Relational Matching for Weakly Semi-Supervised Oriented Object Detection

Wenhao Wu[1], Hau-San Wong[1*], Si Wu[2], and Tianyou Zhang[2]
[1]Department of Computer Science, City University of Hong Kong
[2]School of Computer Science and Engineering, South China University of Technology
wenhaowu5-c@my.cityu.edu.hk, cshswong@cityu.edu.hk, cswusi@scut.edu.cn,
cszhangtianyou@mail.scut.edu.cn

1. Overview

In the supplementary material, we present more experimental results and analysis as follows:

- We provide more details about the implementation of our proposed weakly semi-supervised method for oriented object detection.
- We confirm the effectiveness of the rotational modulation on the relational graph matching method.
- We conduct a category-wise comparison between our proposed method and the competing semi-supervised methods.
- We validate the extensibility of our proposed method to stronger base detectors, multi-scale strategy, and semi-supervised method.
- We qualitatively compare the detection results between the base detector training with direct prediction consistency and our proposed relation-based consistency.

2. Implementation Details

We adopt Faster R-CNN [6] as the base detector, and perform all experiments on MMRotate [11] without using multi-scale training and inference. Experiments are conducted on an NVIDIA V100 with 16G memory. For the proposed Rotation-Modulated Relational Graph Matching and Relational Rank Distribution Matching, we adopt the anchor setting with scales of $[4, 8, 16, 32, 64]$ and aspect ratios of $[0.5, 1.0, 2.0]$ to generate proposals centered on each annotated point, and thus N_k is set to 15 for the DOTA [7] and DIOR-R datasets [1]. The training on the annotated images is under the original view, while the training on point-annotated images is under the augmented views R_T for the teacher model and R_S for the student model, both of which are generated from the original view after random rotation with an angular range of $[-\pi/2, \pi/2]$.

*Corresponding author.

Table 1. Performance comparison between the base model with and without the rotational modulation on the validation set of DOTA-v1.0 under the partially-annotated setting with 5% of annotated images.

Modulated Node	Modulated Edge	mAP
		45.12
✓		46.18
	✓	46.65
✓	✓	48.78

3. Rotational Modulation for Relational Graph Matching

We further conduct a comparison between our proposed method with and without the rotational modulation on the relational graph matching. We perform the experiment without the introduction of the proposed Relational Rank Distribution Matching and weakly supervised learning methods for a fair comparison. The results are shown in Tab. 1. The rotational modulations on nodes and edges focus the matching process on the feature distribution and relational distribution between the predictions with significant deviations in orientation estimation between the teacher model and the student model. Therefore, separately applying the rotational modulation on node and edge matching can each contribute to performance improvement over the base model. With the collaboration of rotational modulations applied to both the node and edge matching, the performance of the base model can be enhanced to a significant degree.

4. Category-Wise Comparison

We provide a detailed category-wise comparison of detection performance between our proposed weakly semi-supervised method and the competing semi-supervised methods under the partially-annotated setting with 5% of the training images manually annotated. The results are

Table 2. Category-wise comparison between our proposed method and the competing semi-supervised methods on the validation set of DOTA-v1.0 under the partially-annotated setting.

Method	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	mAP
SOOD [4]	65.9	34.6	10.8	24.5	53.8	**66.1**	**81.9**	81.0	24.2	51.1	42.3	30.5	21.5	37.8	4.5	42.04
Group R-CNN [10]	60.9	60.0	16.7	46.3	**66.4**	61.8	78.3	89.2	**40.8**	**60.3**	20.3	62.2	33.1	**52.9**	19.5	51.24
Ours	**67.3**	**62.6**	**18.6**	**50.6**	65.4	63.8	79.5	**89.9**	39.4	58.8	**42.4**	**62.3**	**35.5**	49.8	**21.8**	**53.85**

shown in Tab. 2. Our proposed method surpasses SOOD [4], a semi-supervised method specified for oriented object detection, and Group R-CNN [10], a state-of-the-art weakly semi-supervised object detection method, to a significant extent on the evaluation metric of mAP. Specific to each category, our proposed method outperforms the competing methods over most categories, especially for the difficult categories of bridge (BR), ground-track-field (GTF), harbor (HA), and helicopter (HC). We attribute the success over these categories to several aspects: (1) The proposed Rotation-Modulated Relational Graph Matching and Relational Rank Distribution Matching methods succeed in transforming the ambiguous point information into spatial and semantic information, thus improving the recognition of the base model on difficult oriented objects in the complex aerial scenes. (2) The proposed weak supervised learning methods impose positive signals for difficult points to the base model and alleviate the inaccurate occupancy of regression outputs over densely packed oriented objects, both of which facilitate the base model to carefully explore the predicted bounding boxes with accurate categorical information for difficult oriented objects.

However, our proposed method performs slightly weakly in some categories compared to the competing methods. Specifically, our proposed method achieves a weaker performance in the category of swimming-pool (SP) compared to Group R-CNN. It is likely attributed to the random sampling strategy, in which the sampled annotated points are possibly located inside the oriented bounding boxes but outside the oriented objects due to the irregularity. For example, the annotated points of the HA category are possibly located in the seawater, which share a similar appearance as those of the SP category. Therefore, the deviated annotated points from various categories may have an indirect influence on the performance across categories. In addition, the proposed method achieves slightly lower performance in the categories of large vehicle (LV) and ship (SH) compared to SOOD. It may be the case that the significant scale variance of oriented objects in these categories leads to the preset anchors being unable to capture the circumscribed information for constructing reasonable relational graphs and rank distributions.

Table 3. Extensibility study of our proposed method on the stronger base models under the partially-annotated setting.

Method	Backbone	FS	WSS
Faster R-CNN [6]	ResNet-50 [3]	35.90	53.85
	Swin-T [5]	36.83	55.12
RoI Transformer [2]	ResNet-50 [3]	36.21	57.99
	Swin-T [5]	38.48	58.87

5. Extensibility Study

5.1. Extensibility to Stronger Detectors

We also explore the potential of extending the proposed method to stronger detectors. We take RoI Transformer [2] and Swin Transformer [5] based detectors as the stronger detectors, and implement our proposed method under the partially-annotated setting with 5% of box-annotated images. The results are shown in Tab. 3. Even the performance of RoI Transformer and Swin Transformer based detectors deteriorate when the annotated images become insufficient. However, with the point-annotated images introduced, all the base models achieve a significant performance improvement with our proposed relational matching methods and weakly supervised learning methods, confirming the extensibility of the proposed method to further enhance the performance of stronger detectors when given sufficient point-annotated images.

5.2. Extensibility to Multi-Scale Strategy

We introduce the multi-scale strategy, as in [9], into training and inference in DOTA-v1.0 to verify the effectiveness of the proposed method under the stronger augmentation. We implement the experiment under the fully-annotated setting, in which the images of the trainval set are treated as the annotated subset and the images of the trainval set from DOTA-v2.0 are treated as the point-annotated subset. The evaluation is performed on the test set through the online evaluation server. As shown in Tab. 4, the introduction of the multi-scale strategy obviously improves the performance of the base models. With the point-annotated images and our proposed method introduced, the base model can be further enhanced to a significant extent, confirming the effectiveness of our proposed method on the large-scale

Table 4. Extensibility study of our proposed method with multi-scale strategy under the fully-annotated setting.

Method	Type	mAP
Faster R-CNN [6]	FS	77.46
	WSS	78.34
RoI Transformer [2]	FS	79.66
	WSS	80.56

Table 5. Extensibility study of our proposed method on Soft Teacher under the partially-annotated setting.

Configs	5%	10%
Soft Teacher [8]	44.11	50.29
Soft Teacher [8] + Point Anno.	48.05	54.35
Soft Teacher [8] + Ours	52.50	57.17

dataset under the strong augmentation.

5.3. Extensibility to Semi-supervised Method

We introduce our proposed relational matching methods and weakly supervised learning methods into Soft Teacher to confirm the extensibility of the proposed method to other semi-supervised methods. The results are shown in Tab. 5. With the annotated points introduced, the pseudo supervision loss on the pseudo annotations of highest scores over the annotated points improves the performance of Soft Teacher to a limited extent. However, our proposed method can further enhance the detection performance of Soft Teacher under multiple semi-supervised settings, validating the effectiveness and strong extensibility of the relational matching methods and weakly supervised learning methods applied to the state-of-the-art semi-supervised frameworks.

6. Qualitative Results

We qualitatively compare the detection results between the base detector trained with direct prediction consistency and our proposed relation-based consistency under the partially-annotated setting, with 5% of the training images annotated, on the validation set of DOTA-v1.0. The qualitative comparison is shown in Fig. 1. The base detector achieves unsatisfactory performance on the localization and orientation estimation of the oriented objects when limited annotated images are provided. We then introduce the pseudo supervision on the pseudo annotations of the highest scores on the annotated points, which can be treated as the direct prediction consistency. The base model still fails to recognize the difficult instances, due to the inaccurate pseudo annotations induced by the ambiguous annotated points. However, our proposed weakly semi-supervised method, with the proposed relational matching methods to handle the ambiguity associated with point annotations and weakly supervised learning methods to alleviate the inaccurate classification and regression outputs, enhances the base model to spot difficult instances in challenging aerial scenes.

References

[1] Gong Cheng, Jiabao Wang, Ke Li, Xingxing Xie, Chunbo Lang, Yanqing Yao, and Junwei Han. Anchor-free oriented proposal generator for object detection. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–11, 2022. 1

[2] Jian Ding, Nan Xue, Yang Long, Gui-Song Xia, and Qikai Lu. Learning roi transformer for oriented object detection in aerial images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2849–2858, 2019. 2, 3

[3] 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*, pages 770–778, 2016. 2

[4] Wei Hua, Dingkang Liang, Jingyu Li, Xiaolong Liu, Zhikang Zou, Xiaoqing Ye, and Xiang Bai. Sood: Towards semi-supervised oriented object detection. *arXiv preprint arXiv:2304.04515*, 2023. 2

[5] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 2

[6] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 1, 2, 3

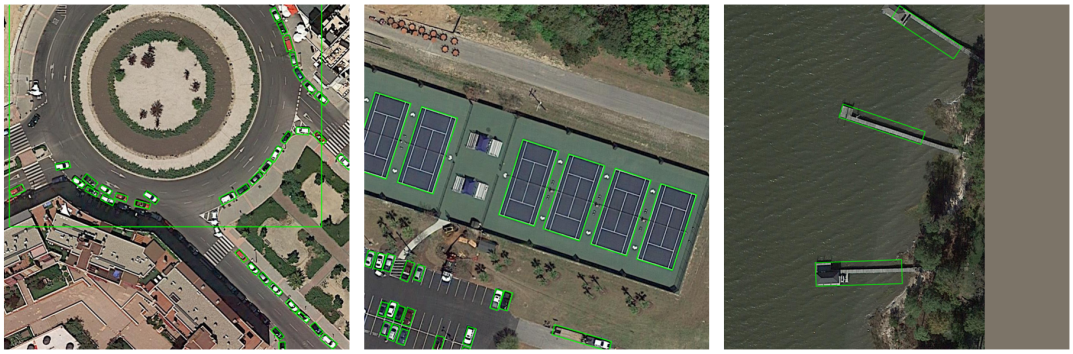
[7] Gui-Song Xia, Xiang Bai, Jian Ding, Zhen Zhu, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liangpei Zhang. Dota: A large-scale dataset for object detection in aerial images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3974–3983, 2018. 1

[8] Mengde Xu, Zheng Zhang, Han Hu, Jianfeng Wang, Lijuan Wang, Fangyun Wei, Xiang Bai, and Zicheng Liu. End-to-end semi-supervised object detection with soft teacher. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3060–3069, 2021. 3

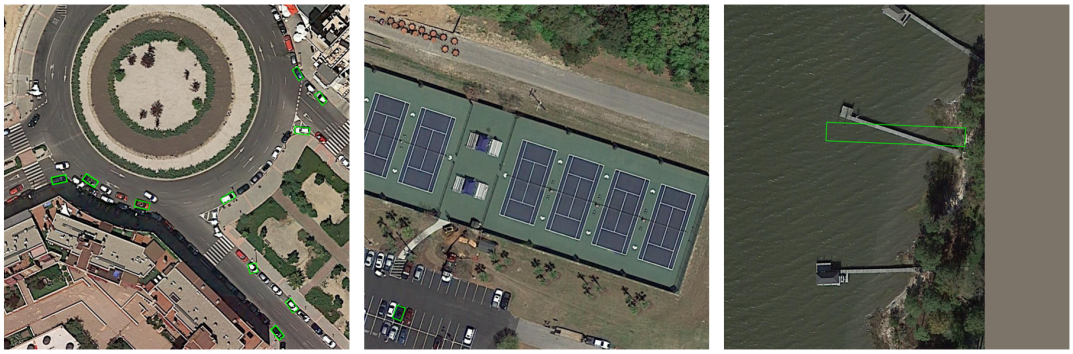
[9] Xue Yang, Gefan Zhang, Wentong Li, Yue Zhou, Xuehui Wang, and Junchi Yan. H2rbox: Horizontal box annotation is all you need for oriented object detection. In *The Eleventh International Conference on Learning Representations*, 2022. 2

[10] Shilong Zhang, Zhuoran Yu, Liyang Liu, Xinjiang Wang, Aojun Zhou, and Kai Chen. Group r-cnn for weakly semi-supervised object detection with points. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9417–9426, 2022. 2

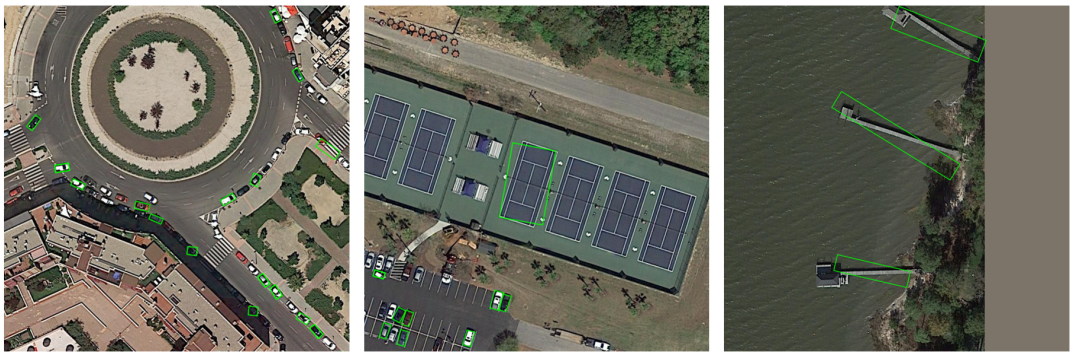
[11] Yue Zhou, Xue Yang, Gefan Zhang, Jiabao Wang, Yanyi Liu, Liping Hou, Xue Jiang, Xingzhao Liu, Junchi Yan, Chengqi



(a) Ground truths.



(b) Qualitative results of the base detector trained on only 5% of fully-annotated data.



(c) Qualitative results of the base detector trained with only prediction consistency on point-annotated images.

(d) Qualitative results of the base detector trained with our proposed weakly semi-supervised method.

Figure 1. Qualitative comparison of the base detector trained under different settings.

Lyu, et al. Mmrotate: A rotated object detection benchmark using pytorch. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 7331–7334, 2022. 1