Improving Distant 3D Object Detection Using 2D Box Supervision Supplementary Material

1. Implementation Detail

1.1. Model Setting

Monocular Baseline. We use FCOS3D [1] as our monocular baseline, of which the implementation is almost the same as the official released one on MMDetection3D [2] at GitHub:fcos3d. Specifically, we use ResNet101 [3] with deformable convolutions [4] as the backbone network, and deploy FPN [5] as the neck to produce multi-scale features. For head, there are five branches with 256 channels for 3D object detection. They are responsible for classification, centerness prediction, 3D box generation, attribute prediction and velocity prediction, respectively. For the KITTI, Cityscapes3D, Waymo and Argoverse 2 experiments, we remove the attribute and velocity branches. For the nuScenes dataset, we keep the exact structure.

LR3D. To deploy the IP-Head, based on FCOS3D, we replace its original depth regression module with three modules for 2D detection f_{2d}, 2D box positional encoding f_{PE}, and weight generation f_g.

In our implementation, f_{2d} is composed of two 3x3 convolutions with channels of 256 and 4 to predict 2D bounding boxes. f_{PE}, which encodes the predicted 2D box sizes on images to high-dimensional embeddings, is a fixed positional encoding [6, 7] with the output channel as 16. f_g includes two 3x3 convolutions with channels of 256 and 272 to extract weights of $f^{(\theta)}$ which is a 2-layer perceptron with channels of 16 and 1 ($272 = 16 \times 16 + 16 \times 1$) to transfer 2D box embeddings to corresponding depth predictions.

Long-range Teacher. For long-range teacher experiments, we utilize LR3D with distance and score thresholds of 40m and 0.1, respectively, to generate pseudo distant 3D annotations to train student models. Note that, for all experiments unless further specifications, we limit the farthest available 3D annotations to 40m.

1.2. Training Schedule

Experiments on the KITTI dataset We train the FCOS3D baseline and our LR3D model using the same training schedule. They are trained with SGD optimizer under the initial learning rate as 1e-3 for total 48 epochs with a batch size of 24 equally distributed on 8 GPUs. We adopt the step-wise learning rate decay strategy and decay the learning rate by 0.1 after 32 epochs and 44 epochs. Random flipping is adopted during training as augmentation strategy.

Experiments on the nuScenes dataset For nuScenes experiments, we follow the same training schedule as GitHub:fcos3d. We train our models with SGD optimizer with an initial learning rate of 2e-3 for 12 epochs. These models are trained on 8 GPUs with a total batch size of 16. We decay the learning rate at 8 and 11 epochs with a rate of 0.1. We adopt random flipping as the data augmentation.

Experiments on the Cityscapes dataset For Cityscapes3D experiments, we use the same training schedule as KITTI experiments. Specifically, We train our models with SGD optimizer with an initial learning rate of 1e-3 for 48 epochs. The total batch size is 24. We decay the learning rate by 0.1 after training 32 epochs and 44 epochs, respectively.

Experiments on the Waymo dataset For experiments on Waymo, we base our implementation on the codebase of GitHub:DFM. We use the same training schedule as the officially provided FCOS3D on Waymo Open Dataset. We train our model for 24 epochs, with a total batch size of 24. The model is optimizer by SGD with an initial learning rate of 8e-3 which is then decayed by 0.1 after 16 and 22 epochs.

Experiments on the Argoverse 2 For Argoverse 2 experiments, we train the FCOS3D with our IP-Head for total 12 epochs. The initial learning rate is 2e-3, and the optimizer is SGD. We use random flipping as the data augmentation.

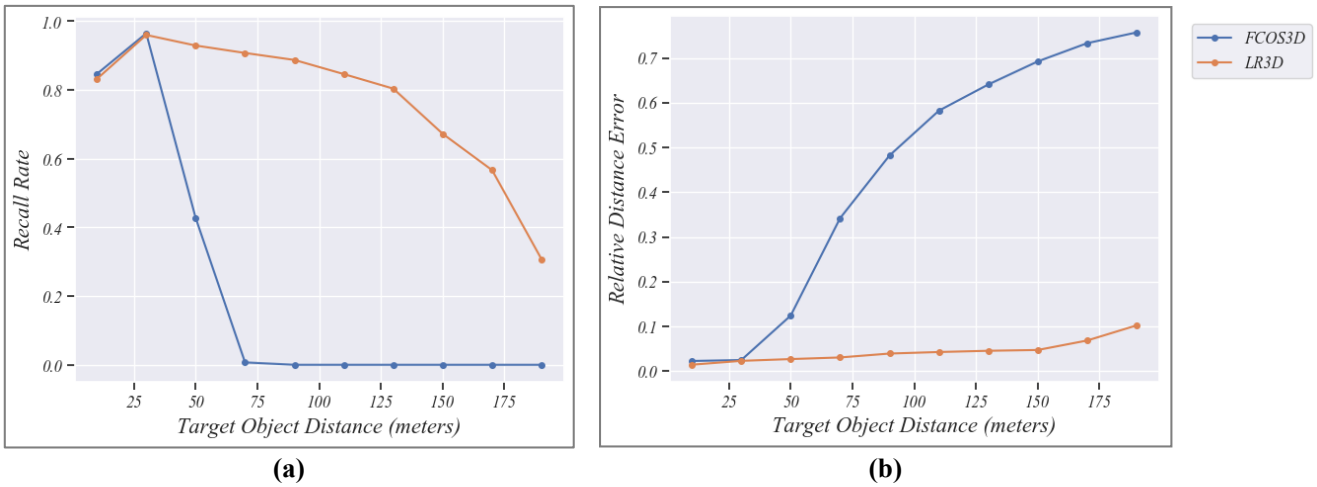
Long-range Teacher. The training schedules and the architectures of student models [8–14] are exactly the same as the official settings without modifications. The only difference is that, for distant 3D annotations of objects over a specific range (40m for KITTI, nuScenes, Cityscapes3D and Argoverse 2; 50m for Waymo Open Dataset), we use the 3D predictions of LR3D instead of the ground truth.

2. Results on Cityscapes3D

Cityscapes3D [15] provides high-quality 3D labels for extremely distant objects among which the farthest instance is more than 250m. It includes 3D box annotations of in-

Method	Distant 3D Groundtruth?	Overall		Close (0m-40m)		Distant (40m-80m)		Distant (80-Inf)	
		LDS (%)	mAP (%)	LDS (%)	mAP (%)	LDS (%)	mAP (%)	LDS (%)	mAP (%)
FCOS3D [1]	√	54.15	47.13	60.36	53.05	50.40	40.49	38.44	29.91
FCOS3D [1]	-	27.85	21.98	59.23	48.99	4.59	2.13	0.0	0.0
LR3D (IP-FCOS3D)		**51.85**	**49.09**	62.86	57.35	**46.82**	**36.08**	**34.54**	**34.88**

Table 1. Comparison on FCOS3D with and without IP-Head supervised by distant 2D ground truth only on the Cityscapes3D validation dataset. Their fully supervised counterparts (with distant 3D ground truth) are also illustrated.



(a) **(b)**

Figure 1. The recall rate and the relative distance error breakdowns of target objects with different depth. Models for comparisons are trained with 3D bounding box annotations within 40m only. **(a).** For objects beyond the 3D supervision range (40m), FCOS3D can hardly retrieve their 3D bounding boxes leading to a huge drop on the recall rate; while our LR3D can well detect those remote objects even much beyond the 3D supervision range. **(b).** The distance estimation error goes larger for FCOS3D when the target objects are beyond the 3D supervision range; while LR3D is capable of accurate distance estimation of far away objects, even though they are beyond the 3D supervision range.

stances in 6 classes, labels 2,975 scenes for training, and leaves 500 scenes for validation. To further analyze the feasibility of only using 2D annotations for detecting 3D objects in farther areas, *e.g.*, farther than 80m, we test LR3D on the Cityscapes3D dataset for quantitative validation.

Similar to the setting of KITTI [16] dataset, we compare the performance of FCOS3D [1] with and without our IP-Head design on instances labeled as "car". We manually mark those objects farther than 40m as distant objects, remove their 3D annotations, and only use their 2D annotations during training. For objects closer than 40m, we keep both their 2D and 3D bounding box labels for training.

As shown in Table 1, compared to the original FCOS3D design, our LR3D outperforms it by a large margin on detecting distant 3D objects beyond the 3D supervision range (40m), *i.e.*, 42.23% LDS improvements for instances from 40m to 80m, and 34.54% LDS improvements for instances farther than 80m. Even compared to the model supervised by human labeled high-quality distant 3D annotations, our LR3D still yields competitive performance.

In Figure 1, we show the breakdowns of target objects with different distance. Specifically, we first cluster objects into 10 groups through their ground truth depth. Objects in each group are within 20m apart. Then, we utilize FCOS3D and LR3D, trained with 3D annotations within 40m only, to obtain 3D detection, and compare the recall rate (Figure 1 **(a)**) and the relative distance error (Figure 1 **(b)**) for each group. As illustrated, when the target objects are beyond the 3D supervision range (40m), the performance of FCOS3D drops significantly. In contrast, our LR3D is still capable of accurately detecting those remote objects, with higher recall rate and smaller relative distance error.

These experimental results show that, by only using 3D annotations within 40m, LR3D is capable of producing accurate 3D detection for objects further than 80m, even up to 250m, which demonstrates the feasibility of our proposal for long-range 3D detection. Moreover, for distant objects with few or no interior points, since drawing their 2D bounding boxes on images is much easier than labeling their 3D annotations, our proposal benefits the scalibility of applications related to long-range 3D detection.

3. Results on Waymo Open Dataset

We further conduct experiments on Waymo Open Dataset [18]. For experiments on the Waymo dataset, in order to keep consistent with the official Waymo testing distance

Method	Distant 3D Groundtruth?	Vehicle (LDS)			Pedestrian (LDS)			Cyclist (LDS)		
		0-30m (%)	30-50m (%)	50-Inf (%)	0-30m (%)	30-50m (%)	50-Inf (%)	0-30m (%)	30-50m (%)	50-Inf (%)
FCOS3D [1]	√	37.64	15.99	12.50	62.00	53.46	46.57	45.20	30.74	20.02
FCOS3D [1]	-	37.81	16.06	2.75	62.20	54.34	5.98	45.47	31.64	2.61
LR3D (IP-FCOS3D)	-	37.56	15.40	**11.58**	61.71	53.18	**38.52**	45.24	30.53	**18.51**
Long-range Teacher										
MV-FCOS3D++ [14]	√	44.43	19.44	11.55	71.24	63.85	55.61	54.19	39.04	26.40
MV-FCOS3D++ [14]	-	44.92	19.29	2.19	71.26	63.94	3.94	54.59	39.97	1.25
+LR3D teacher	-	43.76	18.45	**9.28**	70.41	62.29	**47.22**	53.85	38.77	**23.51**

Table 2. Comparisons of Long-range Detection Score (LDS) on state-of-the-art methods with and without IP-Head or LR3D teacher supervised by distant 2D ground truth only on the Waymo Open Dataset. Their fully supervised counterparts (with distant 3D ground truth) are also illustrated.

Method	Distant 3D Groundtruth?	Vehicle (LET-3D-AP)			Pedestrian (LET-3D-AP)			Cyclist (LET-3D-AP)		
		0-30m (%)	30-50m (%)	50-Inf (%)	0-30m (%)	30-50m (%)	50-Inf (%)	0-30m (%)	30-50m (%)	50-Inf (%)
FCOS3D [1]	√	75.50	60.51	43.14	60.79	40.45	20.60	45.02	14.67	8.95
FCOS3D [1]	-	75.43	61.24	5.71	60.20	41.45	3.14	44.77	14.33	2.60
LR3D (IP-FCOS3D)	-	75.53	60.79	**33.19**	60.88	40.32	**15.90**	45.32	14.11	**5.91**
Long-range Teacher										
MV-FCOS3D++ [14]	√	86.66	69.34	48.75	66.24	35.56	14.34	53.63	14.64	6.93
MV-FCOS3D++ [14]	-	86.73	69.61	5.34	66.37	36.91	2.92	53.92	14.07	3.66
+LR3D teacher	-	86.50	68.99	**37.19**	66.17	34.95	**11.92**	53.60	13.26	**6.77**

Table 3. Comparisons of LET-3D-AP [17], the official Waymo Open Dataset 3D camera-only detection metric, on state-of-the-art methods with and without IP-Head or LR3D teacher supervised by distant 2D ground truth only on the Waymo Open Dataset. Their fully supervised counterparts (with distant 3D ground truth) are also illustrated.

split (0-30m; 30-50m; over 50m), we label objects further than 50m away as distant objects. During training, we discard the 3D box annotations for distant objects and only use their 2D bounding box annotations as supervision. For objects closer than 50m, we use both their 2D and 3D annotations for training.

We base our implementation on the codebase of GitHub:DFM, utilize FCOS3D as our monocular baseline, and enhance it with IP-Head (IP-FCOS3D) for detecting distant 3D objects using their 2D supervisions only. To further demonstrate the effectiveness of our long-range teacher design, we choose MV-FCOS3D++ [14] as the student model and test its performance training without distant 3D supervision. We evaluate different models among all 3 classes on waymo dataset.

As shown in Table 2, compared to the original FCOS3D which heavily relies on abundant 3D annotations, if distant 3D annotations is missing, our proposed LR3D (IP-FCOS3D) achieves 8.83%, 32.54% and 15.90% LDS improvements on distant "Vehicle", "Pedestrian" and "Cyclist" instances farther than 50m, respectively. Furthermore, when using 3D predictions from LR3D (IP-FCOS3D) as pseudo distant 3D annotations, MV-FCOS3D++ [14] outperforms itself, which is trained with only distant 2D annotations, by 7.09% LDS on vehicle instances beyond 50 meters. These experiments further demonstrate the effectiveness of our proposed LR3D in detecting distant 3D objects using 2D box supervision.

We also tabulate the quantitative comparisons of different methods under LET-3D-AP [17] metric, the official Waymo Open Dataset 3D camera-only detection evaluation metric, in Table 3. We have to mention that, the official LET-3D-AP only counts objects within the range of 75 meters. Any ground-truth object farther than this range is ignored during the evaluation. Therefore, LET-3D-AP evaluation only considers partial distant 3D objects, while results in Table 2 are calculated by taking all distant 3D annotations into count. As shown, under the LET-3D-AP, our LR3D still shows superior performance compared to the baseline models on detecting distant 3D objects without using their corresponding 3D box annotations.

4. Results on Argoverse 2 Dataset

Finally, we report our experimental results on Argoverse 2 dataset [19]. Given the absence of a publicly available codebase for camera-based 3D object detection on the Argoverse 2 dataset, we re-implement FCOS3D [1] on this dataset. We then conduct comparisons between FCOS3D with and without the IP-Head, using our own baseline as a reference.

We mark objects farther than 40m as distant objects, re-

Method	Distant 3D Groundtruth?	Overall		Close (0m-40m)		Distant (40m-80m)		Distant (80-Inf)	
		LDS (%)	mAP (%)	LDS (%)	mAP (%)	LDS (%)	mAP (%)	LDS (%)	mAP (%)
FCOS3D [1]	√	16.30	10.91	12.77	5.69	19.64	13.83	20.29	16.55
FCOS3D [1]	-	5.81	3.08	12.60	5.38	4.43	2.90	0.00	0.00
LR3D (IP-FCOS3D)		**15.17**	**9.59**	12.31	5.95	**18.95**	**12.18**	**14.17**	**11.63**

Table 4. Comparison on FCOS3D with and without IP-Head supervised by distant 2D ground truth only on the Argoverse 2 validation dataset. Their fully supervised counterparts (with distant 3D ground truth) are also illustrated.

Score Threshold	Overall (0m-51.2m) LDS (%)	Close (0m-40m) LDS (%)	Distant (40m-51.2m) LDS (%)
0.05	27.8	29.4	13.7
0.1	**29.9**	31.5	13.5
0.2	29.3	31.5	10.5
0.3	28.8	31.7	7.3

Table 5. Effect of the score threshold in LR3D teacher to remove redundant 3D pseudo labels.

Depth Selection Strategy	No. of Augmented Depth	Overall (0m-Inf) LDS (%)	Close (0m-40m) LDS (%)	Distant (40m-Inf) LDS (%)
linspace	10	49.6	51.9	35.4
	20	49.9	52.1	36.1
	40	49.5	51.8	35.0
random	3	**50.0**	52.1	36.2
	10	49.9	51.9	36.2

Table 6. Effect of depth selection strategies and numbers of augmented depth in the projection augmentation.

move their 3D annotations, and only use their 2D annotations for training. For other objects within the 40m range, we utilize their complete annotations, including both 2D and 3D box annotations, for training purposes.

To facilitate the training on the extensive Argoverse 2 dataset, we opt to train our model on a reduced training set comprising only 1/6 of the original Argoverse 2 sensor dataset. We focus specifically on instances labeled as "Regular_vehicle". The experimental results are listed in Table 4, which demonstrates a consistent improvement of our LR3D on boosting existing camera-based 3D detectors to detect distant 3D objects using 2D annotations only.

5. Additional Ablation Studies

Analysis on the Long-range Teacher. In this paragraph, we analyze the impact of score thresholds on removing redundant pseudo distant 3D labels generated by long-range teachers. These experiments are conducted on nuScenes dataset with a quarter of training data. The teacher and the student models are LR3D (IP-FCOS3D) and BEVFormer-S [13], respectively. As illustrated in Table 5, with the score threshold as 0.1, the student model achieves the best balance of 3D detection on close areas and distant areas.

Hyper-Parameters on Projection Augmentation. Projection augmentation generates extra b_{2d}-d training pairs

minimum augmented depth (meters)	maximum augmented depth (meters)	Overall (0m-Inf) LDS (%)	Close (0m-40m) LDS (%)	Distant (40m-Inf) LDS (%)
40	60	49.4	51.6	35.1
40	80	**50.0**	52.1	36.2
40	120	49.9	52.0	36.2

Table 7. Effect of different minimum and maximum augmented depth thresholds in the projection augmentation.

to ensure the network f_g to generate correct implicit inverse function $f^{(\theta)}$. In our implementation, for each close object, we randomly choose 3 depth values from the minimum augmented depth threshold (40m for experiments on all datasets) to the maximum threshold (80m for KITTI, nuScenes, and Waymo; 200m for Argoverse 2 and Cityscapes3D) as augmented depth, calculate the associated projected 2D boxes on image, and utilize those augmented pairs together with ground truth pairs for training.

In this paragraph, we evaluate two different strategies for obtaining augmented depth. One is "random" selection, our default implementation. Another one is "linspace" selection, which returns evenly spaced depth values from the minimum augmented depth threshold to the maximum augmented depth. We conduct this ablation study on KITTI dataset, where the farthest 3D object is typically within 80 meters. According to Table 6, the "random" strategy with 3 randomly augmented depth values performs the best. In Table 7, we further test the impact of different minimum and maximum thresholds for generating augmented depth values on the projection augmentation. As illustrated, the larger the range of randomly augmented depth is, the more precise the implicit inverse function estimates. Therefore, we finally choose the typical distance of the farthest annotated 3D objects as the maximum augmented depth thresholds for each dataset specifically (80m for KITTI, nuScenes, and Waymo; 200m for Argoverse 2 and Cityscapes3D), and set 40m as the minimum depth thresholds for all.

Compared with Object Distance Estimation Methods. In this paper, we propose IP-Head for depth estimation of distant objects without 3D supervision. As a related task, long-range object distance estimation estimates the distance of far away objects, of which IP-Head can serve as the solution, yet without the need of annotations. We conduct experiments to evaluate its effectiveness, and compare it with existing methods on KITTI dataset. We mark objects

Method	LiDAR	Long-range Distance Ground truth?	<5%↑	<10%↑	<15%↑	Abs Rel↓	Sq Rel↓	RMSE↓	RMSE_{log}↓
Zhu & Fang [20]	√	√	41.1	66.5	78.0	8.9	0.97	8.1	0.136
R4D [21]			46.3	72.5	83.9	7.5	0.68	6.8	0.112
LR3D (IP-FastRCNN3D)	-	-	**47.2**	**77.9**	**92.5**	**6.3**	**0.33**	**4.3**	**0.080**

Table 8. Comparisons with state-of-the-art object distance estimation methods of distant objects on KITTI Dataset.

over 40m as distant objects, and compare the performance of different methods for distant objects following [21]. To align the settings of state-of-the-art methods [20, 21], we adopt IP-FastRCNN3D as the detector of LR3D, in which 2D ground truth boxes are used as proposals for distance estimation. As illustrated in Table 8, even without distance annotations for long-range objects and LiDAR hints, LR3D still outperforms existing fully supervised methods.

6. Additional Qualitative Results

In Figure 2, we provide more qualitative results on detecting extremely distant 3D objects. As illustrated, LR3D is capable of accurately detecting 3D bounding boxes of objects over 200m with only 3D supervision within 40m. In Figure 3, we show qualitative comparisons of DID-M3D [12] with and without LR3D long-range teacher on KITTI Dataset [16]. In Figure 4 and Figure 5, we compare the performance of BEVFormer [13] with and without LR3D long-range teacher on nuScenes Dataset [22]. All experiments are conducted under a limited range of available 3D bounding box annotations (40 meters). As shown, LR3D boosts the performance of state-of-the-art 3D detectors on long-range 3D detection and enables them to accurately detect distant objects without corresponding 3D supervision.

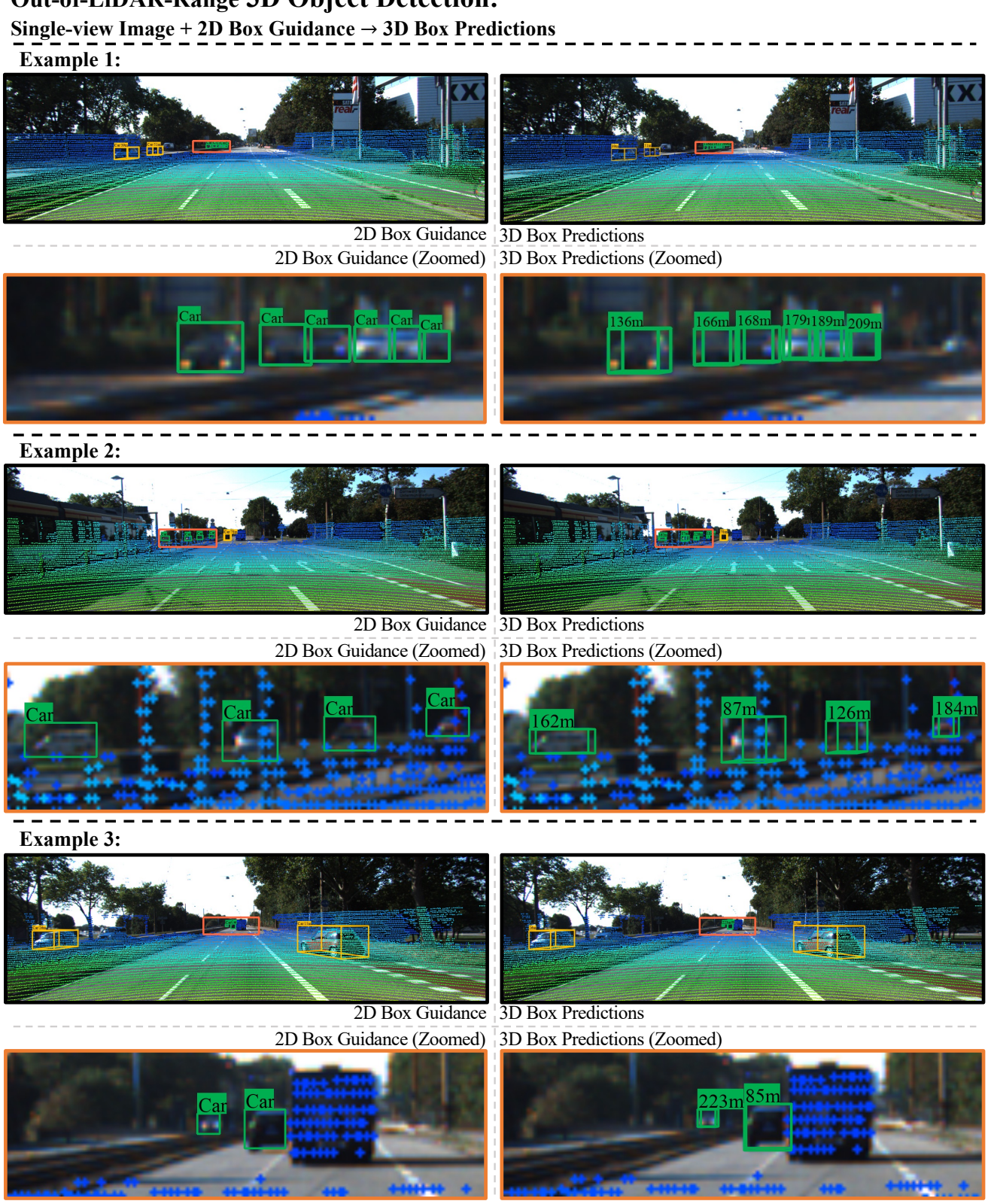
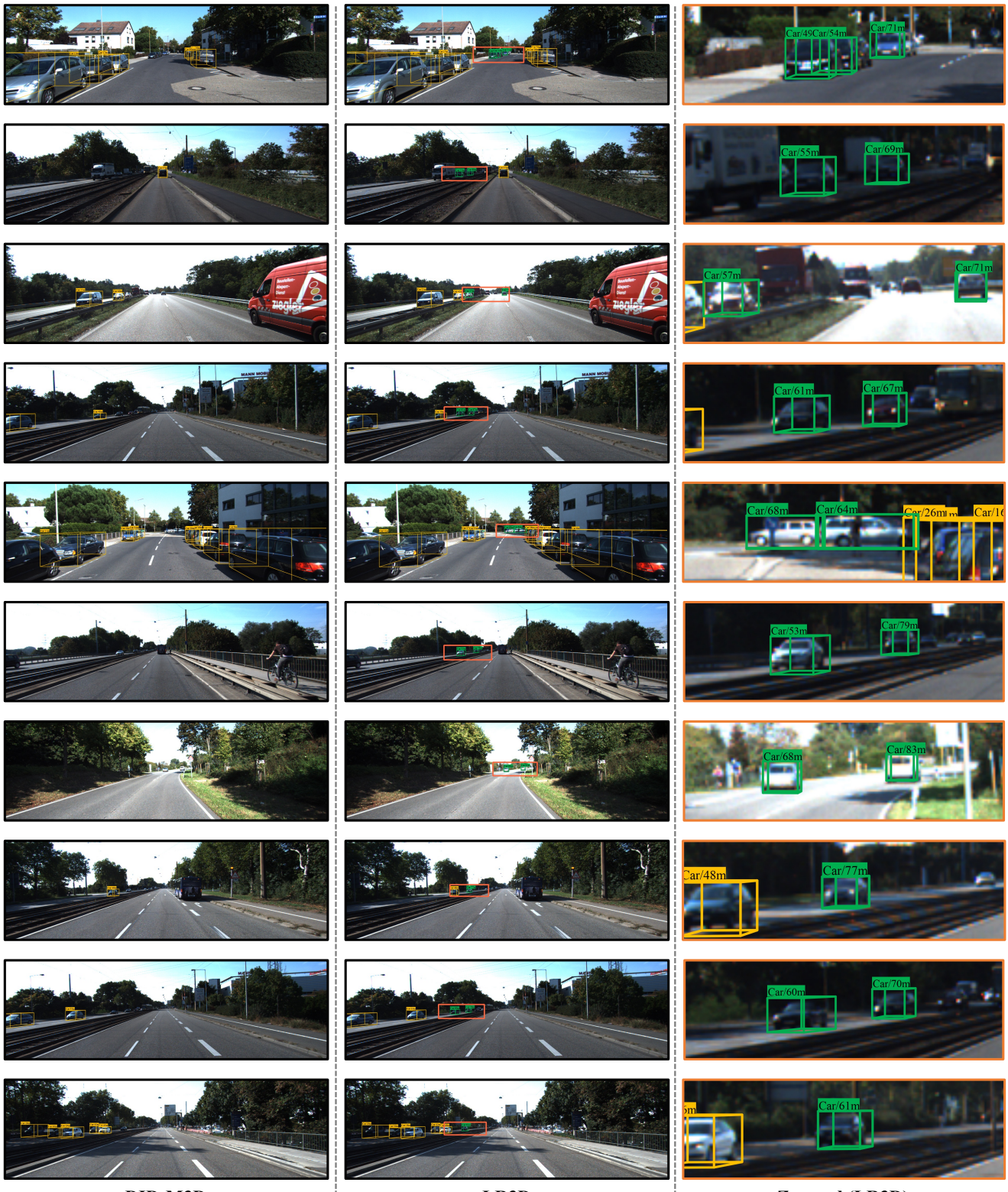


Figure 2. More qualitative results on detecting extremely distant 3D objects. As shown, taking as inputs of 2D bounding box conditions (left), LR3D is capable of predicting associated 3D boxes (right), including locations, sizes and orientations, for distant objects (marked in **green**) much farther than the range of available 3D annotations (40m in our case). LiDAR points are projected on input images with different colors corresponding to different depths.



DID-M3D **LR3D** **Zoomed (LR3D)**

Figure 3. Qualitative comparison results between DID-M3D [12] with (middle, **LR3D**) and without (left, **DID-M3D**) LR3D long-range teacher on detecting distant objects beyond 3D bounding box label range. The newly detected distant objects with their predicted 3D bounding boxes are marked in **green**. As illustrated, though only trained with limited range of available 3D annotations within 40m, LR3D enables DID-M3D to accurately detect distant objects beyond this range.

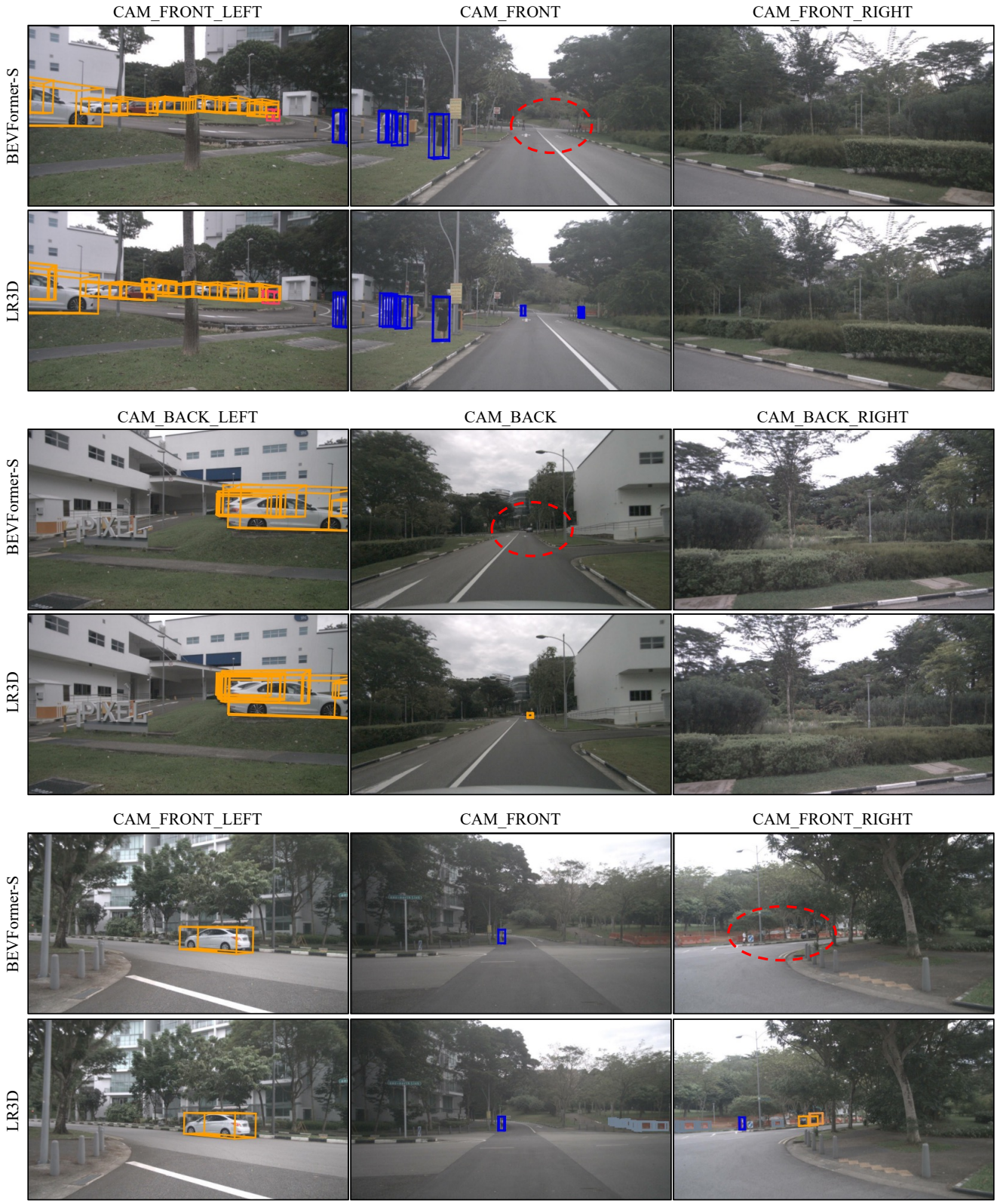


Figure 4. Qualitative comparison results between BEVFormer [13] with (lower, LR3D) and without (upper, BEVFormer-S) LR3D long-range teacher on detecting distant objects beyond 3D bounding box label range. The newly detected distant objects with their predicted 3D bounding boxes are surrounded in **red circle**. As illustrated, though only trained with limited range of available 3D annotations within 40m, LR3D enables BEVFormer to accurately detect distant objects beyond this range on nuScenes Dataset.

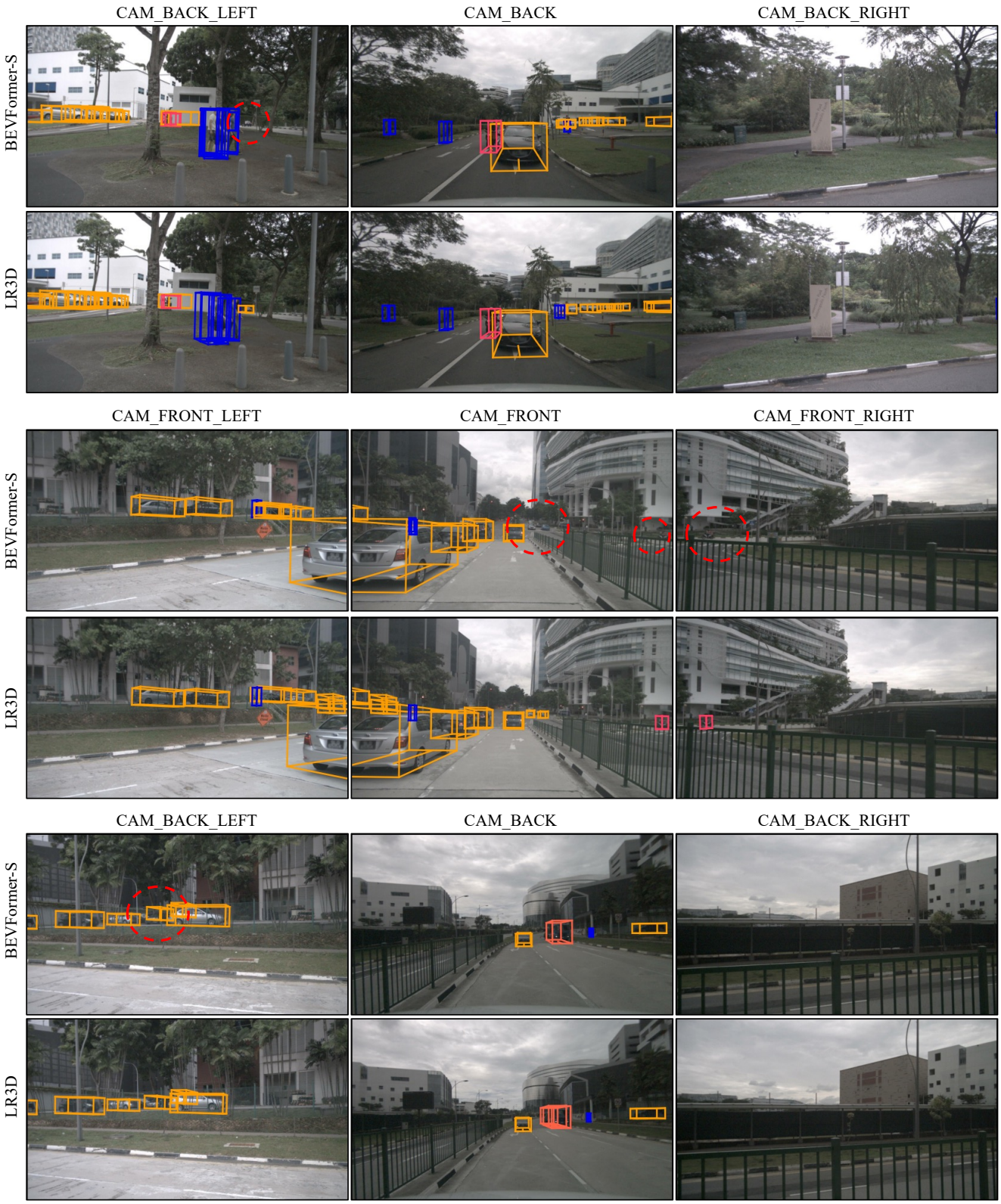


Figure 5. More qualitative comparison results between BEVFormer [13] with (lower, LR3D) and without (upper, BEVFormer-S) LR3D long-range teacher on detecting distant objects beyond 3D bounding box label range (40 meters).

References

[1] Tai Wang, Xinge Zhu, Jiangmiao Pang, and Dahua Lin. FCOS3D: Fully convolutional one-stage monocular 3d object detection. In *ICCV Workshops*, 2021. 1, 2, 3, 4

[2] MMDetection3D Contributors. MMDetection3D: OpenMMLab next-generation platform for general 3D object detection. https://github.com/open-mmlab/mmdetection3d, 2020. 1

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

[4] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. *ICCV*, 2017. 1

[5] Tsung-Yi Lin, Piotr Dollár, Ross B. Girshick, Kaiming He, Bharath Hariharan, and Serge J. Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017. 1

[6] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *NIPS*, 2017. 1

[7] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *ECCV*, 2020. 1

[8] Cody Reading, Ali Harakeh, Julia Chae, and Steven L. Waslander. Categorical depth distributionnetwork for monocular 3d object detection. *CVPR*, 2021. 1

[9] Danila Rukhovich, Anna Vorontsova, and Anton Konushin. Imvoxelnet: Image to voxels projection for monocular and multi-view general-purpose 3d object detection. In *WACV*, 2022.

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

[11] 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. *arXiv preprint arXiv:2107.13774*, 2021.

[12] Liang Peng, Xiaopei Wu, Zheng Yang, Haifeng Liu, and Deng Cai. Did-m3d: Decoupling instance depth for monocular 3d object detection. In *ECCV*, 2022. 5, 7

[13] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Yu Qiao, and Jifeng Dai. Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. *arXiv preprint arXiv:2203.17270*, 2022. 4, 5, 8, 9

[14] Tai Wang, Qing Lian, Chenming Zhu, Xinge Zhu, and Wenwei Zhang. MV-FCOS3D++: Multi-View camera-only 4d object detection with pretrained monocular backbones. *arXiv preprint*, 2022. 1, 3

[15] Nils Gählert, Nicolas Jourdan, Marius Cordts, Uwe Franke, and Joachim Denzler. Cityscapes 3d: Dataset and benchmark for 9 dof vehicle detection. *arXiv preprint arXiv:2006.07864*, 2020. 1

[16] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The KITTI dataset. *I. J. Robotics Res.*, 2013. 2, 5

[17] Wei-Chih Hung, Henrik Kretzschmar, Vincent Casser, Jyh-Jing Hwang, and Dragomir Anguelov. Let-3d-ap: Longitudinal error tolerant 3d average precision for camera-only 3d detection. *arXiv preprint*, 2022. 3

[18] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, Vijay Vasudevan, Wei Han, Jiquan Ngiam, Hang Zhao, Aleksei Timofeev, Scott Ettinger, Maxim Krivokon, Amy Gao, Aditya Joshi, Yu Zhang, Jonathon Shlens, Zhifeng Chen, and Dragomir Anguelov. Scalability in perception for autonomous driving: Waymo open dataset. In *CVPR*, 2020. 2

[19] Benjamin Wilson, William Qi, Tanmay Agarwal, John Lambert, Jagjeet Singh, Siddhesh Khandelwal, Bowen Pan, Ratnesh Kumar, Andrew Hartnett, Jhony Kaesemodel Pontes, Deva Ramanan, Peter Carr, and James Hays. Argoverse 2: Next generation datasets for self-driving perception and forecasting. In *NeurIPS Datasets and Benchmarks 2021*, 2021. 3

[20] Jing Zhu and Yi Fang. Learning object-specific distance from a monocular image. In *ICCV*, 2019. 5

[21] Yingwei Li, Tiffany Chen, Maya Kabkab, Ruichi Yu, Longlong Jing, Yurong You, and Hang Zhao. R4D: utilizing reference objects for long-range distance estimation. In *ICLR*, 2022. 5

[22] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. *CVPR*, 2020. 5