PointOBB: Learning Oriented Object Detection via Single Point Supervision

Supplementary Material

A. More Details

Proposal Setting Details. To generate a bag of initial horizontal proposals from each point label, a set of basic scales $\{s_1, s_2, ..., s_G\}$ and a set of basic ratios $\{r_1, r_2, ..., r_R\}$ are employed, where G and R are the numbers of scales and ratios, respectively. Specifically, the basic scales are set to $\{4, 8, 16, 24, 32, 48, 64, 72, 80, 96\}$ and the basic ratios are set to $\{1/3, 1/2, 1/1.5, 1.0, 1.5, 2.0, 3.0\}$. In the refined Multiple Instance Learning (MIL) head, we follow the settings in [2], employing shake and random jitter to refine the selected horizontal proposals or rotated proposals from the first MIL head, with the shake ratio set to 0.1 and jitter ratios set to $\{1, 1.2, 1.3, 0.8, 0.7\}$. It is important to note that during the reproduction of P2BNet [2], all parameters remained consistent with the aforementioned.

Framework Structure Details. For the self-supervised angle learning branch in the angle acquisition module, we construct 4 parameter-shared 3x3 convolutional layers to process the dense features from the neck (*e.g.*, FPN [5]). Additionally, when employing the Dense-to-Dense assignment strategy for positive sample selection, for each level of feature maps, we set a circular region with a radius of 1.5 times the corresponding stride. All grid points within this region are selected as positive samples. To extract features from rotated proposals generated after Dense-to-Sparse matching or random rotation, rotated RoI extractors are employed to perform rotated RoI alignment.

Detailed Ablation Result. Tab. 1 presents the detailed results of Tab. 3 in the main paper.

B. Detailed Analysis of SSC Loss

In this section, we further analyze the inconsistency present in the MIL approach and the impact of SSC loss, providing additional visual analysis.

Inherent Instability of MIL. As mentioned in the main paper, the MIL fashion tends to focus on the most discriminative part of an object instead of its exact scale and boundary. This phenomenon is referred to as local focus. We have further observed that in aerial scenes, MIL fashion not only suffers from the local focus but also encounters the background prediction issues, the latter means that the predicted boxes may encompass the surrounding background of the object. As shown in Fig. 1, we visualize the prediction results of our method under the original MIL fashion without employing SSC loss. It can be observed that the confidence scores of candidate proposals closest to the ground truth boxes may not be the highest, resulting in inconsistencies between confidence scores and positional precision.

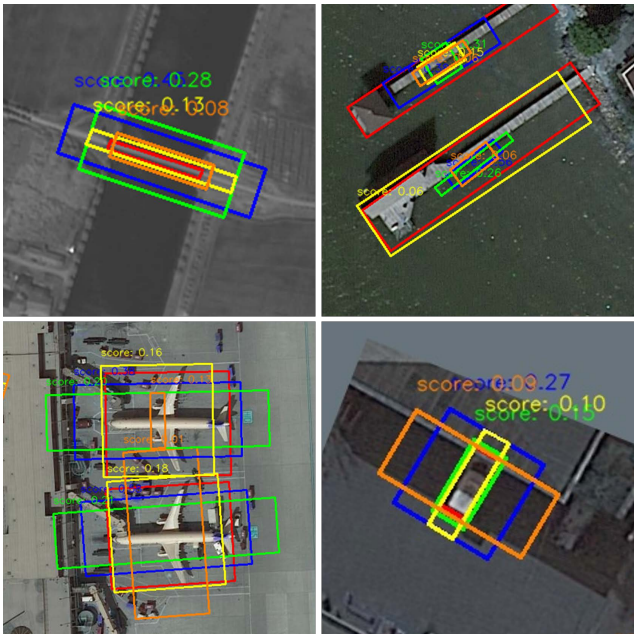

Figure 1. Local focus and background prediction problems in MIL fashion. The red bounding boxes indicate the ground-truth boxes. Bounding boxes with other different colors and the corresponding scores represent the top-k candidate proposals, along with their confidence scores (the product of class scores and instance scores). It can be observed that the candidate proposals with the highest confidence scores may not necessarily be the ones closest to the ground-truth boxes.

More Analysis of SSC Loss. To better showcase the score distributions before and after employing the SSC loss, we employ two-dimensional line charts in addition to the contour plot graph in the main paper.

Specifically, we visualize the score distributions grouped by scale and grouped by ratio, as shown in Fig. 2 and Fig. 3, respectively. From these two figures, we can observe that whether grouping the score distributions by scale or ratio for calculating the SSC loss, the consistency of score distributions between the original view and resized view (indicated by peaks of the same color) is enhanced.

Moreover, the visualization results align with the corresponding experimental results in the main paper. In the ablation study of the grouping type used in the SSC loss, the two grouping types (*i.e.*, ratio and scale) exhibit relatively close results (36.71% and 38.08% in mAP_{50} metric). This indicates that grouping by scale and ratio can both reflect consistency based on the variations in the distribution, whereas grouping by proposal results in a significant per-

SSC	DS	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	mAP_{50}
		28.60	59.36	2.75	52.95	63.75	35.89	28.73	1.44	7.73	15.91	1.38	45.50	18.61	36.86	27.20	28.44
✓		28.91	71.89	4.55	65.79	68.96	46.88	33.05	1.52	10.09	25.29	0.58	44.38	30.01	39.37	23.50	32.98
	✓	29.41	60.36	4.13	58.76	65.33	39.01	30.93	3.76	7.24	17.88	2.61	46.42	26.12	38.22	29.21	30.63
✓	✓	28.29	70.71	1.52	64.94	68.82	46.75	33.85	9.09	9.98	20.06	0.16	47.02	29.72	38.23	30.55	33.31

Table 1. Detailed ablation result on the DOTA-v1.0 testing set.

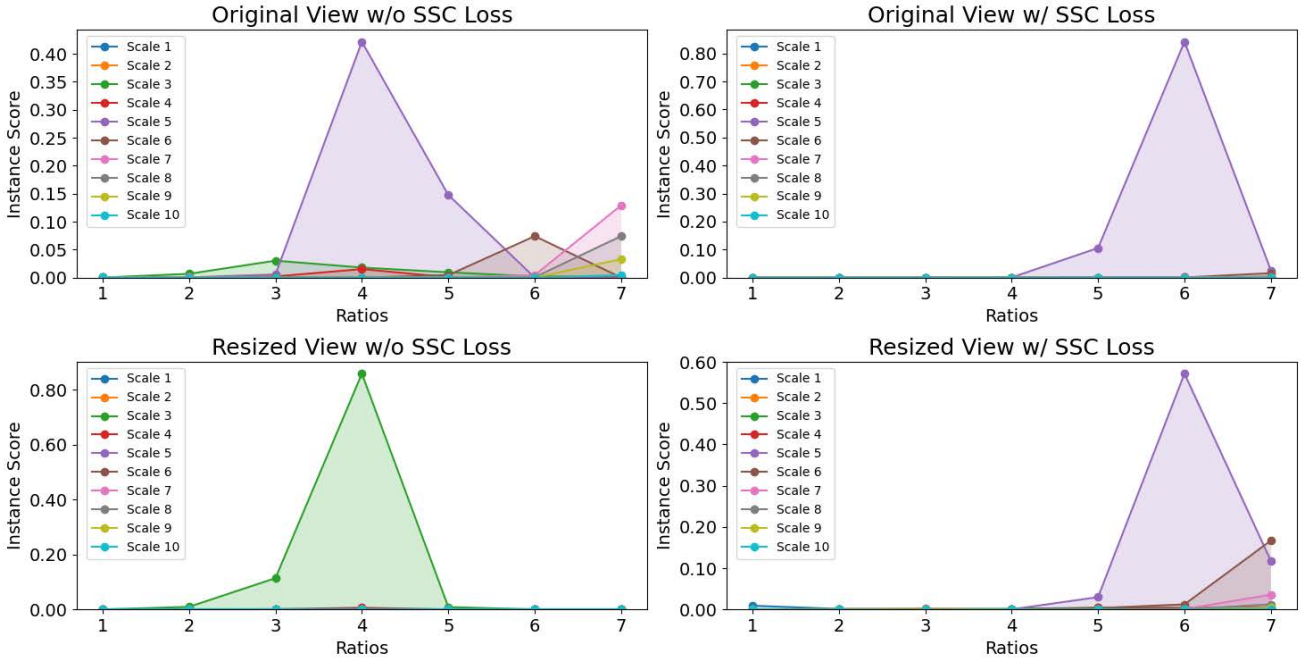

Figure 2. Line graphs of instance score distributions from the original and the resized views before and after employing the proposed SSC loss, the distributions are grouped by scales.

formance decline (30.42% in mAP_{50} metric). We believe that grouping by proposal implies forcibly aligning the confidence scores, neglecting the inherent differences between the two views. In short, grouping by ratio or scale represents that the discrete distribution of confidence scores from a specific dimension is beneficial for performance.

C. Impact of Different Pre-trained Models

In this section, we explore the impact of using different pre-trained models. To investigate whether increasing the volume of data can effectively enhance the MIL network's perceptual ability for object scales, we employ MS COCO [4] as the pre-training dataset and train our model on the general object detection task.

Experiment Settings. MS COCO is a large-scale dataset for common object detection with horizontal bounding boxes. Throughout the pre-training process, we don't create additional enhanced views (*i.e.*, resized view and rotated/flipped view), solely employing the original view for training in the horizontal object detection task. As the parameters in the MIL heads are shared among the three views, the overall network can acquire fundamental perceptual capabilities from the pre-training. During pre-training on the MS COCO dataset, we follow the "1x" training schedule in MMDetection [1], setting the learning rate to 0.02, while keeping the remaining hyper-parameter unchanged.

Experiment Results. As illustrated in Tab. 2 and Tab. 3, after employing the COCO pre-trained model, our approach demonstrates improved performance, reaching 42.50% on the DIOR-R dataset and 46.22% on the DOTA-v1.0 dataset, surpassing the ImageNet pre-trained model by **+4.42**% and **+12.91**%, respectively. This demonstrates the immense potential of our approach.

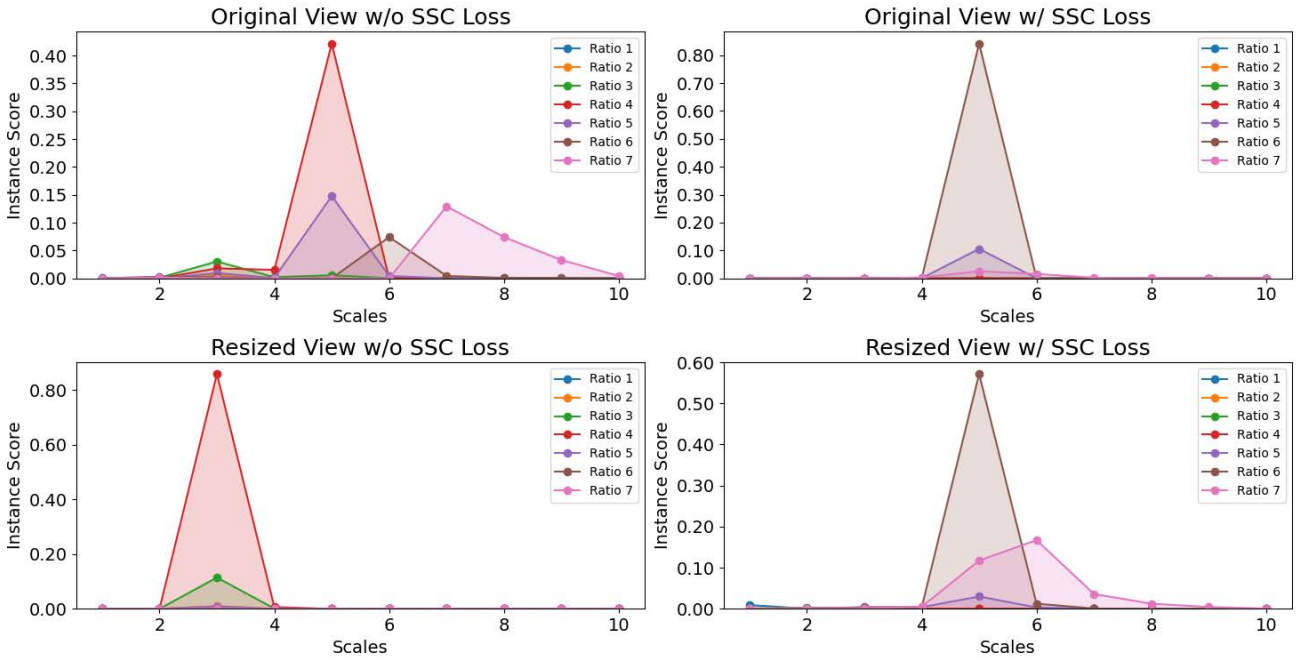

Figure 3. Line graphs of instance score distributions from the original and the resized views before and after employing the proposed SSC loss, the distributions are grouped by scales.

Method	APL	APO	BF	BC	BR	CH	ESA	ETS	DAM	GF	GTF	HA	OP	SH	STA	STO	TC	TS	VE	WM	8-mAP$_{50}$	mAP$_{50}$
Ours (Oriented R-CNN) [3]	**58.2**	**15.3**	70.5	**78.6**	**0.1**	72.2	**69.6**	1.8	**3.7**	0.3	77.3	16.7	**4.0**	**79.2**	39.6	51.7	44.9	**16.8**	33.6	**27.7**	57.38	38.08
Ours* (Oriented R-CNN) [4]	57.9	8.7	**86.1**	57.1	0	**74.7**	**69.6**	**13.5**	2.5	**68.5**	**85.7**	**15.0**	1.0	68.8	**55.6**	**55.6**	**68.5**	6.1	**43.4**	12.3	**60.31**	**42.50**

Table 2. Results of employing different pre-trained models on the DIOR-R testing set. 'Ours" indicates employing backbone pre-trained on the ImageNet. "Ours*" indicates employing the basic network pre-trained on the MS COCO.

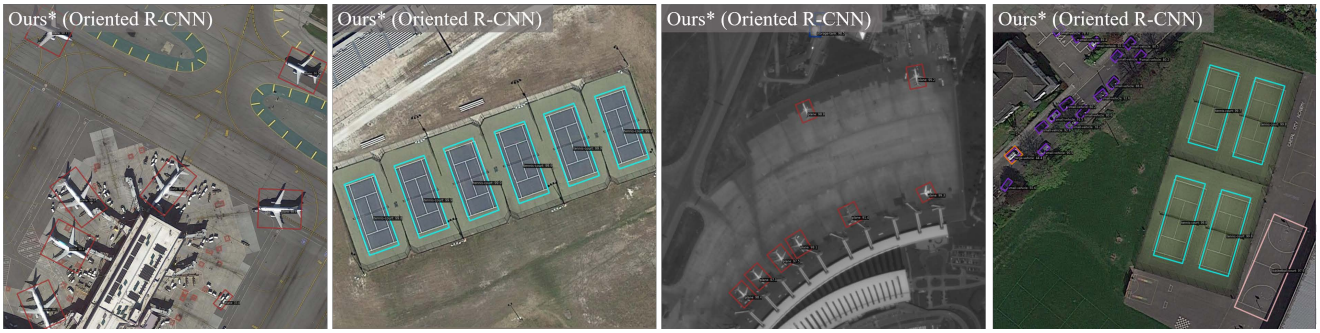

Figure 4. The visual detection results. "Ours*" indicates employing the pre-trained weights on the MS COCO dataset.

Visual Results. For the experimental results employing COCO pre-trained weights, we supplement additional visualization results to illustrate the improvements it brings to prediction accuracy, as shown in Fig. 4.

References

[1] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, et al. Mmdetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019.

Method	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	mAP_{50}
RBox-supervised:																
Rotated RetinaNet [6]	88.7	77.6	38.8	58.2	74.6	71.6	79.1	88.0	80.2	72.3	52.8	58.6	62.6	67.7	59.6	68.69
Rotated FCOS [8]	88.4	76.8	45.0	59.2	79.2	79.0	86.9	88.1	76.6	78.8	58.6	57.5	69.3	72.4	53.5	**71.28**
HBox-supervised:																
Sun et al. [7]	51.5	38.7	16.1	36.8	29.8	19.2	23.4	83.9	50.6	80.0	18.9	50.2	25.6	28.7	25.5	38.60
H2RBox [9]	88.5	73.5	48.8	56.9	77.5	65.4	77.9	88.9	81.2	79.2	55.3	59.9	52.4	57.6	45.3	67.21
H2RBox-v2 [10]	89.0	74.4	51.0	60.5	79.8	75.3	86.9	90.9	86.1	85.0	59.2	63.2	65.2	71.6	49.7	**72.52**
Point-supervised:																
P2BNet [2] + H2RBox [9]	24.7	35.9	7.0	27.9	3.3	12.1	17.5	17.5	0.8	**34.0**	6.3	49.6	11.6	27.2	18.8	19.63
P2BNet [2] + H2RBox-v2 [10]	11.0	44.8	**14.9**	15.4	36.8	16.7	27.8	12.1	1.8	31.2	3.4	50.6	12.6	36.7	12.5	21.87
Ours (Rotated FCOS)	26.1	65.7	9.1	59.4	65.8	34.9	29.8	0.5	2.3	16.7	0.6	49.04	21.8	41.0	**36.7**	30.08
Ours (Oriented R-CNN)	28.3	70.7	1.5	**64.9**	**68.8**	**46.8**	33.9	9.1	10.0	20.1	0.2	47.0	**29.7**	38.2	30.6	33.31
Ours* (Oriented R-CNN)	**75.7**	**80.4**	9.1	55.9	58.5	4.1	**52.0**	**80.7**	**71.5**	28.6	**52.0**	**54.8**	1.0	**46.0**	22.9	**46.22**

Table 3. Complete results along with the result of employing COCO pre-trained model on the DOTA-v1.0 testing set. The categories in DOTA-v1.0 include Plane (PL), Baseball Diamond (BD), Bridge (BR), Ground Track Field (GTF), Small Vehicle (SV), Large Vehicle (LV), Ship (SH), Tennis Court (TC), Basketball court (BC), Storage Tank (ST), Soccer Ball Field (SBF), Roundabout (RA), Harbor (HA), Swimming Pool (SP), Helicopter (HC). "Ours" indicates employing backbone pre-trained on the ImageNet. "Ours*" indicates employing the basic network pre-trained on the MS COCO.

2

[2] Pengfei Chen, Xuehui Yu, Xumeng Han, Najmul Hassan, Kai Wang, Jiachen Li, Jian Zhao, Humphrey Shi, Zhenjun Han, and Qixiang Ye. Point-to-box network for accurate object detection via single point supervision. In *European Conference on Computer Vision*, pages 51–67. Springer, 2022. 1, 4

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. Ieee, 2009. 3

[4] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pages 740–755. Springer, 2014. 2, 3

[5] 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*, pages 2117–2125, 2017. 1

[6] 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*, pages 2980–2988, 2017. 4

[7] Yongqing Sun, Jie Ran, Feng Yang, Chenqiang Gao, Takayuki Kurozumi, Hideaki Kimata, and Ziqi Ye. Oriented object detection for remote sensing images based on weakly supervised learning. In *IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pages 1–6. IEEE, 2021. 4

[8] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9627–9636, 2019. 4

[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 *International Conference on Learning Representations*, 2022. 4

[10] Yi Yu, Xue Yang, Qingyun Li, Yue Zhou, Gefan Zhang, Junchi Yan, and Feipeng Da. H2rbox-v2: Boosting hbox-supervised oriented object detection via symmetric learning. *Advances in Neural Information Processing Systems*, 2023. 4