Supplementary Material: Lidar Point Cloud Guided Monocular 3D Object Detection

Liang Peng[1,2], Fei Liu[3], Zhengxu Yu[1], Senbo Yan[1,2], Dan Deng[2], Zheng Yang[2], Haifeng Liu[1], and Deng Cai[1,2] ✉

[1] State Key Lab of CAD&CG, Zhejiang University, China
{pengliang, senboyan, haifengliu}@zju.edu.cn yuzxfred@gmail.com
dengcai@cad.zju.edu.cn
[2] Fabu Inc., Hangzhou, China
{dengdan, yangzheng}@fabu.ai
[3] State Key Lab of Industrial Control and Technology, Zhejiang University, China
liufei21@zju.edu.cn

Due to the space limitation, some details and experimental results are included in this supplementary material as below:

- Section A : More details and ablations.
  - Section A.1 : Details of disturbed labels.
  - Section A.2 : Ablation on different LiDAR detectors.
  - Section A.3 : Ablation on different number of pseudo labels.
- Section B : Qualitative results.
  - Section B.1 : Qualitative results of monocular model detections.
  - Section B.2 : Qualitative results of pseudo labels.

A More Details and Ablations

A.1 Details of Disturbed Labels

The labels in Figure 1 in the main text are perturbed by randomly shifting the original value within the percentage range. For example, concerning object's 3D location loc, $loc' = loc(1 + uniform(-\frac{p}{2}, \frac{p}{2}))$, where $uniform(-\frac{p}{2}, \frac{p}{2})$ refers to randomly select a sample from the uniform distribution $\mathcal{U}\left[-\frac{p}{2}, \frac{p}{2}\right]$ (p is 5%, 10%, 20%, 40% in Figure 1) and loc' is the perturbed value.

A.2 Ablation on Different LiDAR Detectors

This section investigates the effect of different LiDAR detectors in LPCG in the high accuracy mode. We employ different LiDAR 3D detectors [3,2,5,4], which are trained by the labeled data. They then generate pseudo labels on unlabeled data. As shown in Table 1, we can see that all LiDAR-based methods bring significant improvements for the monocular detector, and the resulting accuracy is close. It indicates that the monocular method is not sensitive to specific LiDAR 3D detectors.

Monocular Learning Paradigm in LPCG	AP_{BEV} /AP_{3D} (IoU=0.7)$\|_{R_{40}}$ Easy	Moderate	Hard
M3D-RPN [1]	20.85/14.53	15.62/11.07	11.88/8.65
PointPillars [2] + M3D-RPN	34.46/26.85	**26.74/20.12**	**23.75**/17.46
Second [5] + M3D-RPN	**34.71/28.04**	26.07/20.01	22.85/17.80
Part-A^2 [4] + M3D-RPN	33.28/26.20	25.51/20.09	22.49/**17.81**
PV-RCNN [3] + M3D-RPN	33.94/26.17	25.20/19.61	22.06/16.80

Table 1. Influences of different LiDAR detectors. We use different LiDAR 3D detectors to generate pseudo labels. The results are close, meaning that LPCG is not sensitive to specific LiDAR detectors.

A.3 Ablation on The Number of Pseudo Labels

In Table 2, we train the monocular 3D detector with different numbers of pseudo labels. We can observe that the performance increases dramatically with more pseudo labels, which means that collecting more unlabeled LiDAR point clouds can further push the performance of monocular 3D detection. Actually, this offline collecting process can be easily achieved in a real-world self-driving system.

Apporoaches	Samples	$AP_{BEV}/AP_{3D}\|_{R_{40}}$ Easy	Moderate	Hard
Under IoU criterion 0.5				
M3D-RPN + LPCG	100	11.18/7.32	7.70/4.73	6.37/4.06
	200	21.71/17.14	16.13/12.32	14.26/10.69
	500	31.82/26.82	23.31/19.82	20.99/17.08
	1000	35.09/29.41	28.33/23.72	25.23/20.93
	3000	54.86/49.75	40.11/36.58	35.45/32.29
	10000	64.57/59.01	47.48/43.97	43.56/39.04
	26057	**67.20/62.92**	**50.52/47.14**	**46.31/42.03**
Under IoU criterion 0.7				
M3D-RPN + LPCG	100	1.29/0.35	0.74/0.21	0.62/0.19
	200	3.83/1.84	3.01/1.20	2.28/1.14
	500	9.14/4.78	6.40/3.35	5.52/2.96
	1000	11.71/7.48	9.23/5.78	8.07/4.89
	3000	20.96/14.83	15.76/10.80	13.78/9.76
	10000	31.51/25.05	22.76/17.87	19.82/15.30
	26057	**33.94/26.17**	**25.20/19.61**	**22.06/16.80**

Table 2. Influences of the number of pseudo labels. "Samples" in the table denotes the number of training samples, which are generated by pseudo labels. All the methods are evaluated with metric $AP|_{R_{40}}$.

B Qualitative Results

B.1 Qualitative Results of Monocular Model Detections

We provide qualitative results in Figure 1. We compare the predictions from the original model [1] with the ones from the model employing our framework (LPCG). It can be easily seen that our predictions are much more accurate, especially for 3D locations. We also show the failure cases, which are usually heavily occluded or faraway objects. These objects are hard to be precisely recovered due to the ill-posed nature of monocular imagery.

B.2 Qualitative Results of Pseudo Labels

To intuitively understand the gap between pseudo labels and manually annotated labels, we illustrate some examples in Figure 2. We can observe that both types of pseudo labels (produced by high accuracy and low cost mode) are close to the manual annotations. This accuracy can be attributed to the highly precise 3D measurements of LiDAR point clouds, which give explicit information for obtaining objects' 3D locations. However, regarding pseudo labels in the low cost mode, they are produced by the geometry-based method, *i.e.*, finding the minimum bounding box of RoI LiDAR points and then filtering invalid boxes (via the dimension prior). When the LiDAR point clouds cannot describe the 3D outline of the object, this geometry-based method will fail since the minimum bounding box is filtered by the dimension constraint. Therefore, many real objects are missed. These objects usually have few LiDAR points or only have one surface that is captured by the LiDAR device. We show these failure cases in Figure 3.

References

1. Brazil, G., Liu, X.: M3d-rpn: Monocular 3d region proposal network for object detection. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 9287–9296 (2019)
2. Lang, A.H., Vora, S., Caesar, H., Zhou, L., Yang, J., Beijbom, O.: Pointpillars: Fast encoders for object detection from point clouds. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 12697–12705 (2019)
3. Shi, S., Guo, C., Jiang, L., Wang, Z., Shi, J., Wang, X., Li, H.: Pv-rcnn: Point-voxel feature set abstraction for 3d object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10529–10538 (2020)
4. Shi, S., Wang, Z., Shi, J., Wang, X., Li, H.: From points to parts: 3d object detection from point cloud with part-aware and part-aggregation network. arXiv preprint arXiv:1907.03670 (2019)
5. Yan, Y., Mao, Y., Li, B.: Second: Sparsely embedded convolutional detection. Sensors **18**(10), 3337 (2018)

Fig. 1. Qualitative results of M3D-RPN [1] trained by our framework (LPCG). **Green**: ground-truths. **Red**: our predictions. **White**: original predictions from M3D-RPN. The bolded side of the box in the bird's-eye-view map refers to the orientation. We can see that our predictions are much more accurate. Note that some predictions are overlapped by ground-truths. We also show failure cases, which are contained in the gray dotted circle. Best viewed in color with zoom in.

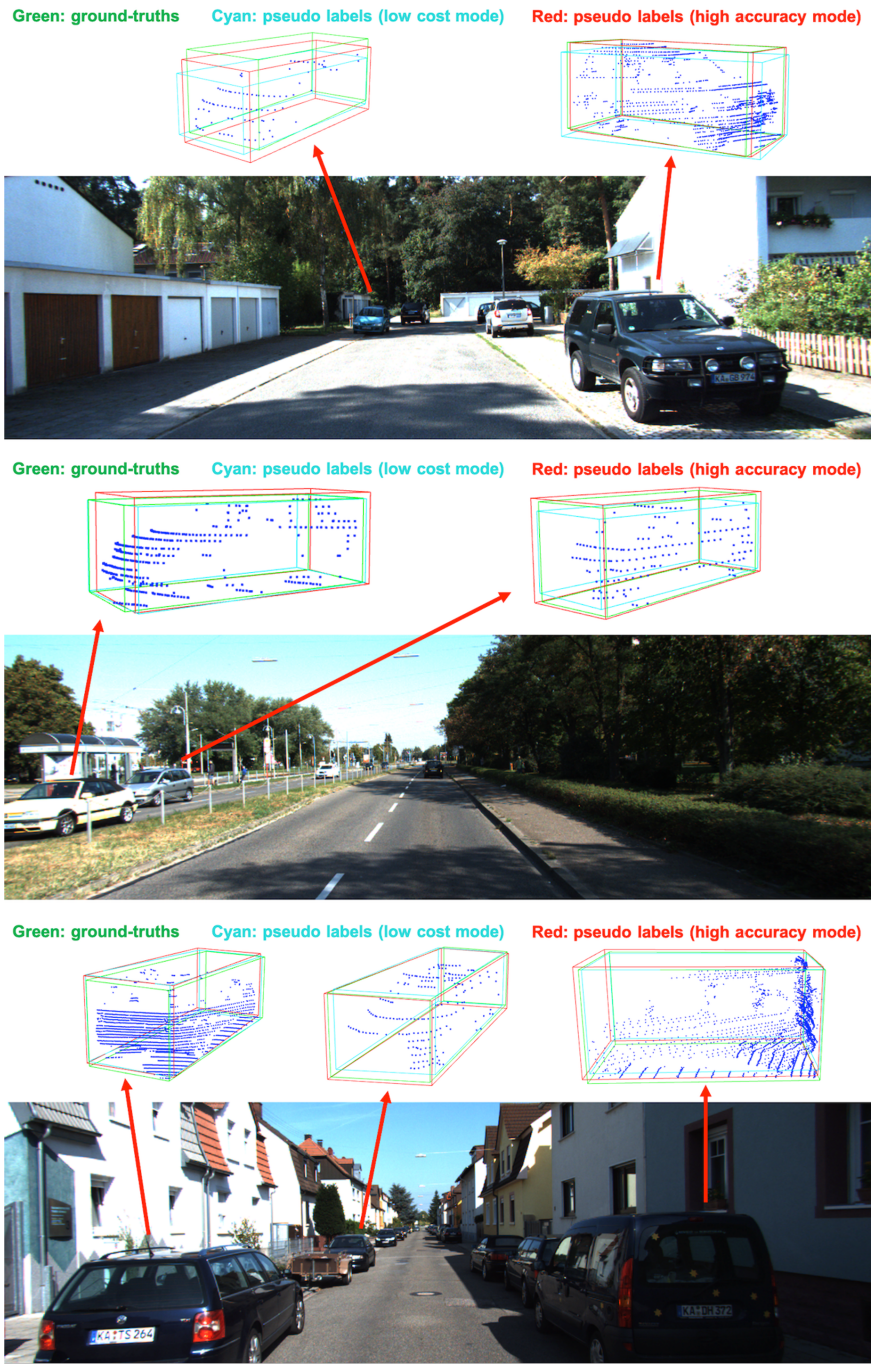

Fig. 2. Qualitative comparisons on different labels. Thanks to the accurate LiDAR 3D measurements, all types of labels are close, especially for 3D locations.

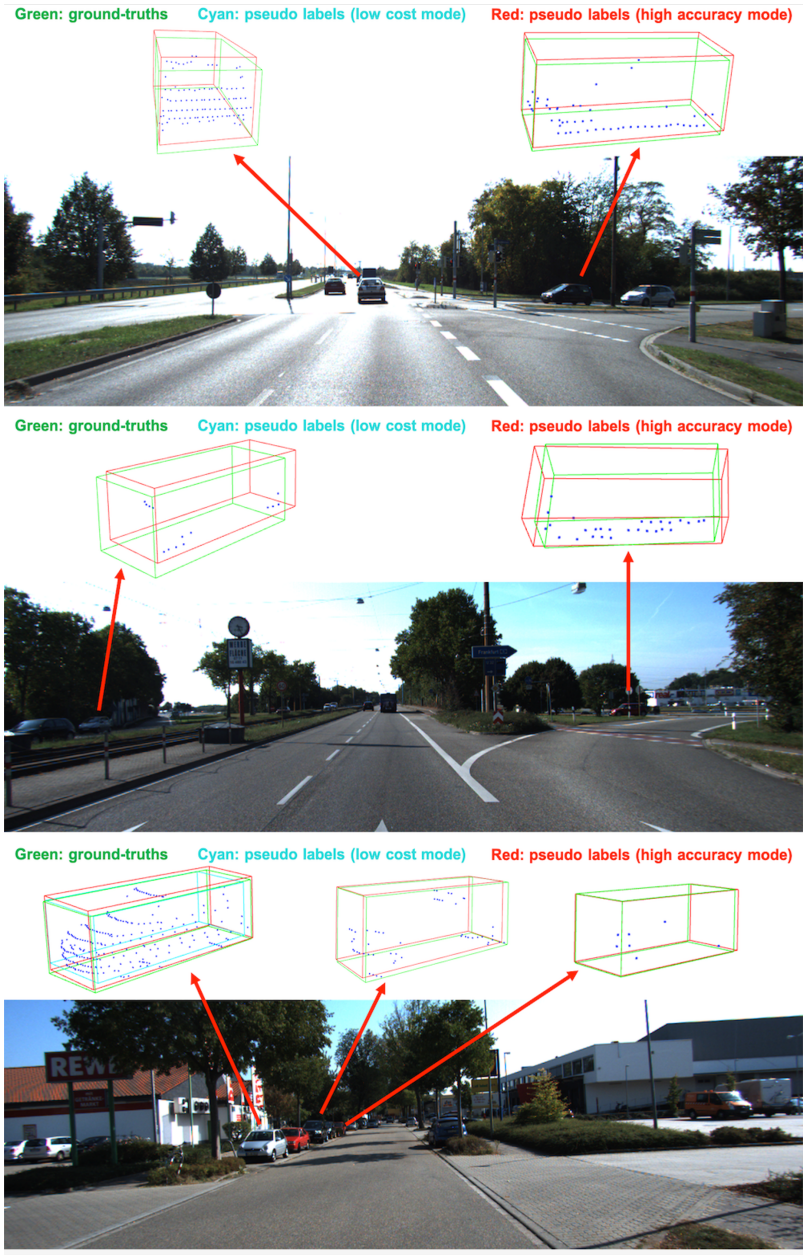

Fig. 3. Failure cases of pseudo labels on the low cost mode. When the RoI LiDAR point clouds cannot fully describe the object's 3D outline, the geometry-based method cannot recover good 3D box pseudo labels, which are filtered by the dimension constraints. Thus many real objects are missed, especially for occluded and faraway objects. By contrast, pseudo labels from high accuracy mode still achieve good results because the LiDAR-based network is well-trained, which has studied the latent object pattern.