Supplementary Material: The Devil is in the Task: Exploiting Reciprocal Appearance-Localization Features for Monocular 3D Object Detection

1. More quantitative experiments

Impact of different depth estimators For generalization ability, we compare DFR-Net (depth-assisted version) with the baseline (D^4LCN [2]) based on different depth estimation methods. We choose the monocular depth estimator DORN [3] (which has been already reported in our paper) as well as the more accurate stereo matching method PSMNet [1] to obtain depth maps for comparison. Note that stereo matching approaches attain higher accuracy in depth estimation than monocular methods. As shown in Table 1, the performance gain with respect to baseline are getting larger with the increasing accuracy of estimated depth.

Depth	Method	AP_{3D}		
		Mod.	Easy	Hard
DORN [3]	D^4LCN [2]	22.32	16.20	12.30
	Ours	**24.81**	**17.78**	**14.41**
	Improvement	+2.49	+1.58	+2.11
PSMNet [1]	D^4LCN [2]	25.24	19.80	16.45
	Ours	**28.97**	**21.12**	**17.19**
	Improvement	+3.73	+1.32	+0.74

Table 1. Comparison of different depth estimators on the KITTI "val1" split set at IoU = 0.7 (R40). The first and second rows show the "Car" results of the baseline (D^4LCN [2]) and our method.

Plug-and-play on the anchor-free monocular 3D detection framework We apply an anchor-free monocular 3D object detection method SMOKE [4] as our encoding backbone to validate the expansion capability. Compared to the anchor-based methods [2] used in our paper, SMOKE reformulates the 3D detection as the coarse keypoints detection task instead of relying on the predefined anchors of 3D bounding boxes. This strategy significantly improves both training convergence and inference time. As shown in Table 2, our method boosts SMOKE [4] by a large margin for all three entries. The inference speed and model size are comparable with SMOKE [4] (33.0 vs. 33.3 FPS; 225.5 vs. 223.0 Mb).

Method	$AP\|_{R_{11}}$			$AP\|_{R_{40}}$		
	Mod.	Easy	Hard	Mod.	Easy	Hard
SMOKE [4]	12.85	14.76	11.50	5.05	7.50	4.49
Ours	**14.90**	**17.59**	**14.20**	**8.67**	**11.49**	**7.27**
Improvement	+2.05	+2.83	+2.70	+3.62	+3.99	+2.78

Table 2. Comparison of our DFR-Net (anchor-free version) and SMOKE [4] on the KITTI "val1" split set at IoU = 0.7 (AP_{3D}).

Effect of the residual mechanism To avoid the negative impact of noisy attention at the initial stage of the network, we design an adaptive residual connection. As shown in Table 3, the "Mod." performance of the proposed model improves from 13.92 to 14.72 as the residual mechanism is adopted. These experimental results demonstrate the effectiveness of the adaptive residual connection.

Method	$AP\|_{R_{40}}$		
	Mod.	Easy	Hard
w/o residual mechanism	13.92	18.21	10.32
w/ residual mechanism	**14.79**	**19.95**	**11.04**
Improvement	+0.87	+1.74	+0.72

Table 3. Effect of the residual mechansim on the KITTI "val1" split set at IoU = 0.7 (AP_{3D}).

2. More qualitative results

We visualize more results of prediction and ground-truth 3D boxes in Figure 1 and Figure 2. It is obvious that the proposed DFR-Net can produce accurate bounding box predictions, especially for distant and occluded objects.

References

[1] Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. In *CVPR*, 2018. 1

[2] Mingyu Ding, Yuqi Huo, Hongwei Yi, Zhe Wang, Jianping Shi, Zhiwu Lu, and Ping Luo. Learning depth-guided convolutions for monocular 3d object detection. In *CVPR*, 2020. 1

Figure 1. More qualitative results on the KITTI dataset. The 3D ground-truth boxes and our DFR-Net predictions are drawn in green and red, respectively.

[3] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. In *CVPR*, 2018. 1

[4] Zechen Liu, Zizhang Wu, and Roland Toth. Smoke: Single-stage monocular 3d object detection via keypoint estimation. In *CVPR workshops*, 2020. 1

Figure 2. More qualitative results on the KITTI dataset. The 3D ground-truth boxes and our DFR-Net predictions are drawn in green and red, respectively.