Learning Object-level Point Augmentor for Semi-supervised 3D Object Detection

Cheng-Ju Ho*[1]
ace52751208@gmail.com
Chen-Hsuan Tai*[1]
derek.0417t@gmail.com
Yi-Hsuan Tsai‡[2]
wasidennis@gmail.com
Yen-Yu Lin[1]
lin@cs.nycu.edu.tw
Ming-Hsuan Yang[3]
mhyang@ucmerced.edu

[1] National Yang Ming Chiao Tung University, Taiwan
[2] Phiar Technologies, United States
[3] University of California at Merced, United States

1 Supplementary Material

This document provides more details and analysis of our proposed method, Object-level Point Augmentor (OPA), and is arranged as follows: We elaborate the experiments using pre-defined augmentations in Section 1.1. The per-class mAP scores of the detector trained with OPA are reported in Section 1.2. We evaluate the performance of the pre-trained detector with the proposed augmentor in Section 1.3. We analyze the displacement distribution in the form of histograms in Sec. 1.4. Finally, the qualitative visualizations are shown in Section 1.5.

1.1 Details of Pre-defined Augmentation Experiments

Table 1 of the main submission reports the results of using pre-defined object-level augmentations. A more detailed description of this experiment is given below. We test the object-level augmentation with 1) three operations, including scale, flip, and rotation, and 2) only displacement. For using the three operations, we follow the same augmentation operations that are used in scene-level augmentations but only apply them to the foreground points with a smaller magnitude. Specifically, we randomly scale an object between 0.85 and 1.15 times the original size, randomly rotate the object from -5 to 5 degrees, and randomly horizontal flip the object with the probability of 0.5. For the experiment using only displacement under a certain range α, we randomly jitter each point in the x-direction, y-direction, and z-direction with a $-\alpha\%$ to $\alpha\%$ ratio of displacement with respect to its corresponding bounding box size.

*The authors contribute equally to this paper
‡Currently at Google

Table 1: Per-class mAP@0.25 (top group) and mAP@0.5 (bottom group) on the ScanNet val set with 10% labeled data.

	cabin	bed	chair	sofa	table	door	wind	bkshf	pic	cntr	desk	curt	refrig	showr	toilet	sink	bath	ofurn
VoteNet [0]	17.9	74.7	74.5	75.3	45.6	18.3	11.7	21.7	0.7	28.4	49.4	21.5	23.2	18.5	79.6	25.7	66.3	11.7
SESS [0]	20.5	75.1	76.2	76.4	48.1	20.0	14.4	19.4	1.2	30.0	51.8	25.0	30.0	26.4	82.2	29.2	72.3	14.1
3DIoUMatch [0]	26.6	82.6	80.9	**83.3**	52.1	28.0	19.9	29.4	3.7	45.0	**61.9**	29.2	34.1	51.2	85.7	32.3	82.8	21.5
OPA	**27.4**	**85.7**	**81.8**	79.6	**53.8**	**32.4**	**27.6**	**37.6**	**5.3**	**53.5**	58.2	**35.7**	**35.2**	**57.1**	**94.4**	**33.3**	**86.3**	**26.1**
VoteNet [0]	3.2	64.6	43.4	49.3	25.1	2.8	1.1	8.7	0.0	2.4	14.7	3.9	7.6	1.1	46.8	11.9	39.4	1.5
SESS [0]	3.7	61.2	48.0	44.8	29.5	3.2	2.8	8.4	0.2	7.5	19.2	5.0	12.2	1.8	48.7	15.3	40.8	2.2
3DIoUMatch [0]	5.9	72.0	60.5	56.6	39.7	10.3	5.2	18.1	**0.7**	16.0	**35.3**	8.3	21.4	6.2	67.5	13.2	67.6	5.2
OPA	**6.6**	**72.8**	**64.2**	**66.2**	**41.2**	**10.7**	**9.7**	**23.4**	0.1	**20.0**	**35.3**	**16.2**	**23.2**	**15.9**	**90.4**	**20.4**	**83.1**	**11.7**

Table 2: Per-class mAP@0.25 (top group) and mAP@0.5 (bottom group) on the SUNRGB-D val set with 5% labeled data.

	bathtub	bed	bookshelf	chair	desk	dresser	nightstand	sofa	table	toilet
VoteNet [0]	67.8	32.2	39.4	58.5	53.5	8.0	1.9	14.7	3.2	20.3
SESS [0]	70.8	34.7	41.9	60.4	63.0	**9.8**	3.7	25.2	4.0	28.0
3DIoUMatch [0]	75.4	37.7	45.2	**64.2**	77.0	6.0	**5.7**	34.6	**4.5**	39.4
OPA	**77.1**	**39.6**	**47.7**	63.4	**81.1**	8.2	4.8	**44.7**	3.6	**49.2**
VoteNet [0]	31.2	6.2	15.5	29.6	14.6	0.5	0.2	2.0	0.3	5.2
SESS [0]	36.7	7.2	19.2	31.8	20.4	0.7	0.5	7.0	0.4	7.1
3DIoUMatch [0]	45.2	**14.4**	27.8	**43.6**	47.2	0.8	**1.9**	15.7	**0.6**	13.4
OPA	**45.6**	14.1	**30.9**	41.8	**52.4**	**1.1**	0.6	**25.3**	0.2	**20.7**

1.2 Per-class Evaluation

We report the per-class average precision on the val set of ScanNet [1] with 10% labeled data and of SUN RGB-D [3] with 5% labeled data in Table 1 and Table 2, respectively. The results in Table 1 show that our method improves the performance in almost all the classes. Similarly, Table 2 indicates that our model achieves more favorable results in most classes. Overall, our method achieves better performance against the 3DIoUMatch [4], the best competing method.

1.3 Pre-trained Detector

We show that jointly training the detector and our proposed augmentor in the pre-training stage can improve the performance. We compare our method with 3DIoUMatch [4] on both ScanNet [1] and SUN RGB-D [3] with 5%, 10%, and 20% labeled training data. The results in Table 3 point out that OPA results in larger performance gains when few training data are available, such as 10% labeled training data on ScanNet and 5% labeled training data on SUN RGB-D.

1.4 Displacements Analysis

We analyze how OPA augments the points along x-direction, y-direction, and z-direction. The three histograms in Figure 1 display the displacement distributions along the three directions, respectively. The x-axis in each histogram represents the ratio of the absolute displacement to its corresponding bounding box dimension while the y-axis gives the frequency. Here, we only count points with displacement ratios larger than 1%, which are considered

Table 3: We report the pre-trained model performance of OPA and 3DIoUMatch with different amounts of labeled training data.

Dataset	Model	5%		10%		20%	
		mAP @0.25	mAP @0.5	mAP @0.25	mAP @0.5	mAP @0.25	mAP @0.5
ScanNet	3DIoUMatch [4]	29.5	13.6	40.6	20.8	47.4	29.1
	OPA	**33.3**	**16.2**	**45.8**	**26.1**	**50.2**	**31.8**
	Gain (%)	3.8↑	2.6↑	5.2↑	5.3↑	2.8↑	2.7↑
SUN RGB-D	3DIoUMatch [4]	31.0	14.5	41.5	21.4	48.0	**26.6**
	OPA	**36.1**	**16.3**	**44.4**	**23.8**	**48.5**	**26.6**
	Gain (%)	5.1↑	1.8↑	2.9↑	2.4↑	0.5↑	0

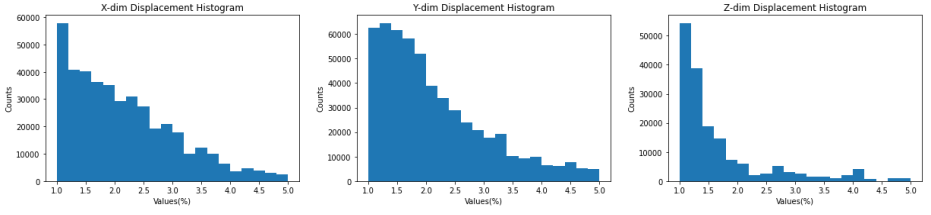

Figure 1: Histograms of the displacement distributions of the augmented points along x-direction (left), y-direction (middle), and z-direction (right).

significant augmentation. This figure shows that our augmentor learns to produce more realistic augmented data. For example, most points are augmented with displacement ratios less than 3% along x-direction and y-direction while less than 1.5% along z-direction. It is the expected result since all objects should be placed on the ground or on top of another object due to gravity. It implies that our augmentor can appropriately augment objects by increasing data variance and maintaining object distinctiveness at the same time.

1.5 Qualitative Visualization

We show qualitative results on the validation set of the model using ScanNet [1] with 10% labeled data in Figure 2 and on the validation set using SUN RGB-D [3] with 5% labeled data in Figure 3. In the results, the green bounding boxes in the scenes indicate proposals with IoU ≥ 0.25, and the red bounding boxes denote proposals with IoU < 0.25. Overall, our method predicts the objects more precisely compared to ground truths and can locate the objects that are highly occluded by other objects.

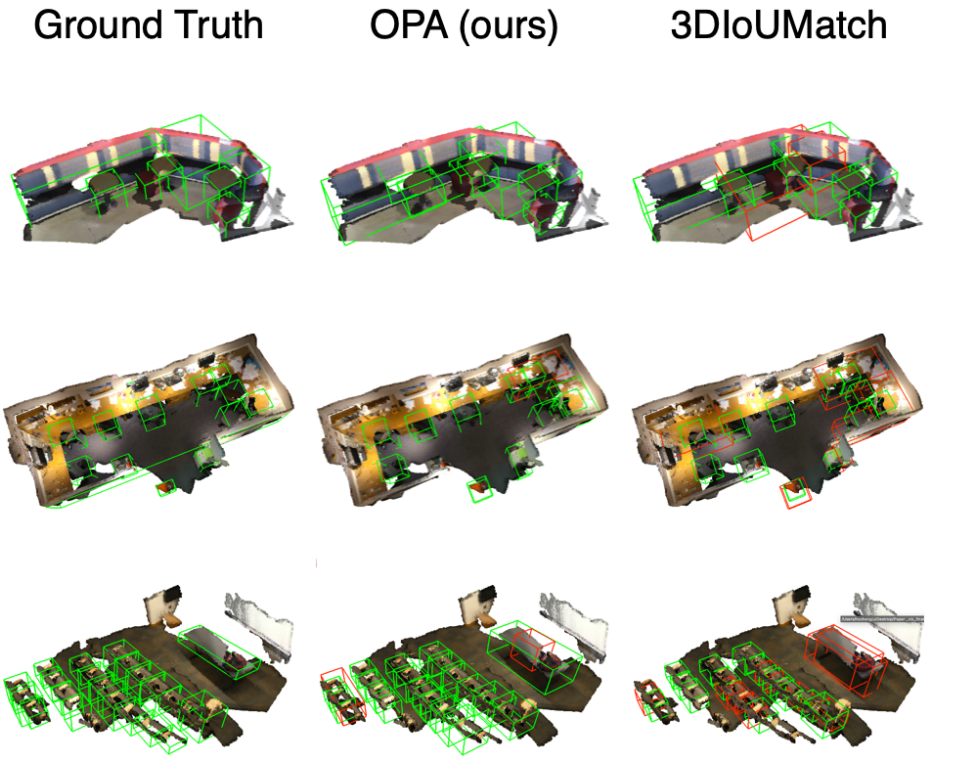

Figure 2: Qualitative results on the ScanNet val set, training with 10% labeled data. The green bounding boxes denote the IoU score of proposals greater than 0.25, while the red bounding boxes indicate the IoU score of the proposal less than 0.25.

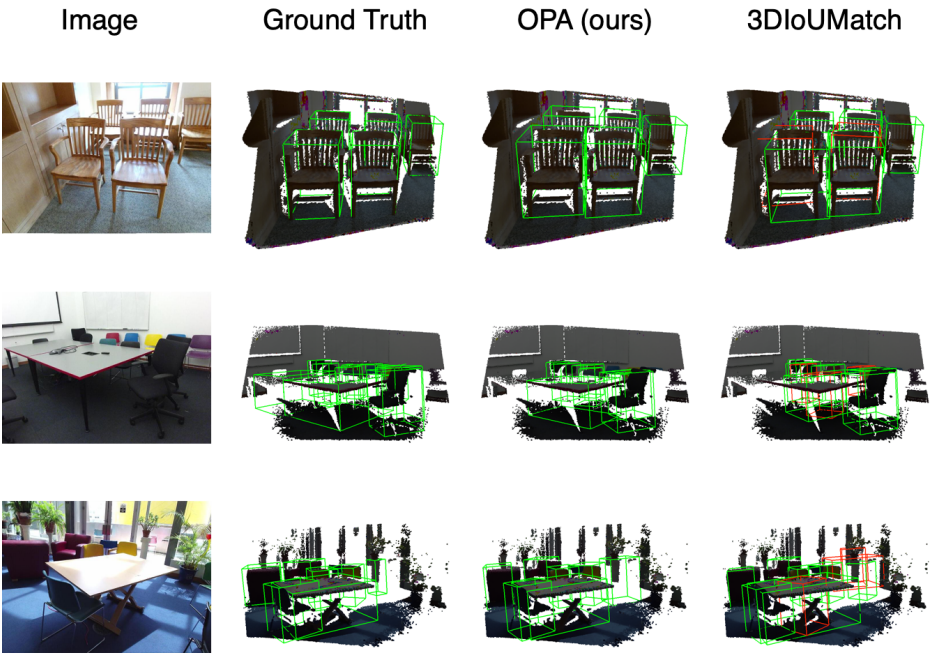

Figure 3: Qualitative results on the SUN RGB-D val set, training with 5% labeled data. The green bounding boxes denote the IoU score of proposals greater than 0.25, while the red bounding boxes indicate the IoU score of the proposal less than 0.25.

References

[1] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Niessner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[2] Charles R Qi, Or Litany, Kaiming He, and Leonidas J Guibas. Deep hough voting for 3d object detection in point clouds. In *proceedings of the IEEE International Conference on Computer Vision*, pages 9277–9286, 2019.

[3] Shuran Song, Samuel P. Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[4] He Wang, Yezhen Cong, Or Litany, Yue Gao, and Leonidas J Guibas. 3dioumatch: Leveraging iou prediction for semi-supervised 3d object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 14615–14624, 2021.

[5] Na Zhao, Tat-Seng Chua, and Gim Hee Lee. Sess: Self-ensembling semi-supervised 3d object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11079–11087, 2020.