Supplementary Materials

Yu Zheng[1,2,3], Danyang Zhang[1], Sinan Xie[1], Jiwen Lu[1,2,3*], and Jie Zhou[1,2,3,4]

[1] Department of Automation, Tsinghua University, China
[2] State Key Lab of Intelligent Technologies and Systems, China
[3] Beijing National Research Center for Information Science and Technology, China
[4] Tsinghua Shenzhen International Graduate School, Tsinghua University, China
{zhengyu19,zhang-dy16,xsn18}@mails.tsinghua.edu.cn
{lujiwen,jzhou}@tsinghua.edu.cn

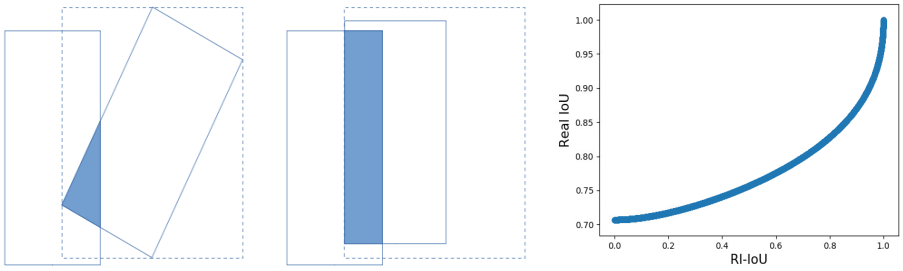
1 More Analysis of RIoU

The necessity of the cosine coefficient. We already show the necessity of the cosine coefficient in (4) in the ablation experiments. Here we further demonstrate the necessity through an illustration example, as shown in Fig. 1(a). We show 2 pairs of bounding boxes with rotation (in solid lines). As the intersection aera of the left pair is smaller than that of the right pair, the left should be penalized more. However, if we remove the cosine coefficient in (4):

$$I_{RIoU}^{'} = \min(I1, I2) \tag{1}$$

the I_{RIoU} of the left is larger than right (we draw the preserved projected rectangle of the left pair after the min function in the dashed line), hence the $\mathcal{L}_{RIoU}$ of the left is smaller than right, which is contradictory to the situation of the real IoU. After imposing the cosine coefficient, the error is mitigated.

The consistency between RIoU and the real IoU. We ran 5000 numerical simulations of 2 identical rotation-free squares centered at the same point. We set the fluctuation rate of each parameter as 10%. As shown in Figure 1(b),



(a) The inconsistency after removing the cosine coefficient. The rectangles in the solid line are of the same shape.

(b) Average IoU curve

Fig. 1. Illustration of the analysis.

Table 1. Comparisons of different loss settings for object detection on the DOTA v1.0 dataset. The abbreviation of each category is defined as follows: PL-Plane, BD-Baseball diamond, BR-Bridge, GTF-Ground field track, SV-Small vehicle, LV-Large vehicle, SH-Ship, TC-Tennis court, BC-Basketball court, ST-Storage tank, SBF-Soccer-ball field, RA-Roundabout, HA-Harbor, SPSwimming pool, and HC-Helicopter.

Method	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	mAP
$\ell1$(baseline)	89.17	71.57	43.70	63.45	65.07	56.78	66.86	**90.80**	78.80	**78.36**	54.54	62.99	58.55	67.16	50.89	66.58
$\ell1$+*RIoU*	89.14	72.80	**44.42**	**67.12**	**65.72**	**64.40**	**67.96**	**90.80**	79.43	77.73	**56.82**	**64.02**	62.47	**68.64**	**55.98**	**68.50**
$\ell1$+*RGIoU*	**89.19**	**73.36**	44.07	63.75	65.40	64.09	67.60	90.79	**82.65**	78.29	56.19	62.66	**63.07**	67.15	52.95	68.08

Table 2. Comparisons of the detection for aerial images with or without the RIoU loss. The baseline model is DrBox-v2 [1].

Method	Airplane		Car		Ship	
	AP	BEP	AP	BEP	AP	BEP
$\ell1$(baseline)	63.65	69.29	66.21	70.04	83.55	**81.63**
$\ell1$+iou	**63.95**	**71.42**	**66.72**	**71.24**	**83.90**	80.64

the scatter plot implies the consistency between the real IoU and our RIoU, so it is reasonable to adopt RIoU as the optimization target. The consistency is preserved during training, because the angle difference is relatively small only after 10 epochs, with real IoU larger than 0.77 and angle difference less than 10 degree.

2 More Experimental results of RIoU

2.1 Aerial Image Detection

We did the experiments of the aerial image detection on 2 released datasets. The DOTA v1.0 [13] dataset contains 2806 large-size aerial images for small oriented aerial object detection. The target instances are divided into 15 different categories, including *Tennis court*, *Swimming pool*, etc. We used a modified RetinaNet [7, 14] as our baseline method. The backbone network was ResNet-50 [4], where the horizontal anchors were generated in the first stage of the RetinaNet. The baseline method employs smooth-$\ell1$ loss function. We incorporated it with the proposed *RIoU* or *RGIoU* loss.

The results are summarized in Table 1. Compared with the baseline method, the detection performance is better except for the Storage tank (ST) category, and the mAP of the Tennis court category is the same with the baseline method. For more experiment validation results, please refer to the supplementary pages.

We then validated RIoU on the remote sensing images of GoogleEarth collected by [8]. The dataset contains three types of targets: vehicles, ships and airplanes. As the raw dataset is too large for training, we randomly sampled 1000 training images and 200 testing images for vehicle, 5000 and 1000 for ship, 5000 and 1000 for airplane. We selected the DRBox-v2 [1] as our baseline method. The evaluation metric as average precision (AP) and break-even point (BEP).

As shown in Table 2, except for the BEP metric of the ship category, the network achieves better detection performance when incororated with $\mathcal{L}_{RIoU}$.

2.2 KITTI 3D Object Detection

Table 3. Comparisons of different loss settings on the KITTI validation set. The baseline model is Frustum-PointNet [9] (F-PointNet). The backbone network is PointNet [10](v1) and PointNet++ [11](v2) respectively. The results are reported on the task of **3D object localization** (Loc.) and **3D object detection** (Det.).

Method		Pedestrians						Cyclists					
		Easy		Moderate		Hard		Easy		Moderate		Hard	
		v1	v2	v1	v2	v1	v2	v1	v2	v1	v2	v1	v2
Loc.	ℓ1(baseline)	70.65	72.38	61.22	66.39	53.46	59.57	81.79	81.82	59.94	60.03	56.15	56.32
	ℓ1+ArIoU	70.54	71.24	61.33	65.34	53.70	58.77	76.71	79.79	56.96	59.45	53.05	56.10
	ℓ1+*RIoU*	73.09	76.66	66.23	68.97	58.62	61.31	**82.29**	82.77	61.83	61.06	**57.53**	58.33
	ℓ1+*RGIoU*	**77.01**	**77.16**	**67.61**	**69.29**	**59.70**	**61.55**	77.71	**83.22**	59.55	**62.59**	56.15	**58.47**
Det.	ℓ1(baseline)	66.73	70.00	56.91	61.32	49.82	53.59	76.38	77.15	55.18	56.49	50.97	53.37
	ℓ1+ARIoU	63.27	67.29	54.97	59.19	47.95	52.03	71.35	77.02	52.39	56.87	48.51	53.16
	ℓ1+*RIoU*	69.72	71.57	60.02	65.22	52.44	57.28	**78.45**	79.98	**57.96**	58.66	**53.94**	55.01
	ℓ1+*RGIoU*	**70.71**	**71.65**	**61.05**	**65.29**	**53.22**	**57.46**	74.70	**83.30**	55.01	**59.07**	**53.94**	**55.06**

The experimental results of 3D object detection on the pedestrain and cyclist categories of the KITTI [3] validation dataset are summarized in Table 3. We can see that both RIoU and RGIoU loss improves the detection mAP by an obvious margin, while the performance of $\mathcal{L}_{RGIoU}$ is comparably better. Note that $\mathcal{L}_{RGIoU}$ lower the cyclists detection performance under PointNet backbone network. This is partially due to fewer instances of cyclists in the KITTI dataset, which could make the AP metric unstable. And the rich local feature encoded by PointNet++ [11] backbone network mitigates such effect.

2.3 Nuscenes 3D Object Detection

We jointly trained PointPillars [6] on a 9-class subset of the nuScenes dataset [2]. The subset includes *Car, Bicycle, Bus, Construction Vehicle, Motorcycle, Pedestrian, Traffic Cone, Trailer, Truck*. The per-class results are summarized in Table. 4. Note that we omitted the result of *Bicycle* class because all AP values are zero. We calculated the average AP with regard to the Center Distance (D). As shown in Table. 5, the $\mathcal{L}_{RI\text{-}IoU}$ still outperforms all the other methods. Due to the positive matching protocol of the PointPillars[6], fewer samples are penalized by the $\mathcal{L}_{RI\text{-}GIoU}$, which is similar to the situation of single class training.

3 More Visualized Results

We show the visualized comparison of predicted results on the nuScenes dataset (predictions in red, ground-truths in green) in Figure 2, and the SUN RGB-D

Table 4. Comparisons of different loss settings for **3D object detection** on a 9-class subset of the nuScenes validation set. The Average Precision (AP) metric of 3D object detection is based on the matching of center distance(D). The baseline model is PointPillars [6]. The *Bicycle* class is omitted because all the AP values are zero.

Method	Car		Bus		Construction		Motorcycle	
	D=2.0	D=4.0	D=2.0	D=4.0	D=2.0	D=4.0	D=2.0	D=4.0
$\ell 1$(baseline)	**76.83**	**78.91**	37.25	40.08	0.52	1.59	14.93	15.11
$\ell 1$+giou	75.86	78.24	**39.16**	**43.33**	**2.80**	**5.25**	17.19	17.65
$\ell 1$+iou	75.87	78.36	38.30	42.64	1.76	4.10	**19.88**	**20.23**
Method	Pedestrian		Traffic Cone		Trailer		Truck	
	D=2.0	D=4.0	D=2.0	D=4.0	D=2.0	D=4.0	D=2.0	D=4.0
$\ell 1$(baseline)	65.01	67.15	14.98	**20.75**	**17.73**	22.84	28.02	31.33
$\ell 1$+giou	64.68	66.95	**15.00**	19.40	11.00	20.06	**30.06**	33.59
$\ell 1$+iou	**65.41**	**67.61**	14.28	18.89	13.00	**25.40**	29.33	**33.68**

Table 5. The average AP with regard to the center distance (D) in the Table. 4

Method	$Avg\%D = 2.0$	$Avg\%D = 4.0$
$\ell 1$(baseline)	31.94	34.72
$\ell 1$+giou	31.98	35.56
$\ell 1$+iou	**32.23**	**36.36**

dataset (predictions in green, ground-truths in red). In the examples, the inaccuracy of localization exists in the **upper** baseline results. When incorporated with our $\mathcal{L}_{RI\text{-}IoU}$ (**bottom**), the localization performance is notably improved.

We illustrate the experimental results on the ICDAR2015 [5] dataset. The baseline method is EAST [15]. As shown in Figure 4, when incorporated with $RIoU$ loss function, the missing (in the pink circles), inaccurate and merged (in the red circle) detection instances in the baseline are mitigated.

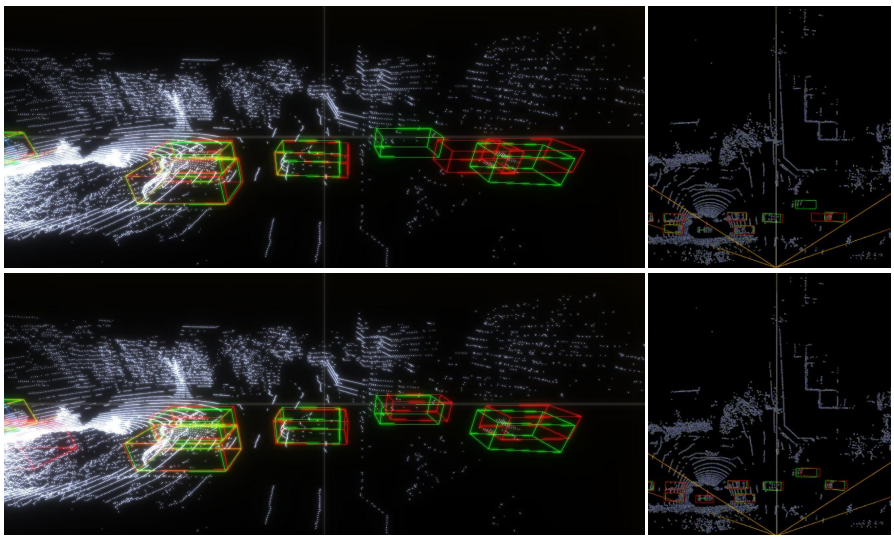


Fig. 2. A visualized comparison of predicted results on the nuScenes (predictions in red, ground-truths in green) before NMS of baseline with or without our proposed loss function. Bounding boxes with confidence score larger than 0.3 are preserved. The detection results are from the single class training of PointPillars. (*best viewed in color pdf file*)

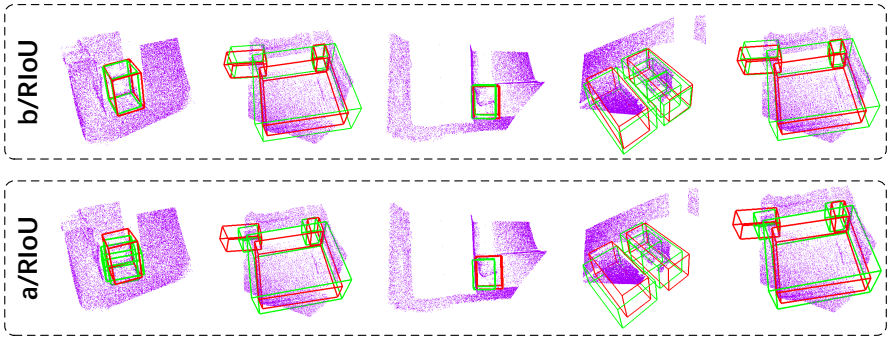


Fig. 3. Visualized comparison of predicted results on the SUN RGB-D [12] dataset (predictions in green, ground-truths in red). The detection results are from VoteNet. (*best viewed in color pdf file*)

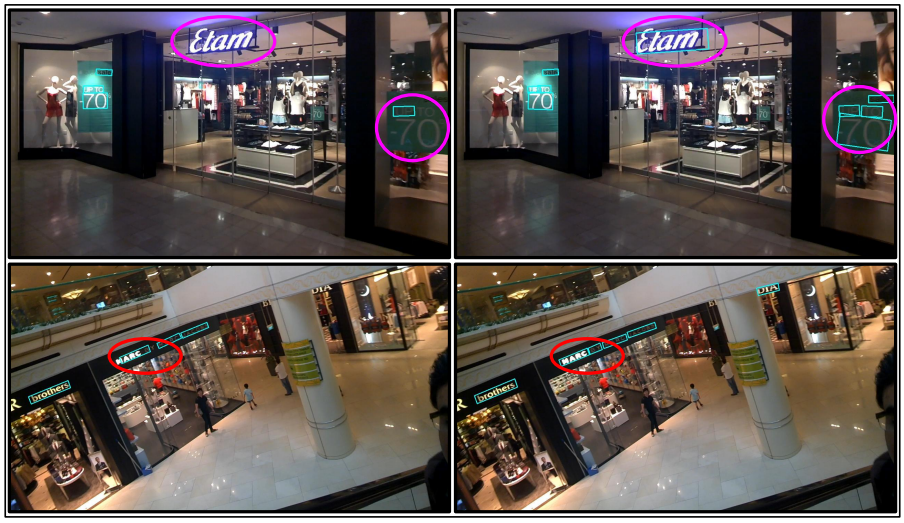


Fig. 4. Visualized comparision of EAST [15] with (right) or without (left) the proposed *RIoU* loss funciton. (*best viewed in color pdf file*)

References

1. An, Q., Pan, Z., Liu, L., You, H.: Drbox-v2: An improved detector with rotatable boxes for target detection in sar images. IEEE Transactions on Geoscience and Remote Sensing **57**(11), 8333–8349 (2019)
2. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuscenes: A multimodal dataset for autonomous driving. arXiv preprint arXiv:1903.11027 (2019)
3. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: CVPR. pp. 3354–3361 (2012)
4. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016)
5. Karatzas, D., Gomez-Bigorda, L., Nicolaou, A., Ghosh, S., Bagdanov, A., Iwamura, M., Matas, J., Neumann, L., Chandrasekhar, V.R., Lu, S., et al.: Icdar 2015 competition on robust reading. In: ICDAR. pp. 1156–1160 (2015)
6. Lang, A.H., Vora, S., Caesar, H., Zhou, L., Yang, J., Beijbom, O.: Pointpillars: Fast encoders for object detection from point clouds. In: CVPR. pp. 12697–12705 (2019)
7. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: ICCV. pp. 2980–2988 (2017)
8. Liu, L., Pan, Z., Lei, B.: Learning a rotation invariant detector with rotatable bounding box. arXiv preprint arXiv:1711.09405 (2017)
9. Qi, C.R., Liu, W., Wu, C., Su, H., Guibas, L.J.: Frustum pointnets for 3d object detection from rgb-d data. In: CVPR. pp. 918–927 (2018)
10. Qi, C.R., Su, H., Mo, K., Guibas, L.J.: Pointnet: Deep learning on point sets for 3d classification and segmentation. In: CVPR. pp. 652–660 (2017)
11. Qi, C.R., Yi, L., Su, H., Guibas, L.J.: Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In: NIPS. pp. 5099–5108 (2017)
12. Song, S., Lichtenberg, S.P., Xiao, J.: Sun rgb-d: A rgb-d scene understanding benchmark suite. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 567–576 (2015)
13. Xia, G.S., Bai, X., Ding, J., Zhu, Z., Belongie, S., Luo, J., Datcu, M., Pelillo, M., Zhang, L.: Dota: A large-scale dataset for object detection in aerial images. In: CVPR. pp. 3974–3983 (2018)
14. Yang, X., Liu, Q., Yan, J., Li, A.: R3det: Refined single-stage detector with feature refinement for rotating object. arXiv preprint arXiv:1908.05612 (2019)
15. Zhou, X., Yao, C., Wen, H., Wang, Y., Zhou, S., He, W., Liang, J.: East: an efficient and accurate scene text detector. In: CVPR. pp. 5551–5560 (2017)