Alleviating Foreground Sparsity for Semi-Supervised Monocular 3D Object Detection - Supplementary Material

Weijia Zhang[1] Dongnan Liu[1] Chao Ma[2] Weidong Cai[1]
[1]University of Sydney [2]Shanghai Jiao Tong University
{wzha0649, dongnan.liu, tom.cai}@sydney.edu.au chaoma@sjtu.edu.cn

In this supplementary material, we provide additional experimental details, as well as quantitative and qualitative results of our proposed method.

1. Additional Experimental Details

Network details. The ODM3D cross-modal distillation framework consists of a SECOND [21] teacher and a CaDDN [19] student. The SECOND network voxelises point clouds $\mathbf{L}$ within the range of $[x_{min} = 2m, y_{min} = -30.08m, z_{min} = -3m, x_{max} = 46.8m, y_{max} = 30.08m, z_{max} = 1m]$ with a resolution of $[0, 04m, 0.04m, 0.1m]$. It produces an intermediate bird's-eye view (BEV) feature map $\mathbf{F}_{\text{BEV}}^{\text{Tea}} \in \mathbb{R}^{140\times188\times128}$. CaDDN takes as input RGB images $\mathbf{I} \in \mathbb{R}^{W_{\text{RGB}}\times H_{\text{RGB}}\times3}$. It employs ResNet-101 [3] and DeepLabV3 [1] for its RGB feature extraction backbone and categorical depth estimation backbone, respectively. CaDDN produces an intermediate BEV feature map $\mathbf{F}_{\text{BEV}}^{\text{Stu}} \in \mathbb{R}^{140\times188\times128}$, whose shape aligns with $\mathbf{F}_{\text{BEV}}^{\text{Tea}}$. Following [4], we stack five calibration blocks from [13] to refine these BEV features. The occupancy mask is a 2D mask that has the same width and height as $\mathbf{F}_{\text{BEV}}^{\text{Tea}}$ and $\mathbf{F}_{\text{BEV}}^{\text{Stu}}$.

Training details. First, to prepare the pre-trained LiDAR-based teacher, we follow the same settings as in [4] and train a SECOND [21] detector for 80 epochs on labelled data using the quality focal loss (QFL) [8]. Next, we train our cross-modal distillation framework using the Adam [6] optimiser with weight decay and a one-cycle learning rate policy. Cross-modal distillation is trained using both labelled and unlabelled data and in two stages. Specifically, we train the framework with only feature distillation in stage 1 for 30 epochs and with both feature and response distillations in stage 2 for 15 epochs. This strategy is intended to ease the multi-task learning of both types of distillations, by training the student to first produce BEV features of adequate quality. CMAug is applied in both stages. A total of 15 "Car", 10 "Pedestrian", and 10 "Cyclist" objects are sampled for each scene and filtered with an IoU-based BEV collision threshold of 0.5, an OAIS-based perspective-view (PV) collision threshold of 0.5, and a PV size threshold of 600, before pasted into the scene. Ground plane data are not utilised when pasting objects. Besides, we also apply random scene-level horizontal flipping to both images and point clouds. During inference, we apply non-maximum suppression (NMS) with an IoU threshold of 0.01, before filtering predicted boxes with a score threshold of 0.2. Test-time augmentation (TTA) is not applied.

Method	Venue	*Val* AP_{BEV}@IoU=0.7		
		Easy	Mod.	Hard
MonoDTR [5]	CVPR'22	33.33	25.35	21.68
DEVIANT [7]	ECCV'22	32.60	23.04	19.99
GUPNet [14]	ICCV'21	31.07	22.94	19.75
MonoDETR [23]	ICCV'23	37.86	26.95	22.80
MonoDistill [2]	ICLR'22	33.09	25.40	22.16
DID-M3D [18]	ECCV'22	31.10	22.76	19.50
MonoDDE [10]	CVPR'22	35.51	26.48	23.07
ADD [20]	AAAI'23	<u>40.38</u>	29.07	25.05
MonoATT [25]	CVPR'23	38.93	<u>29.76</u>	**<u>25.73</u>**
Mix-Teaching* [22]	CSVT'23	37.45	28.99	25.31
ODM3D* (Ours)	-	**41.24**	**30.53**	25.70
Improvements	-	*+0.86*	*+0.77*	*-0.03*

Table 1. $AP_{BEV}|_{R_{40}}$ results of "Car" objects on KITTI *val*. * denotes semi-supervised methods. "*Improvements*" indicates absolute AP improvements compared to the highest results reported by previous methods (underlined). Best results within each subcategory are marked in **bold**.

2. Additional Quantitative Results

KITTI *Val* AP_{BEV} results. We provide AP_{BEV} results on KITTI *val* in Tab. 1. As can be seen, on "Easy" and the most important "Moderate" difficulty levels, our method consistently outperforms CMKD [4] and all previous methods. Our method also achieves competitive results on "Hard" objects, second only to MonoATT [25] with a very small margin.

Method	Venue	Extra Data	Ped. AP_{3D}@IoU=0.5			Cyc. AP_{3D}@IoU=0.5		
			Easy	Mod.	Hard	Easy	Mod.	Hard
MonoFlex [24]	CVPR'21	-	9.43	6.31	5.26	4.17	2.35	2.04
MonoDDE [10]	CVPR'22	-	11.13	7.32	6.67	5.94	3.78	3.33
MonoJSG [12]	CVPR'22	-	11.02	7.49	6.41	5.45	3.21	2.57
MonoDETR [23]	ICCV'23	-	12.54	7.89	6.65	7.33	4.18	2.92
CaDDN [19]	CVPR'21	LiDAR	12.87	8.14	6.76	7.00	3.41	3.30
MonoDistill [2]	ICLR'22	LiDAR	12.79	8.17	7.45	5.53	2.81	2.40
DEVIANT [7]	ECCV'22	-	13.43	8.65	7.69	5.05	3.13	2.59
DD3D [15]	ICCV'21	Depth	13.91	9.30	8.05	2.39	1.52	1.31
GUPNet [14]	ICCV'21	-	14.95	9.76	8.41	5.58	3.21	2.66
MonoDTR [5]	CVPR'22	LiDAR	15.33	10.18	8.61	5.05	3.27	3.19
DD3Dv2 [16]	ICRA'23	LiDAR	**16.25**	**10.82**	**9.24**	8.79	5.68	4.75
LPCG* [17]	ECCV'22	LiDAR	7.21	5.53	4.46	4.83	2.65	2.62
3DSeMo* [11]	arXiv'23	LiDAR	10.78	7.26	6.05	7.04	4.24	3.56
Mix-Teaching* [22]	CSVT'23	LiDAR	11.67	7.47	6.61	8.04	4.91	4.15
CMKD* [4]	ECCV'22	LiDAR	13.94	8.79	7.42	12.52	**6.67**	**6.34**
ODM3D* (Ours)	-	LiDAR	15.28	8.98	7.80	**13.73**	6.54	6.21
Improvements	-	-	*+1.34*	*+0.19*	*+0.42*	*+1.21*	*-0.12*	*-0.13*

Table 2. $AP_{3D}|_{R_{40}}$ results of "Pedestrian" and "Cyclist" objects on KITTI *test*. * denotes semi-supervised methods. "*Improvements*" indicates absolute AP improvements compared to a CMKD baseline. Best results within each sub-category are marked in **bold**.

Method	*Test* AP_{3D}@IoU=0.7		
	Easy	Mod.	Hard
MonoFlex [24]	19.94	13.89	12.07
MonoFlex+3DSeMo [11]	23.55	15.25	13.24
Improvements	*+18.1%*	*+9.8%*	*+9.7%*
MonoFlex [24]	19.94	13.89	12.07
MonoFlex+LPCG [17]	25.56	17.80	15.38
Improvements	*+28.2%*	*+28.1%*	*+27.4%*
MonoFlex [24]	19.94	13.89	12.07
MonoFlex+Mix-Teaching [22]	26.89	18.54	15.79
Improvements	*+34.9%*	*+33.5%*	*+30.8%*
CaDDN [19]	19.17	13.41	11.46
CaDDN+CMKD [4]	28.55	18.69	16.77
Improvements	*+48.9%*	*+39.4%*	*+46.3%*
CaDDN [19]	19.17	13.41	11.46
CaDDN+ODM3D (Ours)	29.75	19.09	16.93
Improvements	***+55.2%***	***+42.4%***	***+47.7%***

Table 3. Relative improvements of semi-supervised methods over base detectors for "Car" objects on KITTI *test*.

KITTI *test* "Pedestrian" and "Cyclist" results. "Pedestrian" and "Cyclist" objects in the KITTI dataset suffer from smaller sizes (in terms of numbers of pixels and LiDAR points), non-rigid appearance, and a significantly limited and unbalanced number of samples (2,207 "Pedestrians" and 734 "Cyclists" compared to 14,357 "Cars" in the KITTI 3D training set), leading to large fluctuations in results. For these reasons, some methods do not report results on these two categories. In Tab. 2, we present results on these two categories of our method along with other methods for which these results are available.

Relative improvements. Tab. 3 computes the relative improvements of semi-supervised M3OD methods over their respective supervised base detectors. As can be seen, our method yields the largest performance improvements across "Car" objects of all difficulties by dint of effective utilisation of unlabelled training samples and point cloud data.

3. Additional Qualitative Results

Detection visualisation. In Fig. 1, we showcase detections by our method compared to a CaDDN [19] base detector, CMKD [4], and the ground-truth annotations. It can be seen that our method more accurately detects objects, especially challenging ones (*e.g.* objects with a smaller apparent size, occluded, or in shadows).

CMAug visualisation. In Fig. 2, we provide more examples of point cloud and image scenes augmented by our CMAug strategy and MixedAug [9]. It is clear that MixedAug produces augmented scenes with objects that are extremely challenging (*i.e.* severely occluded) or even impossible (*i.e.* fully occluded) to learn, while our method successfully mitigates such issues.

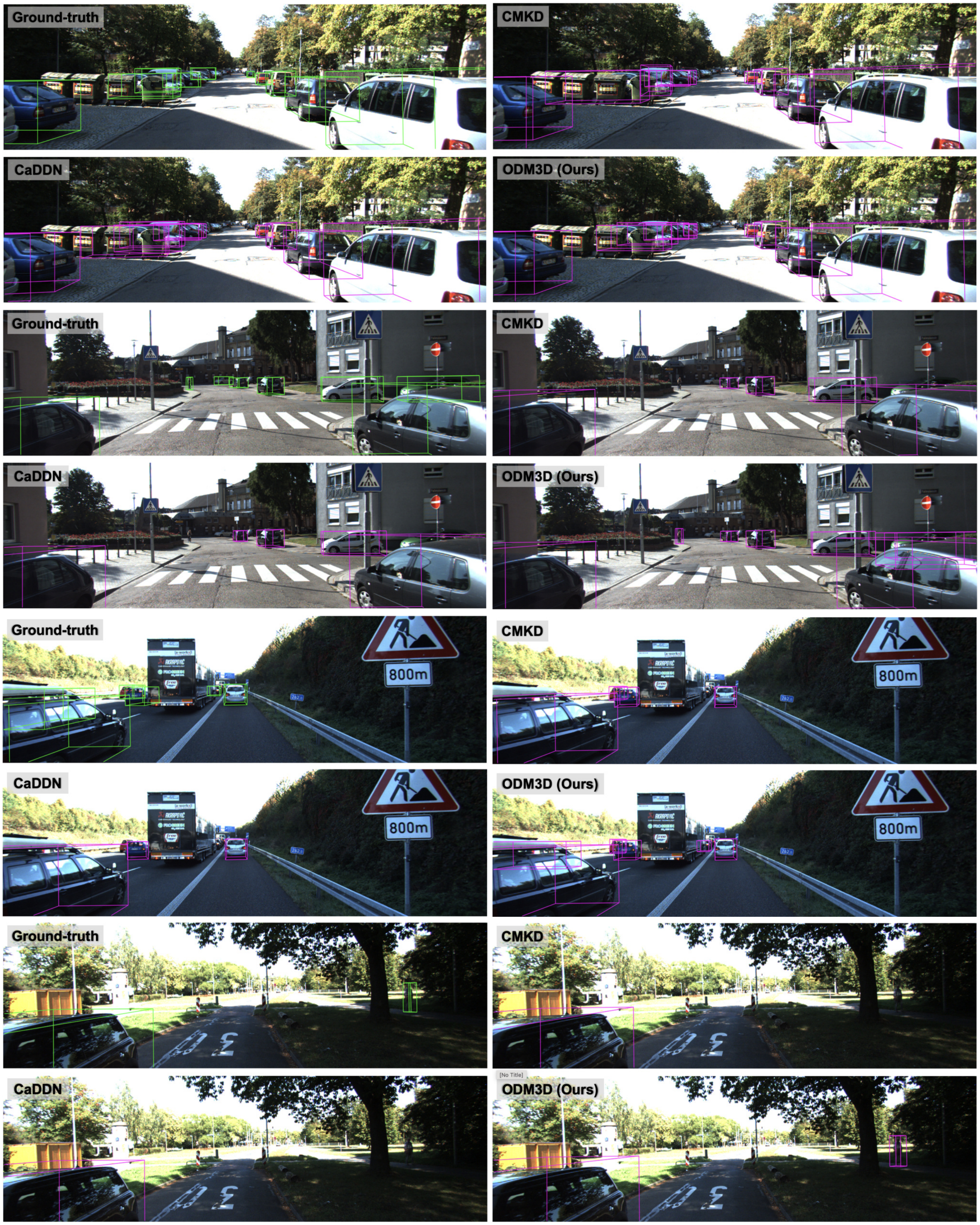

Figure 1. Qualitative comparison of detection results by CaDDN, CMKD, and our method.

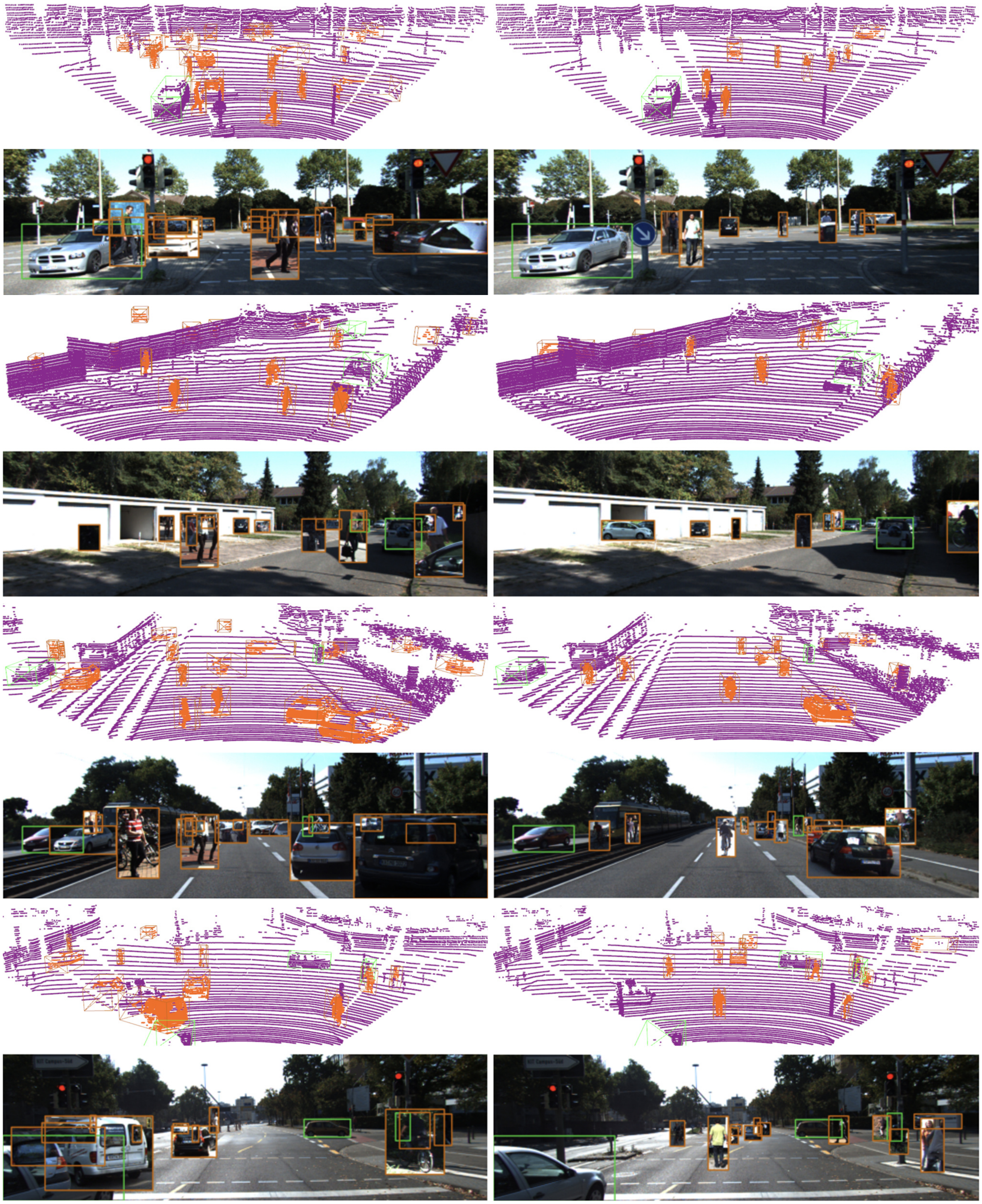

Figure 2. Visualisation of training scenes augmented by MixedAug [9] (left) and our CMAug (right). Pasted objects and original objects are marked with brown and lime green boxes, respectively; pasted points in LiDAR are coloured in orange.

References

[1] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. In *CVPR*, 2017. 1

[2] Zhiyu Chong, Xinzhu Ma, Hong Zhang, Yuxin Yue, Haojie Li, Zhihui Wang, and Wanli Ouyang. Monodistill: Learning spatial features for monocular 3d object detection. In *ICLR*, 2022. 1, 2

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 1

[4] Yu Hong, Hang Dai, and Yong Ding. Cross-modality knowledge distillation network for monocular 3d object detection. In *ECCV*, 2022. 1, 2

[5] Kuan-Chih Huang, Tsung-Han Wu, Hung-Ting Su, and Winston H. Hsu. Monodtr: Monocular 3d object detection with depth-aware transformer. In *CVPR*, 2022. 1, 2

[6] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 1

[7] Abhinav Kumar, Garrick Brazil, Enrique Corona, Armin Parchami, and Xiaoming Liu. Deviant: depth equivariant network for monocular 3d object detection. In *ECCV*, 2022. 1, 2

[8] Xiang Li, Wenhai Wang, Lijun Wu, Shuo Chen, Xiaolin Hu, Jun Li, Jinhui Tang, and Jian Yang. Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection. In *NeurIPS*, 2020. 1

[9] Yanwei Li, Xiaojuan Qi, Yukang Chen, Liwei Wang, Zeming Li, Jian Sun, and Jiaya Jia. Voxel field fusion for 3d object detection. In *CVPR*, 2022. 2, 4

[10] Zhuoling Li, Zhan Qu, Yang Zhou, Jianzhuang Liu, Haoqian Wang, and Lihui Jiang. Diversity matters: Fully exploiting depth clues for reliable monocular 3d object detection. In *CVPR*, 2022. 1, 2

[11] Zhenyu Li, Zhipeng Zhang, Heng Fan, Yuan He, Ke Wang, Xianming Liu, and Junjun Jiang. Augment and criticize: Exploring informative samples for semi-supervised monocular 3d object detection. In *arXiv preprint arXiv:2303.11243*, 2023. 2

[12] Qing Lian, Peiliang Li, and Xiaozhi Chen. Monojsg: Joint semantic and geometric cost volume for monocular 3d object detection. In *CVPR*, 2022. 2

[13] Jiang-Jiang Liu, Qibin Hou, Ming-Ming Cheng, Changhu Wang, and Jiashi Feng. Improving convolutional networks with self-calibrated convolutions. In *CVPR*, 2020. 1

[14] Yan Lu, Xinzhu Ma, Lei Yang, Tianzhu Zhang, Yating Liu, Qi Chu, Junjie Yan, and Wanli Ouyang. Geometry uncertainty projection network for monocular 3d object detection. In *ICCV*, 2021. 1, 2

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

[16] Dennis Park, Jie Li, Dian Chen, Vitor Guizilini, and Adrien Gaidon. Depth is all you need for monocular 3d detection. In *ICRA*, 2023. 2

[17] Liang Peng, Fei Liu, Zhengxu Yu, Senbo Yan, Dan Deng, Zheng Yang, Haifeng Liu, and Deng Cai. Lidar point cloud guided monocular 3d object detection. In *ECCV*, 2022. 2

[18] Liang Peng, Xiaopei Wu, Zheng Yang, Haifeng Liu, and Deng Cai. Did-m3d: Decoupling instance depth for monocular 3d object detection. In *ECCV*, 2022. 1

[19] Cody Reading, Ali Harakeh, Julia Chae, , and Steven L Waslander. Categorical depth distribution network for monocular 3d object detection. In *CVPR*, 2021. 1, 2

[20] Zizhang Wu, Yunzhe Wu, Jian Pu, Xianzhi Li, and Xiaoquan Wang. Attention-based depth distillation with 3d-aware positional encoding for monocular 3d object detection. In *AAAI*, 2023. 1

[21] Yan Yan, Yuxing Mao, and Bo Li. Second: Sparsely embedded convolutional detection. In *Sensors*, 2018. 1

[22] Lei Yang, Xinyu Zhang, Li Wang, Minghan Zhu, Chuan-Fang Zhang, and Jun Li. Mix-teaching: A simple, unified and effective semi-supervised learning framework for monocular 3d object detection. In *IEEE TCSVT*, 2023. 1, 2

[23] Renrui Zhang, Han Qiu, Tai Wang, Ziyu Guo, Ziteng Cui, Yu Qiao, , Hongsheng Li, and Peng Gao. Monodetr: Depth-guided transformer for monocular 3d object detection. In *ICCV*, 2023. 1, 2

[24] Yunpeng Zhang, Jiwen Lu, and Jie Zhou. Objects are different: Flexible monocular 3d object detection. In *CVPR*, 2021. 2

[25] Yunsong Zhou, Hongzi Zhu, Quan Liu, Shan Chang, and Minyi Guo. Monoatt: Online monocular 3d object detection with adaptive token transformer. In *CVPR*, 2023. 1