h_o, \theta_o)$. Then we can obtain the center region box $(x_o, y_o, \alpha w_o, \alpha h_o, \theta_o)$, where we set α = 0.2 in this paper.

Partial rotated anchors. Traditional anchors generating over a feature map of size $W \times H$ with a stride of S result in excessively heavy computational cost, because many anchors are placed in regions where the vehicles of interest are unlikely to exist. We rethink the relationship between segmentation tasks and detection tasks. As far as we are concerned, segmentation results can provide coarse location and boundary information, and what we want to get in detection tasks is the information of location. So we think whether the network can only generate anchors where the vehicles exist. However, it is not feasible to generate anchors according to the segmentation results, as it has the order reversed and is still time-consuming. Therefore we design a more efficient scheme to arrange these anchors. Since sufficient semantic features are generated from ResNet50 and FPN, we are trying to infer the geometry center of vehicle from feature maps in centerness estimation branch. This branch applies a 1×1 convolution to the feature maps to obtain maps of centerness scores, which are then converted to existence probability values via an element-wise sigmoid function. Based on the resultant centerness probability map, we then determine the positive regions where vehicles may possibly exist by selecting those centerness whose corresponding probability values are above a predefined centerness threshold ε. The centerness threshold ε is set to 0.001 in this work. In order to get a good balance between efficiency and accuracy, we add a centerness mask branch in the training phase to get precise estimation of centerness. Significantly, the network can directly generate bounding boxes without the centerness mask branch in the inference phase.

2.3. Loss Functions

We minimize the multi-task loss in an end to end fashion. The loss can be formulated as

$$L = \lambda_1 L_{ce} + \lambda_2 L_{cm} + L_{cls} + L_{reg} \quad (1)$$

Fig. 3. Visualization of vehicle detection results on DOTA testing dataset. Results marked in magenta are small vehicles, while the other results marked in green are large vehicles. We select the predicted bounding boxes with scores above 0.3, and a NMS with threshold 0.5 is applied for duplicate removal.

where L_{ce} and L_{cm} represent the losses for centerness estimation branch and centerness mask branch, respectively. For L_{ce} and L_{cm}, we use the focal loss [11]. L_{cls} is classification loss function which is cross entropy. L_{reg} is location loss functions which is smooth L1 loss defined in [12]. λ_1 and λ_2 weight the importance among four losses.

In addition, we use the following approach to perform bounding box regression:

$$\begin{aligned} t_x &= (x - x_a)/w_a, t_y = (y - y_a)/h_a \\ t_w &= log(w/w_a), t_h = log(h/h_a) \\ t_\theta &= \theta - \theta_a \end{aligned} \tag{2}$$

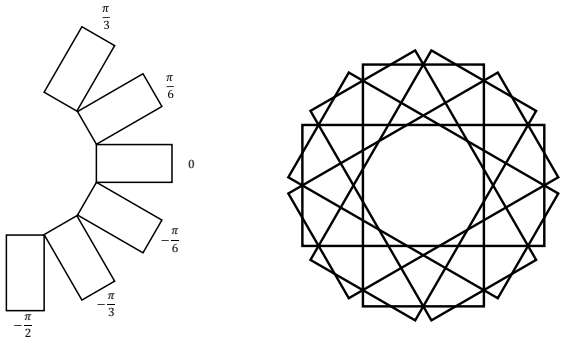
$$\begin{aligned} t_x^* &= (x^* - x_a)/w_a, t_y^* = (y^* - y_a)/h_a \\ t_w^* &= log(w^*/w_a), t_h^* = log(h^*/h_a) \\ t_\theta^* &= \theta^* - \theta_a \end{aligned} \tag{3}$$

where x, x_a, x^* are for the predicted box, anchor box and ground-truth box respectively; and it is the same for y, w, h, θ.

3. EXPERIMENTS

3.1. Dataset and Settings

To comprehensively demonstrate the effectiveness of our proposed detection framework, we conduct experiments on DOTA [8], the largest dataset for object detection in aerial images with oriented bounding box annotations. Though the testing images have no labels, We can submit the test results to DOTA Evaluation Server. We only focus on vehicle detection, so we do experiments on these images which contain small vehicle(SV) and large vehicle(LV) instances. For both training and testing, we resize images at three scales (0.5, 1.0, 1.5) and split images into the blocks of 1024×1024 with the overlap of 512 pixels using the official DOTA development kit. During testing, we combine outputs using rotated Non-Maximum Suppression(NMS) as the final results.



(a) (b)

Fig. 4. Rotated anchors. (a) Anchors with different angles. (b) An example of anchors used in our work.

The rotated anchors own ranges of $\{32^2, 64^2, 128^2, 256^2, 512^2\}$ on feature maps from five stages of ResNet50 and FPN. On each feature map, there are two aspect ratios of 1:2 and 1:4, and six angles of $-\pi/2$, $-\pi/3$, $-\pi/6$, 0, $\pi/6$, $\pi/3$, illustrated in Fig. 4. As a result, there are a total of 12 different kinds of anchors per location.

Our network is initialized by the pre-trained ResNet50 model, which is optimized by stochastic gradient descent(SGD). In the multi-task loss function, we set $\lambda_1 = 1$ and $\lambda_2 = 0.5$ to balance different branches. We train 12 epoches in total with batchsize 2 on 2 GTX 1080Ti GPUs. The weights of the network are updated by using an initial learning rate of 0.005, which is decreased by 0.1 at epoch 8 and 11.

3.2. Result and Analysis

The experimental results of our method compared with other methods are given in Table 1. FR-O indicates the Faster R-CNN OBB detector, which is the official baseline provided by DOTA. Ours[1] means our detection method only with partial anchors. Ours[2] means our detection method with partial anchors and centerness mask branch. Ours[3] means our detection method with partial anchors and centerness mask branch, which uses one more scale anchor. Our method outperforms other methods in both mAP and time cost. Our proposed method obtains the 76.9% mAP, which is 40.2% higher than the official baseline[8]. Compared with our reimplemented RRPN, our performance of vehicle detection is 6.2% higher. Our proposed method armed with the centerness mask branch has 3.4% mAP improvement over the method without this branch. In addition, the time cost of our method is also reported in Table 1. Benefiting from partial anchors, our method takes less time for each image in average.

Table 1. Quantitative results of FR-O, RRPN and our methods. The best result in each case is highlighted in bold.

Method	SV	LV	mAP	Time cost
FR-O[8]	35.3	38.0	36.7	-
RRPN[5]	69.7	71.7	70.7	0.29s
Ours[1]	68.8	71.2	70.0	**0.13s**
Ours[2]	72.7	74.1	73.4	0.15s
Ours[3]	**77.5**	**76.3**	**76.9**	0.21s

4. CONCLUSION

In this paper, we propose a novel method named PADeN to detect vehicles in RS images, which outperforms the conventional rotation region proposal network based detection methods. The design of detecting with partial anchors brings significant inference speed improvements for oriented vehicle detection with competitive mAP. The results demonstrate that PADeN benefits a lot from the centerness mask branch. The low-GPU-cost design also allow multiple scale anchors to boost the performance of PADeN.

5. REFERENCES

[1] Lars Sommer, Tobias Schuchert, and Jürgen Beyerer, "Comprehensive analysis of deep learning based vehicle detection in aerial images," *IEEE Transactions on Circuits and Systems for Video Technology*, 2018.

[2] Ziyi Chen, Cheng Wang, Huan Luo, Hanyun Wang, Yiping Chen, Chenglu Wen, Yongtao Yu, Liujuan Cao, and Jonathan Li, "Vehicle detection in high-resolution aerial images based on fast sparse representation classification and multiorder feature," *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 8, pp. 2296–2309, 2016.

[3] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems*, 2015, pp. 91–99.

[4] Jiaqi Wang, Kai Chen, Shuo Yang, Chen Change Loy, and Dahua Lin, "Region proposal by guided anchoring," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2965–2974.

[5] Jianqi Ma, Weiyuan Shao, Hao Ye, Li Wang, Hong Wang, Yingbin Zheng, and Xiangyang Xue, "Arbitrary-oriented scene text detection via rotation proposals," *IEEE Transactions on Multimedia*, vol. 20, no. 11, pp. 3111–3122, 2018.

[6] Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, and Qi Tian, "Centernet: Keypoint triplets for object detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 6569–6578.

[7] Hei Law and Jia Deng, "Cornernet: Detecting objects as paired keypoints," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 734–750.

[8] Gui-Song Xia, Xiang Bai, Jian Ding, Zhen Zhu, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liangpei Zhang, "Dota: A large-scale dataset for object detection in aerial images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3974–3983.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[10] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2117–2125.

[11] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2980–2988.

[12] Ross Girshick, "Fast R-CNN," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1440–1448.