Supplementary Material for "Exploring Geometric Consistency for Monocular 3D Object Detection"

Qing Lian [1], Botao Ye[2,3], Ruijia Xu[1], Weilong Yao[3], Tong Zhang[1]
[1]The Hong Kong University of Science and Technology,
[2]Institute of Computing Technology, Chinese Academy of Sciences, China [3] Autowise.AI
qlianab@connect.ust.hk, botao.ye@vipl.ict.ac.cn, rxuaq@connect.ust.hk,
yaoweilong@autowise.ai, tongzhang@ust.hk

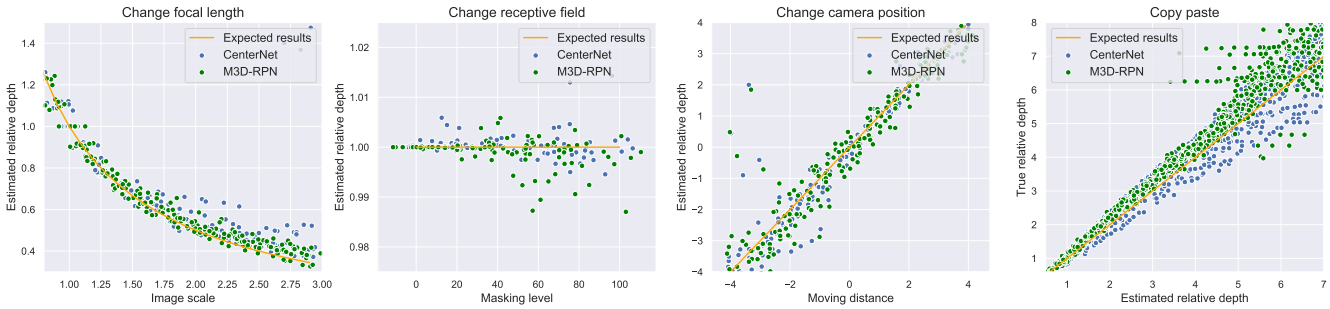

Figure 7. Empirical analysis of augmentation enhanced detectors under geometric manipulations.

The content of supplementary material is organized as follows:

- Section 1 conducts more evaluations of the geometry-aware strategy.
- Section 3 and 4 introduce the implementation details of the data augmentation methods.
- Section 5 presents the details of semi-supervised training settings.

1. More experimental results

We display the experimental results of our geometry-aware augmentation in Table 6. As illustrated, our augmentation methods effectively enhance the model robustness under different kinds of perturbation. Compared to the vanilla version in the main paper, the performance of augmentation enhanced detectors is much better in the perturbation settings.

1.1. Stability of augmentation enhanced detectors

In Figure 7, we also display the empirical analysis we conducted in Section [4] to evaluate whether our proposed data augmentation methods can enhance the stability. Compared with the baseline results in Figure[4], the results from the augmentation enhanced detectors are more fixed with the expected results and have less deviation.

Table 6. Experimental results of Anchor-based (M3D-RPN) and Anchor-free (CenterNet) detectors under different manipulation techniques. Except the baseline setting, we replace the ground-truth with estimated results. For example, "Depth*" denotes replacing the ground truth depth with the estimation and setting all other components with ground truth. (Results of $AP|_{40}$ with IoU$\geq$0.5 on car (easy) are reported.)

Network	Method	Base	Depth*	Dim*	Pos*
M3D-RPN	Origin	65.9	70.2	99.2	99.0
	Random scale	60.1	68.1	98.5	98.6
	Random crop	59.2	62.3	96.6	96.7
	Moving cam	52.8	62.8	93.9	92.1
	Copy-paste	53.2	58.3	89.4	98.2
CenterNet	Origin	60.3	65.3	99.1	99.0
	Random scale	55.3	62.3	98.9	98.8
	Random crop	58.8	64.2	97.3	98.2
	Moving cam	50.3	59.8	91.7	88.6
	Copy-paste	49.2	52.1	90.0	98.8

2. nuScenes datasets

We first introduce the detailed experimental setting on the nuScenes dataset and provide additional results of Cen-

Table 7. Experimental results of the anchor-free detector on the nuScenes validation set.

Pretrained	Setting	mAP↑	mATE↓	mASE↓	NDS↑
ImageNet	Vanilla aug	33.2	0.69	0.28	38.4
	Geo aug	34.5	0.68	0.27	39.4
ImageNet+ DDAD	Vanilla aug	34.6	0.67	0.27	39.4
	Geo aug	35.6	0.66	0.26	40.6

terNet [5] with different augmentation strategies. In the nuScenes dataset, we utilize the AdamW optimizer to train the models with 48 epochs. The initial learning rate is 4e-2 and downscaled with 0.1 in the 32^{th} and 44^{th} epoch. To save the memory occupation, we rescale the input resolution from 1600×900 to 1200×675 in both training and inference, where the batch size is set as 80 during training.

In Table 7, we provide the experimental results of CenterNet with different pre-trained weights. DDAD [3] denotes the private datasets reported in DD3D [3]. We utilize the provide pre-trained models to initialize the modified DLA-34 backbone in the detection model. Experimental results illustrate the effectiveness of our geometry-aware strategy in a stronger baseline setting.

3. Details about geometry-aware data Hyper-parameters in data augmentation

The hyper-parameters for the data augmentation are represented as follows: 1). Random Crop: we randomly crop the image with size of 960×320. 2). Random Scale: we randomly resize the image with a range from 0.8 to 1.2, with fixing the size ratio. 3). Camera position: To alleviate generate artifact, we control the change distance of camera position from -5 to 5 meters. 4). Copy-paste: We first utilizes an instance segmentation method [4] to crop the foreground objects with around 12,581 instances. After that, we randomly select two cropped instances and insert them into every training image with sampling new depth from 0 - 70.

4. Details of Copy-paste augmentation method

Generating bounding boxes For the step 7 in the Algorithm 1, we utilize the acquired object dimension, location, orientation to get the final bounding boxes 2D coordinates. The procedure is similar in [2]. We first calculate the rotation matrix R with using the egocentric orientation angle:

$$R = \begin{pmatrix} \cos\theta & 0 & \sin\theta \\ 0 & 1 & 0 \\ -\sin\theta & 0 & \cos\theta \end{pmatrix}. \tag{1}$$

The 8 corner points in the object coordinate is:

$$P^{3d}_{4\times 8} = \begin{pmatrix} \frac{L}{2} & \frac{L}{2} & -\frac{L}{2} & -\frac{L}{2} & \frac{L}{2} & \frac{L}{2} & -\frac{L}{2} & -\frac{L}{2} \\ 0 & 0 & 0 & 0 & -H & -H & -H & -H \\ \frac{W}{2} & -\frac{W}{2} & -\frac{W}{2} & \frac{W}{2} & \frac{W}{2} & -\frac{W}{2} & -\frac{W}{2} & \frac{W}{2} \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix}.$$

For the coordinate of point i, it is calculated as follows:

$$P^{2d}_{3\times 8} = K_{3\times 4} \begin{pmatrix} R & T \\ 0^T & 1 \end{pmatrix} P^{3d}_{4\times 8}, \tag{2}$$

where T is the 3D location matrix with $[X, Y, Z]$, and $P^{2d}_{3\times 8}$ is the coordinates in the images.

5. Implementation details in the semi-supervised setting

In this paper, we adopt the mean-teacher framework [1] to regularize the output consistency of monocular detectors. Following existing work [1], we first select the candidate bounding boxes based on the pooling module (in CenterNet) or nms module (in M3D-RPN). Then we select the candidates with confidence score larger than 0.7 for regularization. The teacher network is the momentum version of the student network with factor of 0.9. We fed the teacher network with origin image and the student network with augmented images. The weight of the regularization loss is set as 1.

References

[1] Jisoo Jeong, Seungeui Lee, Jeesoo Kim, and Nojun Kwak. Consistency-based semi-supervised learning for object detection. In *NeurIPS*, 2019. 2

[2] Lijie Liu, Jiwen Lu, Chunjing Xu, Qi Tian, and Jie Zhou. Deep fitting degree scoring network for monocular 3d object detection. In *CVPR*, 2019. 2

[3] Dennis Park, Rares Ambrus, Vitor Guizilini, Jie Li, and Adrien Gaidon. Is pseudo-lidar needed for monocular 3d object detection? In *ICCV*, 2021. 2

[4] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019. 2

[5] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *arXiv preprint arXiv:1904.07850*, 2019. 2