Learning Depth-Guided Convolutions for Monocular 3D Object Detection – Supplemental Material –

Conv module	AP$\|_{R_{11}}$ Easy	**Moderate**	Hard	AP$\|_{R_{40}}$ Easy	**Moderate**	Hard
Dynamic [7]	23.01	17.67	15.85	17.47	12.18	09.53
Dynamic Local [7]	25.15	18.42	16.27	21.09	13.93	11.31
Deformable [5]	23.98	18.24	16.11	19.05	13.42	10.07
D^{4}LCN (ours)	**26.97**	**21.71**	**18.22**	**22.32**	**16.20**	**12.30**

Table 1. Comparisons of different convolutional modules for *car* 3D detection on the KITTI split1.

A. Evaluation of Convolutional Appoaches

To show the effectiveness of our guided filtering module for 3D object detection, we compare it with several alternatives: Dynamic Convolution [7], Dynamic Local Filtering [7], and Deformable Convolution [5]. Our method belongs to dynamic networks but yields less computation cost and stronger representation. For the first two methods, we conduct experiments using the same depth map as ours. For the third method, we apply deformable convolution on both RGB and depth branches and merge them by element-wise product. From Table 1, we can observe that our method performs the best. This indicates that our method can better capture 3D information from RGB images due to the special design of our D^{4}LCN.

B. Definition of 3D Corners

We define the eight corners of each ground truth box as follows:

$$C^{(m)} = \begin{bmatrix} x^{(m)} \\ y^{(m)} \\ 1 \end{bmatrix}_P \cdot z_{3D}^{(m)} = \left(r_y \cdot \begin{bmatrix} \pm w/2 \\ \pm h/2 \\ \pm l/2 \\ 0 \end{bmatrix}_{3D} + \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix}_{3D} \right) \quad (1)$$

where $m \in (int)[1,8]$ in a defined order, and r_y is the egocentric rotation matrix. Note that we use allocentric pose for regression.

C. Comparisons between Two Rotation Definitions

As shown in Figure 1, while egocentric poses undergo viewpoint changes towards the camera when translated, allocentric poses always exhibit the same view, independent

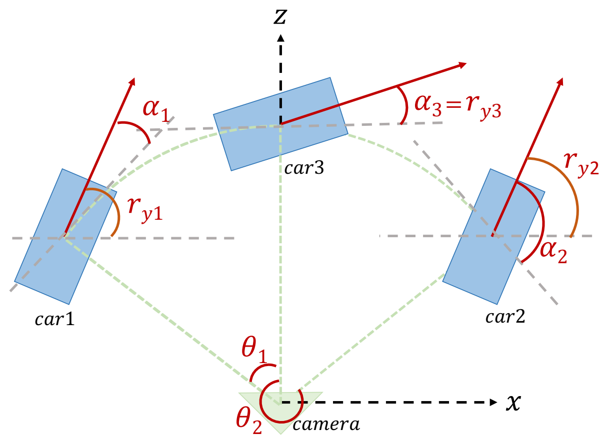

Figure 1. Comparisons between egocentric (r_y) and allocentric (α) poses. The car1 and car2 have the same egocentric pose, but they are observed on different sides (views). We use allocentric pose to keep the same view (car1 and car3).

of the object's location. The allocentric pose α and the egocentric pose r_y can be converted to each other according to the viewing angle θ.

$$\alpha = r_y - \theta \quad (2)$$

D. Ablative Results for Convolutional Methods

The Depth-guided filtering module in our D^{4}LCN model can be decomposed into basic convolutional components:

- Traditional Convolutional Network
- Depth-guided ConvNet (CN)
- Depth-guided Local CN (LCN)
- Depth-guided Depth-wise LCN (DLCN)
- Depth-guided DLCN with Shift-pooling (SP-DLCN)
- D^{4}LCN (Our full model)

Conv Method	Dynamic	Local	Depth-wise	Shift-pooling	Dilated	$AP\|_{R_{11}}$ Easy	$AP\|_{R_{11}}$ **Moderate**	$AP\|_{R_{11}}$ Hard	$AP\|_{R_{40}}$ Easy	$AP\|_{R_{40}}$ **Moderate**	$AP\|_{R_{40}}$ Hard
ConvNet						20.66	15.57	13.41	17.10	12.09	09.47
Depth-guided CN	✓					23.01	17.67	15.85	17.47	12.18	09.53
Depth-guided LCN	✓	✓				25.15	18.42	16.27	21.09	13.93	11.31
Depth-guided DLCN	✓	✓	✓			23.25	17.92	15.58	18.32	13.50	10.61
Depth-guided SP-DLCN	✓	✓	✓	✓		25.30	19.02	17.26	19.69	14.44	11.52
D^{4}LCN	✓	✓	✓	✓	✓	**26.97**	**21.71**	**18.22**	**22.32**	**16.20**	**12.30**

Table 2. Comparisons of different convolutional methods for *car* 3D detection on the KITTI split1.

The ablative results for these convolutional methods are shown in Table 2. We can observe that: (1) Using the depth map to guide the convolution of each pixel brings a considerable improvement. (2) Depth-wise convolution with shift-pooling operator not only has fewer parameters (Section 3.2 of our main paper) but also gets better performance than the standard convolution. (3) The main improvement comes from our adaptive dilated convolution, which allows each channel of the feature map to have different receptive fields.

E. Distributions of Different Dilation

We show the average ratio of different channels with different dilation rates in three blocks of our model over the validation set of split1 (Figure 2). It can be seen that: (1) For the first block with insufficient receptive field, the model tends to increase the receptive field by large dilation rate, and then it uses small receptive field for the second block. (2) In the third block, the model uses three different dilation rates evenly to deal with the object detection of different scales. We also show the active maps corresponding to different filters of the third block of our D^{4}LCN in our main paper (Figure 5).

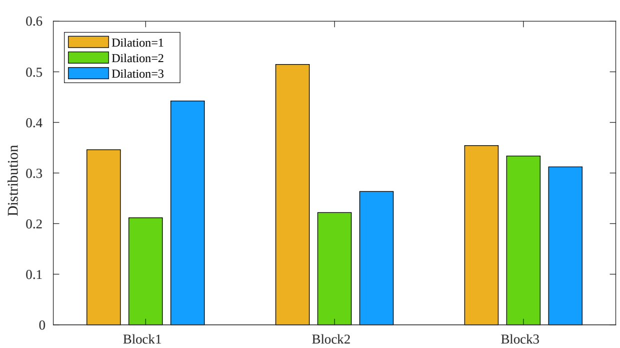

Figure 2. The average ratio of different channels with different dilation rates in three blocks.

References

[1] Garrick Brazil and Xiaoming Liu. M3d-rpn: Monocular 3d region proposal network for object detection. In *ICCV*, pages 9287–9296, 2019.

[2] Florian Chabot, Mohamed Chaouch, Jaonary Rabarisoa, Céline Teulière, and Thierry Chateau. Deep manta: A coarse-to-fine many-task network for joint 2d and 3d vehicle analysis from monocular image. In *CVPR*, pages 2040–2049, 2017.

[3] Xiaozhi Chen, Kaustav Kundu, Ziyu Zhang, Huimin Ma, Sanja Fidler, and Raquel Urtasun. Monocular 3d object detection for autonomous driving. In *CVPR*, pages 2147–2156, 2016.

[4] Xiaozhi Chen, Kaustav Kundu, Yukun Zhu, Huimin Ma, Sanja Fidler, and Raquel Urtasun. 3d object proposals using stereo imagery for accurate object class detection. *TPAMI*, 40(5):1259–1272, 2017.

[5] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *ICCV*, pages 764–773, 2017. 1

[6] Tong He and Stefano Soatto. Mono3d++: Monocular 3d vehicle detection with two-scale 3d hypotheses and task priors. *arXiv preprint arXiv:1901.03446*, 2019.

[7] Xu Jia, Bert De Brabandere, Tinne Tuytelaars, and Luc V Gool. Dynamic filter networks. In *NeurIPS*, pages 667–675, 2016. 1

[8] Eskil Jörgensen, Christopher Zach, and Fredrik Kahl. Monocular 3d object detection and box fitting trained end-to-end using intersection-over-union loss. *arXiv preprint arXiv:1906.08070*, 2019.

[9] Jason Ku, Alex D Pon, and Steven L Waslander. Monocular 3d object detection leveraging accurate proposals and shape reconstruction. In *CVPR*, pages 11867–11876, 2019.

[10] Buyu Li, Wanli Ouyang, Lu Sheng, Xingyu Zeng, and Xiaogang Wang. Gs3d: An efficient 3d object detection framework for autonomous driving. In *CVPR*, pages 1019–1028, 2019.

[11] Lijie Liu, Jiwen Lu, Chunjing Xu, Qi Tian, and Jie Zhou. Deep fitting degree scoring network for monocular 3d object detection. In *CVPR*, pages 1057–1066, 2019.

[12] Xinzhu Ma, Zhihui Wang, Haojie Li, Pengbo Zhang, Wanli Ouyang, and Xin Fan. Accurate monocular 3d object detection via color-embedded 3d reconstruction for autonomous driving. In *ICCV*, pages 6851–6860, 2019.

[13] Fabian Manhardt, Wadim Kehl, and Adrien Gaidon. Roi-10d: Monocular lifting of 2d detection to 6d pose and metric shape. In *CVPR*, pages 2069–2078, 2019.

[14] Arsalan Mousavian, Dragomir Anguelov, John Flynn, and Jana Kosecka. 3d bounding box estimation using deep learning and geometry. In *CVPR*, pages 7074–7082, 2017.

[15] Andretti Naiden, Vlad Paunescu, Gyeongmo Kim, Byeong-Moon Jeon, and Marius Leordeanu. Shift r-cnn: Deep monocular 3d object detection with closed-form geometric constraints. *arXiv preprint arXiv:1905.09970*, 2019.

[16] Zengyi Qin, Jinglu Wang, and Yan Lu. Monogrnet: A geometric reasoning network for monocular 3d object localization. In *AAAI*, volume 33, pages 8851–8858, 2019.

[17] Thomas Roddick, Alex Kendall, and Roberto Cipolla. Orthographic feature transform for monocular 3d object detection. *arXiv preprint arXiv:1811.08188*, 2018.

[18] Andrea Simonelli, Samuel Rota Rota Bulò, Lorenzo Porzi, Manuel López-Antequera, and Peter Kontschieder. Disentangling monocular 3d object detection. *arXiv preprint arXiv:1905.12365*, 2019.

[19] Yan Wang, Wei-Lun Chao, Divyansh Garg, Bharath Hariharan, Mark Campbell, and Kilian Q Weinberger. Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving. In *CVPR*, pages 8445–8453, 2019.

[20] Xinshuo Weng and Kris Kitani. Monocular 3d object detection with pseudo-lidar point cloud. *arXiv preprint arXiv:1903.09847*, 2019.

[21] Yu Xiang, Wongun Choi, Yuanqing Lin, and Silvio Savarese. Data-driven 3d voxel patterns for object category recognition. In *CVPR*, pages 1903–1911, 2015.

[22] Bin Xu and Zhenzhong Chen. Multi-level fusion based 3d object detection from monocular images. In *CVPR*, pages 2345–2353, 2018.