2021 IEEE International Conference on Multimedia & Expo Workshops (ICMEW) | 978-1-6654-4989-2/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICMEW53276.2021.9455957

ORIENTED OBJECT DETECTION FOR REMOTE SENSING IMAGES BASED ON WEAKLY SUPERVISED LEARNING

Yongqing Sun[1], *Jie Ran*[2], *Feng Yang*[2,✉], *Chenqiang Gao*[2], *Takayuki Kurozumi*[1], *Hideaki Kimata*[1], *Ziqi Ye*[2]

[1]NTT Media Intelligence Laboratories, Yokosuka 239-0847, Japan
[2]School of Communication and Information Engineering,
Chongqing University of Posts and Telecommunications, Chongqing 400065, China
yongqing.sun.fb@hco.ntt.co.jp, yangfeng@cqupt.edu.cn

ABSTRACT

Object detection of remote sensing images (RSIs) is an active yet challenging task because of the complex appearance of ground objects and the particular imaging views. One of the difficulties in RSI object detection is the orientation variation, where the objects could take on arbitrary orientations due to the birdview shot from high altitudes. For oriented object detection, existing methods rely on largescale dense oriented annotations for training deep networks under full supervision, which are resource-intensive. To address this problem, we propose a kind of weakly supervised oriented object detection method in this paper. With only the horizontal-object supervision, we rotate object proposals via an angle search strategy to align them as horizontally as possible and detect the oriented objects just like the horizontal ones. We aim to mine more oriented objects and thus can train the Rotational RCNN framework. Experimental results demonstrate that our method can achieve significant performance improvement on the oriented object detection and outperforms the state-of-the-art methods.

Index Terms— oriented object detection, weakly supervised learning, remote sensing images

1. INTRODUCTION

Object detection is one of the fundamental tasks for high-resolution optical remote sensing images (RSIs), which can be widely used in many kinds of applications, e.g., environmental monitoring, urban planning, etc. In contrast to natural images taken from horizontal perspectives, remote sensing images are typically taken with bird's-eye views, where the objects are presented in arbitrary orientations as shown in Fig. 1(a). A primary challenge in object detection for RSIs is orientation variation across object instances. To address this issue, largescale orientation annotations were introduced to train deep networks that can detect rotational objects. Nevertheless, annotating *oriented bounding boxes* (OBBs) for object instances is particularly labor-intensive and time-consuming [1], compared to the annotation of *horizontal bounding boxes* (HBBs) for horizontal or vertical objects (named as easy samples below). Therefore, investigating the potentials of weakly supervised oriented object detection with only easy samples annotation can effectively mitigate the labor cost, showing great practical significance. Given a training set containing only easy samples, e.g., horizontal or vertical objects in Fig. 1(b), we aim to detect rotational objects with arbitrary orientations.

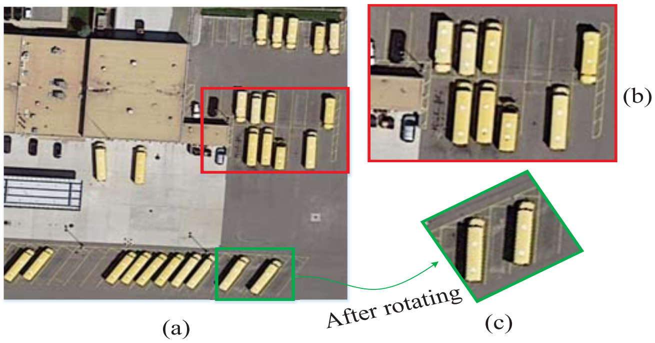


Fig. 1. An example of orientation variations of objects in remote sensing images. (a) Large vehicles with arbitrary orientation. (b) Horizontal or vertical objects are easy to annotate. (c) Oriented objects that are difficult to annotate can be rotated towards horizontally or vertically.

Many of recent progresses on object detection in RSIs have benefited a lot from the deep learning frameworks, which almost rely on the training data from largescale datasets with dense annotations. Early methods follow and transfer object detection algorithms developed for natural scenes to the RSIs domain [2–4]. Recently, oriented object detection methods [5–7] have been proposed to eliminate the misalignments between the bounding boxes and objects. Approaches of oriented object detection trained in full supervision have mainly followed the RCNN frameworks [8–14], where rotated proposals are generated first and then classified. In the cost of enormous additional OBB annotations [1], these methods have reported promising detection performances, by introducing rotated anchors [6, 15–18] or rotated RoI learning module [5]. However, these methods use fully oriented annotations, which are resource intensive.

Despite the promising results achieved by state-of-the-art fully-supervised oriented object detection work [5, 19], the necessary annotations are expensive and multiplying. In order to bridge this gap, we are motivated to utilize HBB supervision: for remote sensing images, only easy samples (horizontal or vertical objects) are annotated. The annotation process for HBB supervision is almost the same as it in the weakly-supervised annotation, where the annotators only select a small part of horizontal or vertical objects easy to annotate and there is no need for angle annotations. It significantly reduces annotation resources compared to full supervision.

We propose a weakly supervised oriented object detection method in this paper. First, we use easy samples with annotated ground truth to train deep networks. Second, the RSIs are rotated with a series of angles towards horizontally or vertically as in Fig. 1(c), and then we use the trained deep networks to predict OBBs, and leverage these pseudo ground truths as additional supervision. Third, when labeling pseudo ground truths of OBBs, besides the HBBs, we aim to mine more oriented objects and thus can train the rotational RCNN framework. To our best knowledge, this is the first work to use HBB supervision for the challenging problem of detecting oriented objects. We show that the horizontal annotation significantly saves annotation time compared to fully-supervised annotation. The proposed method is evaluated on the DOTA dataset [1], and the experimental results reveal that our method can achieve significant improvement on the oriented object detection.

2. METHODOLOGY

In this section, we define our task, present the architecture of our oriented object detection network via a weak supervision strategy, and finally discuss details of training and inference, respectively.

2.1. Overall framework

Training images can contain multiple categories objects, and include a variety of annotation information. Unlike the full supervision setting for oriented object detection, which provides the position, shape and direction information of the whole objects, in our weakly supervised oriented object detection setting, only horizontal and vertical objects are included in the training images, and the supervision information only contains position and shape with no need for angle.

Fig. 2 shows the proposed two-stage framework for weakly supervised object detection in this paper. The first step is to train our network with the annotated horizontal and vertical objects. The second step is to use our trained model to mine the rotation objects in the training image. The details of rotation objects mining are described in Section 2.2. The third step is to use all the data after mining as a supplement to ground truth and rewrite ground truth. Finally we use the rewritten ground truth to train the network again. It should be noted that the orientation of the object is predicted in both of the training stages.

2.2. Oriented object mining

We rotate all the training images with no labels in the DOTA dataset [1] to align the oriented objects as horizontally or vertically as possible. Then the rotated images are put into the detection model trained on the annotated horizontal and vertical objects to generate more detected results as pseudo labels. The angle search space ranges from -π/4 to π/4, where the interval of angle space is π/36. After that, we rewrite the ground truth according to the test result and the corresponding angle. For an image, we record the results generated by each angle, and then perform NMS processing. To ensure the high quality of the ground truth after rewriting, we set a threshold 0.7 for the score of the test results. When the detection result score is greater than threshold, it will be written to the ground truth. Finally, when we get the rewritten ground truth, we can retrain our framework with additional information. The result of oriented object mining is shown in Fig. 4, from which we can easily observe that most of oriented objects are robustly detected, automatically providing almost correct annotation information for retraining of the proposed model.

2.3. Oriented object detection network

In order to ensure the robustness of the object detection model, our basic network is R2CNN [19], which is built on Faster R-CNN [9] and further improved based on the data characteristics. The framework of Faster R-CNN is composed of three components: feature extraction (ResNet-50-FPN module), proposal generation (RPN module), and bounding box localization and recognition (Detection module). First, we adopt ResNet-50 with FPN as the backbone to extract features. And then, RPN is utilized to generate proposals for each image. For learning the direction information of objects, R2CNN [19] add the branch that predicts the object direction to original detection branch for joint training. Inspired by Mask R-CNN [10], we replace the RoI Pooling with RoI Align. Through the RoI Align operation, each proposal becomes a fixed-size feature map. Finally, the oriented bounding box location and recognition is implemented simultaneously based on the feature map.

2.3.1. Oriented bounding box regression

In the training phase, we first only use the horizontal and vertical objects (the initial angle is -π/2) to train the object detection network. Then we use orientation mining to generate

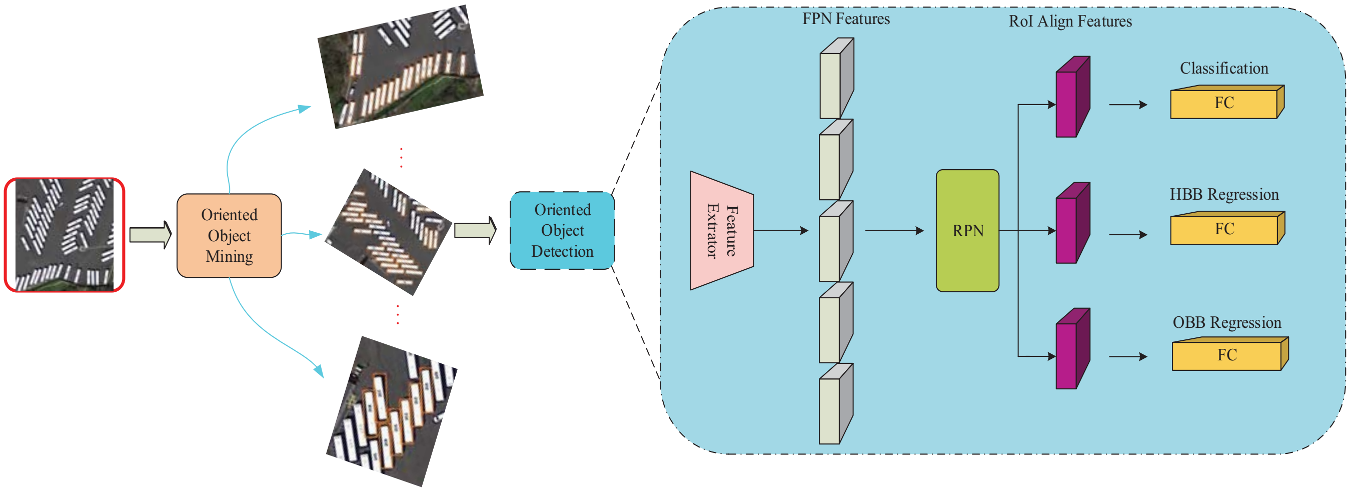


Fig. 2. The overall architecture of our framework. The whole network structure consists of two parts: oriented objects mining module, which aims at mining the information of oriented objects, and detection module.

pseudo ground truths containing rotational objects to train our model.

Considering that our model needs to learn the direction information of the objects, inspired by RoI Transformer [5], we define the regression targets of offsets relative to *Rotated ROIs* (RRoIs) as,

$$
\begin{aligned}
t_x^* &= \frac{1}{w_r}\left((x^* - x_r)\cos\theta_r + (y^* - y_r)\sin\theta_r\right),\\
t_y^* &= \frac{1}{h_r}\left((y^* - y_r)\cos\theta_r - (x^* - x_r)\sin\theta_r\right),\\
t_w^* &= \log\frac{w^*}{w_r}, \quad t_h^* = \log\frac{h^*}{h_r},\\
t_\theta^* &= \frac{1}{2\pi}\left((\theta^* - \theta_r) \mod 2\pi\right).
\end{aligned} \tag{1}
$$

where (x_r, y_r, w_r, h_r, θ_r) is a stacked vector for representing location, width, height and orientation of a RRoI and (x^*, y^*, w^*, h^*, θ^*) is the *rotated ground truth* (RGT) parameters of an OBB. The mod is used to adjust the angle offset target t_θ^* in [0, 2π) for the convenience of computation.

2.3.2. *IoU between polygons*

When matching between RRoI and RGT, we still use IoU as the criterion to judge positive and negative samples. If a RRoI has an IoU more than 0.5 with any RGT, it is considered to be True Positive (TP). We use Eq. (2) to calculate the IoU between RRoI and RGT. It has a similar form with the IoU calculation between horizontal bounding boxes. The only difference is that the IoU calculation for RRoIs is performed within polygons. The B_r means the bounding box of a RRoI. The B_{gt} represents the bounding box of a ground truth. The *area* is a function for calculating the area of an arbitrary polygon.

$$
IOU = \frac{\text{area}(B_r \cap B_{gt})}{\text{area}(B_r \cup B_{gt})} \tag{2}
$$

2.3.3. *Loss function*

For the calculation of loss function, each training proposal p is labeled with a ground-truth category label c, and a ground truth bounding box regression target v. We use a multi-task loss L on each labeled proposal to jointly train for classification and bounding box regression:

$$
\begin{aligned}
L_{total} &= L_{cls} + L_{bbox},\\
L_{bbox} &= L_{Hbbox} + L_{Obbox},
\end{aligned} \tag{3}
$$

where L_{cls} and L_{bbox} are defined in Eq. (4) and Eq. (5), respectively. L_{Hbbox} and L_{Obbox} are the regression loss of the horizontal box and the oriented one respectively.

$$
L_{cls}(p, u) = -\log p_u, \tag{4}
$$

$$
L_{bbox}(t, v) = \sum_i smoothL_1(t - v), \tag{5}
$$

$$
\text{smooth } L_1(x) = \begin{cases} 0.5x^2 & \text{if } |x| < 1,\\ |x| - 0.5 & \text{otherwise .} \end{cases} \tag{6}
$$

Here, t is the predicted bounding box regression offsets.

3. EXPERIMENT

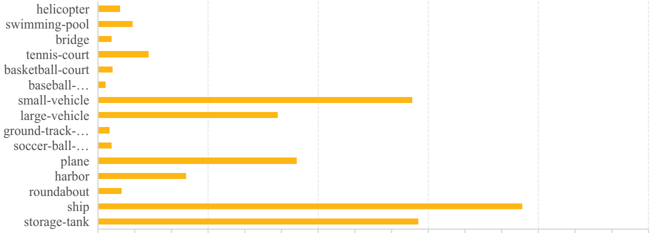
3.1. Dataset and metrics

Our method aims to train network by weakly supervised learning method with only horizontal and vertical objects annotations, and has the ability of robustly detecting the rotation

Table 1. Comparisons with state-of-the-art methods. The short names for each category can be found in Section 3.1. The R2CNN [19] means Rotational Region CNN, R2CNN-A means that data is augmented by rotating the image and the RoITrans [5] means Learning RoI Transformer, which used a design of RRoI learner.

Method	*mAP*	*PE*	*BD*	*BDE*	*GTF*	*SV*	*LV*	*ship*	*TC*	*BC*	*ST*	*SBF*	*RA*	*HB*	*SP*	*HC*
R2CNN [19]	30.6	28	33.7	9.3	20.9	21.1	16.1	24.4	36.1	49.1	76.3	13.7	51.4	15.3	33	30.7
R2CNN-A	34.4	42.2	35	14.8	15.9	31.7	17.7	27	57.8	44.9	75.8	21.9	54.8	21	34.5	22.4
ROITrans [5]	32	23.6	35.1	10.5	24.5	20.2	16.4	24	41.4	50.2	80.6	18.7	57.5	15.3	32.4	29.8
Ours	**38.6**	51.5	38.7	16.1	36.8	29.8	19.2	23.4	83.9	50.6	80	18.9	50.2	25.6	28.7	25.5

objects. To evaluate the proposed method, our experiments are performed on the dataset, sampled from the DOTA [1], and the annotations just include horizontal and vertical objects. The training dataset contains a total of 6817 remote sensing images. The size of the image is 1024 × 1024. It includes 15 categories, including Plane (PE), Baseball diamond (BD), Bridge (BDE), Ground track filed (GTF), Small vehicle (SV), Large vehicle (LV), Tennis court (TC), Basketball court (BC), Storage tank (ST), Soccer-ball field (SBF), Roundabout (RA), Harbor (HB), Swimming pool (SP), and Helicopter (HC). There are a total of 77,843 instance objects after sampling. The distribution of the number of instance in each category in the training dataset is shown in Fig. 3.



	storage-tank	ship	roundabout	harbor	plane	soccer-ball-field	ground-track-field	large-vehicle	small-vehicle	baseball-diamond	basketball-court	tennis-court	bridge	swimming-pool	helicopter
■ instances	14542	19263	1064	3993	9023	604	521	8160	14260	324	643	2288	595	1570	993

Fig. 3. Distribution of the numbers of instances in each category in the training dataset.

For performance evaluation, we adopt the widely used metrics in aerial object detection community: mAP, and we use the test dataset in DOTA [1] to test and evaluate the model in the official evaluation system of DOTA.

3.2. Implementation details

We train the proposed method in a batch size of 4 on 4 GPUs. For RPN [9], we used 15 anchors according to original Light-Head R-CNN [20]. There are 2000 RoIs from RPN [9] before Nonmaximum Suppression (NMS) and 800 RoIs after using NMS. The 512 RoIs are sampled for the training of R-CNN [21]. The learning rate is initial to 0.0005, that is decreased by a factor of 10 at the 9th epoch and the 11th epoch. For testing, we adopt 2000 RoIs before NMS, while 1000 after NMS processing. For oriented objects mining, the angle space ranges from -π/4 to π/4, and the interval is π/36. The threshold used to filter the result of mining is set to 0.6.

3.3. Comparisons with state-of-the-art methods

We use the R2CNN [19] and RoI Transformer [5] as the baseline methods for comparison. The results of different methods are shown in Fig. 5. From the Fig. 5, we can see that the False Positive (FP) rate of RoI Transformer (the first row) is lower than R2CNN (the second row). However, for both methods, the oriented objects cannot be processed. That is because there are only horizontal or vertical object annotations in the training dataset in the training process, R2CNN and RoI Transformer cannot learn any information of the oriented objects. But for ours, the oriented objects can be detected robustly. In Table. 1, we compare the performance of different methods. From the table, we can see that the performance of object detection network after weakly supervised learning has been greatly improved. Compared with R2CNN and RoI Transformer, the performance of our method is improved by 8% and 6.6% mAP, respectively. The improvement of network performance is due to that our method uses weakly supervised learning strategy to mine more supervised information of oriented objects. Thus, our method have stronger generalization performance.

3.4. Comparison with data augmentation

To differentiate our method from data augmentation, we manually rotate the training data consisting of horizontal and vertical objects in the same interval of π/36, and then feed the augmented data into R2CNN for training, in comparison with our method. The results are shown in Table 1, where R2CNN-A means data augmentation with rotation. Compared with R2CNN-A, the performance of our method is improved by 4.2%. This is because simply rotating the training objects cannot introduce additional pseudo ground truth information beyond the annotated ones, and our method is to mine new oriented objects with variant texture and appearance to make full use of the weak supervision.

4. CONCLUSION

In this paper, we propose an oriented object detection method based on weakly supervised learning, which can handle the problem with only horizontal annotations available. We rotate

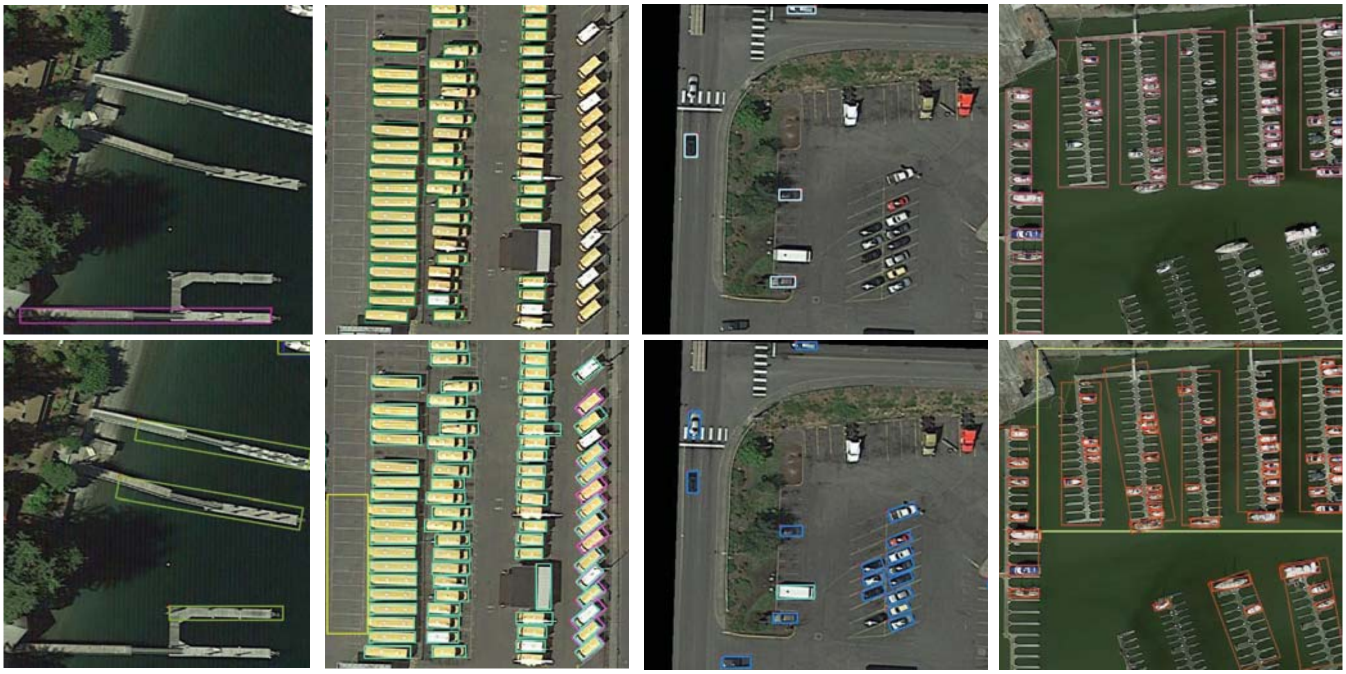


Fig. 4. The top: the images of the training dataset. The bottom: the images of the training dataset after oriented object mining. By using oriented object mining, we extend the information of oriented objects on the original training dataset.

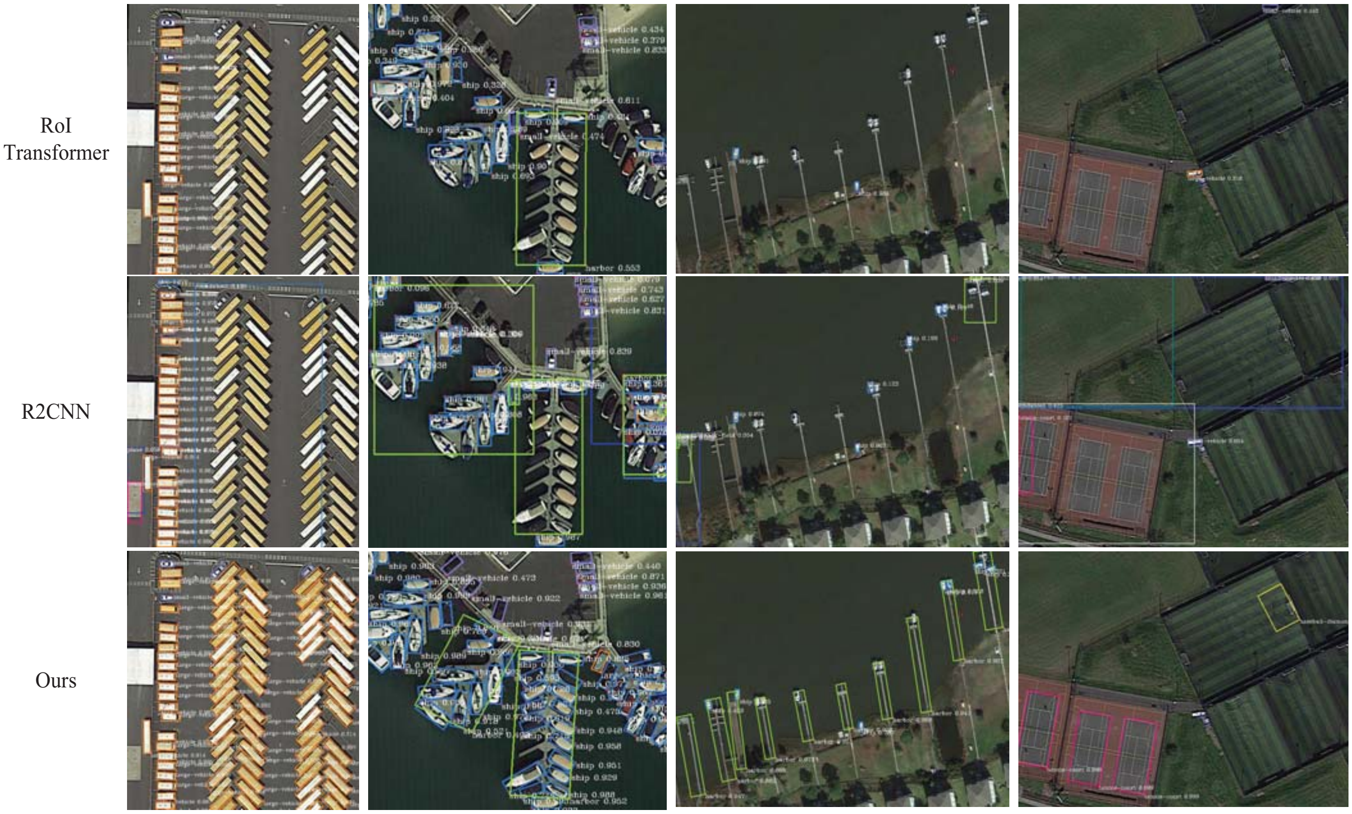


Fig. 5. From top to bottom are the results of RoI Transformer [5], R2CNN [19] and our method, respectively. We can see that R2CNN and RoI Transformer trained only with horizontal and vertical objects cannot almost detect oriented objcets, while our method can work well.

the unlabeled oriented objects to match the labeled objects in horizontal direction, thus can detect the oriented objects and generate bounding boxes for them. The proposed method automatically mines the pseudo labels for the oriented objects to train the Rotational RCNN in a weakly supervision. Experimental results reveal that our method outperforms state-of-the-art methods, and has obvious advantage on the oriented object detection.

5. ACKNOWLEDGMENT

This work is supported by the National Natural Science Foundation of China (No. 61571071, 61906025), Chongqing Research Program of Basic Research and Frontier Technology (No. cstc2018jcyjAX0227, cstc2020jcyj-msxmX0835), the Science and Technology Research Program of Chongqing Municipal Education Commission under Grant (No.KJQN201900607, KJQN202000647).

6. REFERENCES

[1] 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 *CVPR*, 2018, pp. 3974–3983.

[2] Yang Long, Yiping Gong, Zhifeng Xiao, and Qing Liu, "Accurate object localization in remote sensing images based on convolutional neural networks," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 5, pp. 2486–2498, 2017.

[3] Guoli Wang, Xinchao Wang, Bin Fan, and Chunhong Pan, "Feature extraction by rotation-invariant matrix representation for object detection in aerial image," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 6, pp. 851–855, 2017.

[4] Zhipeng Deng, Hao Sun, Shilin Zhou, Juanping Zhao, and Huanxin Zou, "Toward fast and accurate vehicle detection in aerial images using coupled region-based convolutional neural networks," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 10, no. 8, pp. 3652–3664, 2017.

[5] Jian Ding, Nan Xue, Yang Long, Gui-Song Xia, and Qikai Lu, "Learning roi transformer for oriented object detection in aerial images," in *CVPR*, 2019, pp. 2849–2858.

[6] Qingpeng Li, Lichao Mou, Qizhi Xu, Yun Zhang, and Xiao Xiang Zhu, "R ˆ3-net: A deep network for multi-oriented vehicle detection in aerial images and videos," *arXiv preprint arXiv:1808.05560*, 2018.

[7] Jiaming Han, Jian Ding, Jie Li, and Gui-Song Xia, "Align deep features for oriented object detection," *arXiv preprint arXiv:2008.09397*, 2020.

[8] Ross Girshick, "Fast r-cnn," in *ICCV*, 2015, pp. 1440–1448.

[9] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *NeurIPS*, 2015, pp. 91–99.

[10] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick, "Mask r-cnn," in *ICCV*, 2017, pp. 2961–2969.

[11] Guangting Wang, Zhiwei Xiong, Dong Liu, and Chong Luo, "Cascade mask generation framework for fast small object detection," in *ICME*, 2018, pp. 1–6.

[12] Yu Hao, Yanwei Fu, Yu-Gang Jiang, and Qi Tian, "An end-to-end architecture for class-incremental object detection with knowledge distillation," in *ICME*, 2019, pp. 1–6.

[13] Jiale Cao, Hisham Cholakkal, Rao Muhammad Anwer, Fahad Shahbaz Khan, Yanwei Pang, and Ling Shao, "D2det: Towards high quality object detection and instance segmentation," in *CVPR*, 2020, pp. 11485–11494.

[14] Jiaqi Wang, Kai Chen, Shuo Yang, Chen Change Loy, and Dahua Lin, "Region proposal by guided anchoring," in *CVPR*, 2019, pp. 2965–2974.

[15] Zenghui Zhang, Weiwei Guo, Shengnan Zhu, and Wenxian Yu, "Toward arbitrary-oriented ship detection with rotated region proposal and discrimination networks," *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 11, pp. 1745–1749, 2018.

[16] Zikun Liu, Hongzhen Wang, Lubin Weng, and Yiping Yang, "Ship rotated bounding box space for ship extraction from high-resolution optical satellite images with complex backgrounds," *IEEE Geoscience and Remote Sensing Letters*, vol. 13, no. 8, pp. 1074–1078, 2016.

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

[18] Zikun Liu, Jingao Hu, Lubin Weng, and Yiping Yang, "Rotated region based cnn for ship detection," in *ICIP*. IEEE, 2017, pp. 900–904.

[19] Yingying Jiang, Xiangyu Zhu, Xiaobing Wang, Shuli Yang, Wei Li, Hua Wang, Pei Fu, and Zhenbo Luo, "R 2 cnn: Rotational region cnn for arbitrarily-oriented scene text detection," in *ICPR*. IEEE, 2018, pp. 3610–3615.

[20] Zeming Li, Chao Peng, Gang Yu, Xiangyu Zhang, Yangdong Deng, and Jian Sun, "Light-head r-cnn: In defense of two-stage object detector," *arXiv preprint arXiv:1711.07264*, 2017.

[21] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *CVPR*, 2014, pp. 580–587.