# Densely Constrained Depth Estimator for Monocular 3D Object Detection
Supplementary Material

Yingyan Li[1,2,4,5], Yuntao Chen, Jiawei He[1,2,4], and Zhaoxiang Zhang[1,2,3,4,5]

[1] Institute of Automation, Chinese Academy of Sciences (CASIA)
[2] University of Chinese Academy of Sciences (UCAS)
[3] Centre for Artificial Intelligence and Robotics, HKISI_CAS
[4] National Laboratory of Pattern Recognition (NLPR)
[5] School of Future Technology, UCAS
{liyingyan2021,hejiawei2019,zhaoxiang.zhang}@ia.ac.cn,
chenyuntao08@gmail.com

## 1 WOD detailed results

We also provide the detailed results on WOD [4] as Table 1 and Table 2 show.

| Difficulty | Method | 3D mAP | | | | 3D mAPH | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Overall | 0-30m | 30-50m | 50-∞ | Overall | 0-30m | 30-50m | 50m-∞ |
| LEVEL_1 @IoU 0.7 | M3D-RPN‡ [1] | 0.35 | 1.12 | 0.18 | 0.02 | 0.34 | 1.10 | 0.18 | 0.02 |
| | PatchNet†[2] | 0.39 | 1.67 | 0.13 | 0.03 | 0.37 | 1.63 | 0.12 | 0.03 |
| | PCT [5] | 0.89 | 3.18 | 0.27 | 0.07 | 0.88 | 3.15 | 0.27 | 0.07 |
| | CaDDN [3] | 5.03 | 14.54 | 1.47 | 0.10 | 4.99 | 14.43 | 1.45 | 0.10 |
| | *MonoFlex** [6] | 11.70 | 30.64 | 5.29 | 1.05 | 11.64 | 30.48 | 5.27 | 1.04 |
| | *DCD(Ours)* | **12.57** | **32.47** | **5.94** | **1.24** | **12.50** | **32.30** | **5.91** | **1.23** |
| LEVEL_2 @IoU 0.7 | M3D-RPN‡ [1] | 0.33 | 1.12 | 0.18 | 0.02 | 0.33 | 1.10 | 0.17 | 0.02 |
| | PatchNet†[2] | 0.38 | 1.67 | 0.13 | 0.03 | 0.36 | 1.63 | 0.11 | 0.03 |
| | PCT [5] | 0.66 | 3.18 | 0.27 | 0.07 | 0.66 | 3.15 | 0.26 | 0.07 |
| | CaDDN [3] | 4.49 | 14.50 | 1.42 | 0.09 | 4.45 | 14.38 | 1.41 | 0.09 |
| | *MonoFlex** [6] | 10.96 | 30.54 | 5.14 | 0.91 | 10.90 | 30.37 | 5.11 | 0.91 |
| | *DCD(Ours)* | **11.78** | **32.30** | **5.76** | **1.08** | **11.72** | **32.19** | **5.73** | **1.08** |

**Table 1.** The IoU@0.7 result on WOD [4] *val* set. *Italics*: These methods utilize the whole *train* set, while the others uses 1/3 amount of images in *train* set. ‡: M3D-RPN is re-implemented by [3]. †: PatchNet is re-implemented by [5]. *: MonoFlex is our baseline and re-implemented ourselves.

## 2 Keypoints visualization

Fig. 1 visualizes the 63 semantic keypoints.

| Difficulty | Method | 3D mAP | | | | 3D mAPH | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Overall | 0-30m | 30-50m | 50-$\infty$ | Overall | 0-30m | 30-50m | 50m-$\infty$ |
| LEVEL_1 @IoU 0.5 | M3D-RPN‡ [1] | 3.79 | 11.14 | 2.16 | 0.26 | 3.63 | 10.70 | 2.09 | 0.21 |
| | PatchNet†[2] | 2.92 | 10.03 | 1.09 | 0.23 | 2.74 | 9.75 | 0.96 | 0.18 |
| | PCT [5] | 4.20 | 14.70 | 1.78 | 0.39 | 4.15 | 14.54 | 1.75 | 0.39 |
| | CaDDN [3] | 17.54 | 45.00 | 9.24 | 0.64 | 17.31 | 44.46 | 9.11 | 0.62 |
| | *MonoFlex*$^*$ [6] | 32.26 | 61.13 | 25.85 | 9.03 | 32.06 | 60.75 | 25.71 | 8.95 |
| | *DCD(Ours)* | **33.44** | **62.70** | **26.35** | **10.16** | **33.24** | **62.35** | **26.21** | **10.09** |
| LEVEL_2 @IoU 0.5 | M3D-RPN‡ [1] | 3.61 | 11.12 | 2.12 | 0.24 | 3.46 | 10.67 | 2.04 | 0.20 |
| | PatchNet†[2] | 2.42 | 10.01 | 1.07 | 0.22 | 2.28 | 9.73 | 0.94 | 0.16 |
| | PCT [5] | 4.03 | 14.67 | 1.74 | 0.36 | 3.99 | 14.51 | 1.71 | 0.35 |
| | CaDDN [3] | 16.51 | 44.87 | 8.99 | 0.58 | 16.28 | 44.33 | 8.86 | 0.55 |
| | *MonoFlex*$^*$ [6] | 30.31 | 60.91 | 25.11 | 7.92 | 30.12 | 60.54 | 24.97 | 7.85 |
| | *DCD(Ours)* | **31.43** | **62.48** | **25.60** | **8.92** | **31.25** | **62.13** | **25.46** | **8.86** |

**Table 2.** The IoU@0.5 result on WOD[4] *val* set. *Italics*: These methods utilize the whole *train* set, while the others uses 1/3 amount of images in *train* set. ‡: M3D-RPN is re-implemented by [3]. †: PatchNet is re-implemented by [5]. $^*$: MonoFlex is our baseline and re-implemented ourselves.

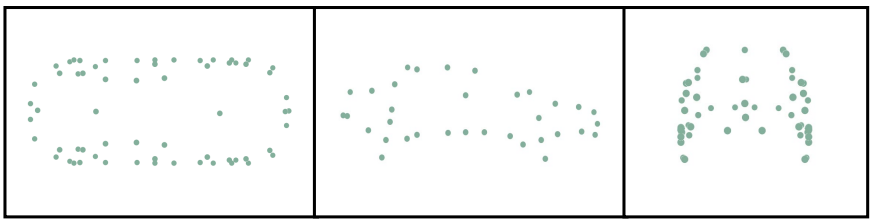

**Fig. 1.** This figure shows the three views of 63 semantic keypoints of the car.

## References


1. Brazil, G., Liu, X.: M3d-rpn: Monocular 3d region proposal network for object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9287–9296 (2019) 1, 2
2. Ma, X., Liu, S., Xia, Z., Zhang, H., Zeng, X., Ouyang, W.: Rethinking pseudo-lidar representation. In: European Conference on Computer Vision. pp. 311–327. Springer (2020) 1, 2
3. Reading, C., Harakeh, A., Chae, J., Waslander, S.L.: Categorical depth distribution network for monocular 3d object detection. arXiv preprint arXiv:2103.01100 (2021) 1, 2
4. Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B., Vasudevan, V., Han, W., Ngiam, J., Zhao, H., Timofeev, A., Ettinger, S., Krivokon, M., Gao, A., Joshi, A., Zhang, Y., Shlens, J., Chen, Z., Anguelov, D.: Scalability in perception for autonomous driving: Waymo open dataset. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020) 1, 2
5. Wang, L., Zhang, L., Zhu, Y., Zhang, Z., He, T., Li, M., Xue, X.: Progressive coordinate transforms for monocular 3d object detection. Advances in Neural Information Processing Systems **34** (2021) 1, 2
6. Zhang, Y., Lu, J., Zhou, J.: Objects are different: Flexible monocular 3d object detection. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021. pp. 3289–3298. Computer Vision Foundation / IEEE (2021) 1, 2