Multivariate Probabilistic Monocular 3D Object Detection
Supplementary File

	[LEVEL_1 / LEVEL_2] ↑	
	Overall AP	Overall APH
MonoRCNN [6]	10.45 / 9.92	10.39 / 9.86
MonoRCNN++ (ours)	11.37 / 10.79	11.31 / 10.73

Table 1: **Comparisons on the Waymo Open val set** [7]. We evaluate on the vehicle class and use overall 3D AP and 3D APH (IoU > 0.5) as metrics.

1. Learned Covariances on Waymo

In the main paper, we show how learned covariances behave on the KITTI dataset [1]. To show the scalability of our method, we further show how learned covariances behave on the Waymo Open dataset [7] in Fig. 1. We can see predicted covariances are negative and their magnitudes increase with the increase of the distance. This shows our model can also predict covariances as expected effectively on the much more diverse and challenging Waymo Open dataset [7]. We also show predicted uncertainties in Fig. 2. We can see predicted uncertainties are larger for distant small objects. Moreover, we notice that the magnitudes of covariances and uncertainties on the Waymo Open dataset [7] are larger than those on the KITTI dataset [1]. We assume this is because Waymo [7] is much more diverse and challenging than KITTI [1].

2. Results of MonoRCNN on Waymo

We evaluate MonoRCNN [6] on the Waymo dataset [7] , shown in Tab. 1. For 3D AP (LEVEL_1 / LEVEL_2), our method outperforms MonoRCNN [6] by $8.80\%/8.77\%$. For 3D APH (LEVEL_1 / LEVEL_2), our method outperforms MonoRCNN [6] by $8.85\%/8.82\%$. This shows that our method is more accurate than MonoRCNN [6].

3. More Results

In the main paper, we compare our method with other methods using the average precision (AP). To show the superiority of our method, we further compare using a more strict metric, i.e., the average precision weighted by heading (APH). Following [8, 3, 2], we benchmark using the vehicle class on the val set of the Waymo Open dataset [7], shown in Tab. 2. We can see MonoRCNN++ still achieves the best accuracy under the APH metric. 1) When the IoU threshold is 0.7, our method achieves the best overall 3D AP and surpasses the second [2] by a large margin. Specifically, MonoRCNN++ surpasses DEVIANT [2] by 59.55% / 61.60% in LEVEL_1 / LEVEL_2, respectively. This shows our MonoRCNN++ is significantly better than GUPNet [4], DEVIANT [2], and MonoJSG [3] under the strict evaluation (APH with IoU > 0.7). Our method also achieves the best accuracy for nearby objects within 30 meters, and the second best accuracy for objects beyond 30 meters. 2) When the IoU threshold is 0.5, our method achieves the best overall 3D AP. For nearby objects within 30 meters, our method also achieves the best accuracy. For faraway objects beyond 50 meters, our method achieves the second best accuracy. We also visualize some qualitative examples in Fig. 3.

References

[1] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the KITTI vision benchmark suite. In *CVPR*, 2012.

[2] Abhinav Kumar, Garrick Brazil, Enrique Corona, Armin Parchami, and Xiaoming Liu. DEVIANT: Depth EquiVarIAnt NeTwork for Monocular 3D Object Detection. In *ECCV*, 2022.

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

[4] 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.

[5] Xinzhu Ma, Shinan Liu, Zhiyi Xia, Hongwen Zhang, Xingyu Zeng, and Wanli Ouyang. Rethinking pseudo-lidar representation. In *ECCV*, 2020.

[6] Xuepeng Shi, Qi Ye, Xiaozhi Chen, Chuangrong Chen, Zhixiang Chen, and Tae-Kyun Kim. Geometry-based distance decomposition for monocular 3d object detection. In *ICCV*, 2021.

[7] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, Vijay Vasudevan, Wei Han,

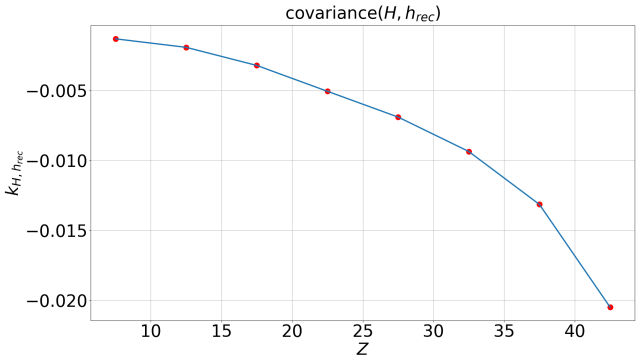

Figure 1: **Predicted covariances** of the vehicle class on the Waymo Open val set [7]. We uniformly divide the distance range into 8 intervals and show the average covariance of each interval. Predicted covariances are negative and their magnitudes increase with the increase of the distance Z.

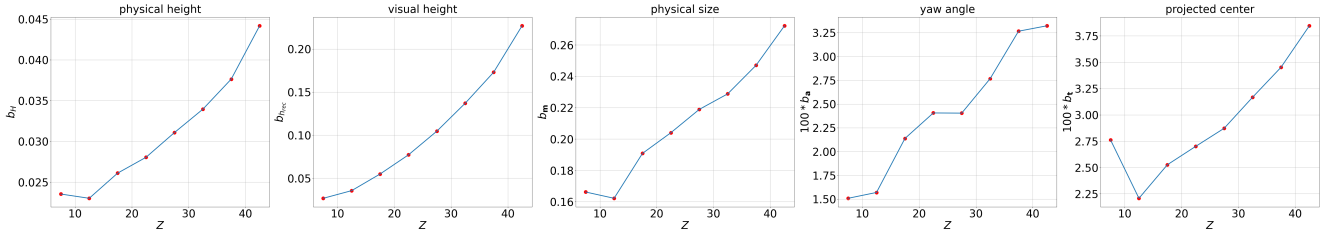

Figure 2: **Predicted uncertainties** of the vehicle class on the Waymo Open val set [7]. We uniformly divide the distance range into 8 intervals and show the average uncertainty of each interval. Predicted uncertainties are larger for faraway objects.

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.

[8] Li Wang, Li Zhang, Yi Zhu, Zhi Zhang, Tong He, Mu Li, and Xiangyang Xue. Progressive coordinate transforms for monocular 3d object detection. In *NeurIPS*, 2021.

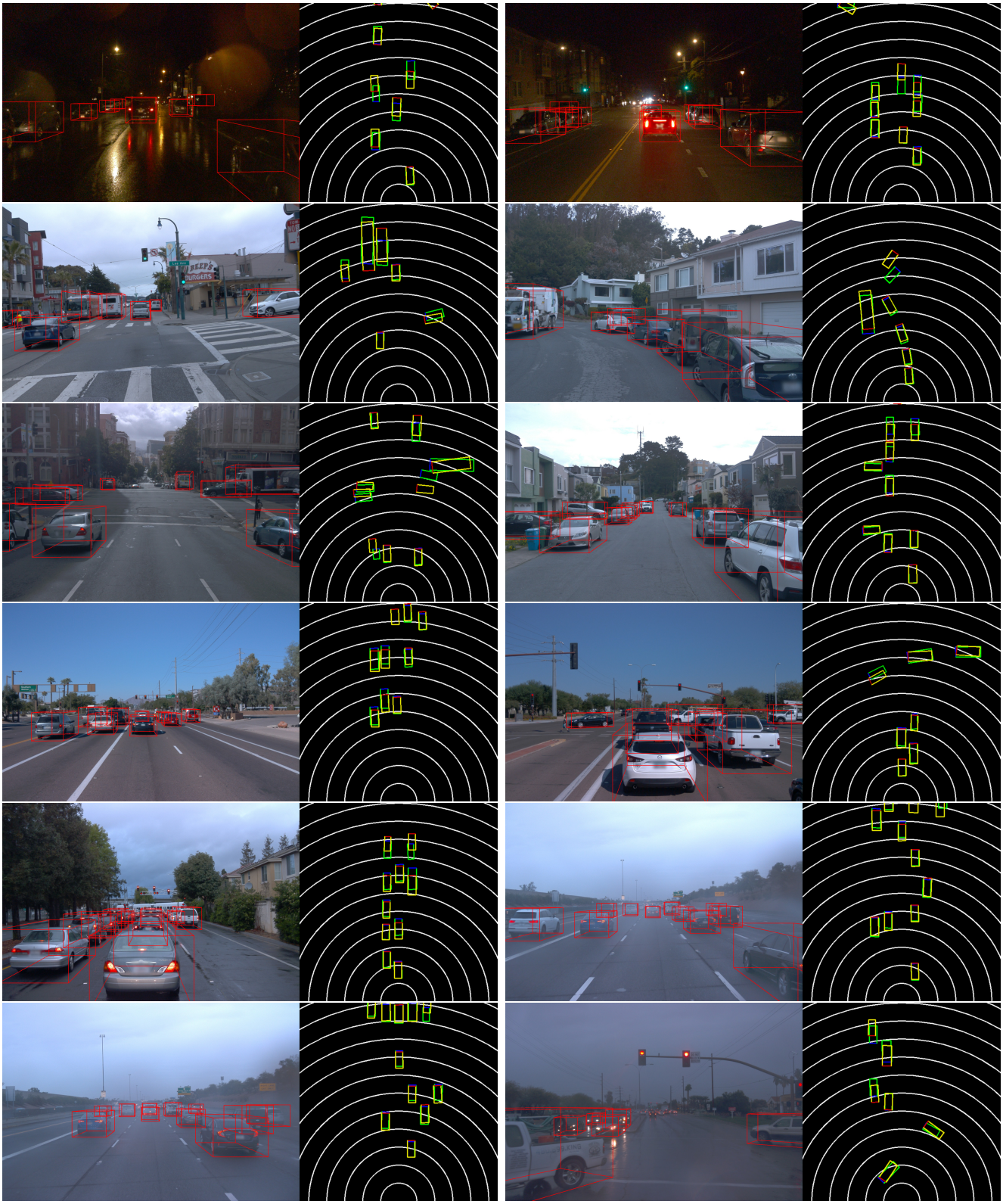

Figure 3: **3D detection results of MonoRCNN++** on the val set of the Waymo Open dataset [7]. MonoRCNN++ predicts accurate 3D bounding boxes for various challenging cases. The red boxes in the image planes represent the 2D projections of the predicted 3D bounding boxes. The yellow / green boxes in the bird's eye views represent the predictions and groundtruths, respectively, and the red / blue lines indicate the yaw angle. The radius difference between two adjacent white circles is 5 meters.

Method	Input	LEVEL_1 (IoU >0.5) $\uparrow$				LEVEL_2 (IoU >0.5) $\uparrow$			
		Overall	0 - 30m	30 - 50m	50m - ∞	Overall	0 - 30m	30 - 50m	50m - ∞
PatchNet (ECCV 20) [5]	I+D	2.74	9.75	0.96	0.18	2.28	9.73	0.94	0.16
PCT (NeurIPS 21) [8]	I+D	4.15	14.54	1.75	0.39	3.99	14.51	1.71	0.35
GUPNet (ICCV 21) [4]	I	9.94	24.59	4.78	0.22	9.31	24.50	4.62	0.19
MonoJSG (CVPR 22) [3]	I	5.47	20.26	3.79	0.92	5.17	20.19	3.67	0.82
DEVIANT (ECCV 22) [2]	I	10.89	26.64	5.08	0.18	10.20	26.54	4.90	0.16
MonoRCNN++ (Ours)	I	11.31	27.81	4.04	0.42	10.73	27.74	3.95	0.38
Method	Input	LEVEL_1 (IoU >0.7) $\uparrow$				LEVEL_2 (IoU >0.7) $\uparrow$			
		Overall	0 - 30m	30 - 50m	50m - ∞	Overall	0 - 30m	30 - 50m	50m - ∞
PatchNet (ECCV 20) [5]	I+D	0.37	1.63	0.12	0.03	0.36	1.63	0.11	0.03
PCT (NeurIPS 21) [8]	I+D	0.88	3.15	0.27	0.07	0.66	3.15	0.26	0.07
GUPNet (ICCV 21) [4]	I	2.27	6.11	0.80	0.03	2.12	6.08	0.77	0.02
MonoJSG (CVPR 22) [3]	I	0.95	4.59	0.53	0.09	0.89	4.65	0.53	0.09
DEVIANT (ECCV 22) [2]	I	2.67	6.90	0.98	0.02	2.50	6.87	0.94	0.02
MonoRCNN++ (Ours)	I	4.26	9.80	0.90	0.09	4.04	9.76	0.88	0.08

Table 2: **Comparisons on the Waymo Open val set** [7]. We evaluate on the vehicle class and use 3D APH (IoU >0.5 and 0.7) as metric. 'Input' means the input data modality used during training and inference. 'I' denotes image and 'D' denotes depth. Red / blue indicate the best / second, respectively. The results of [5] and [4] are from [8] and [2], respectively.