MonoCD: Monocular 3D Object Detection with Complementary Depths - Supplementary Material

Longfei Yan[1] Pei Yan[1] Shengzhou Xiong[1] Xuanyu Xiang[1] Yihua Tan[1*]
[1]Hubei Engineering Research Center of Machine Vision and Intelligent Systems,
School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, China
{longfeiyan, yanpei}@hust.edu.cn, xiongshengzhou@126.com, {xuanyuxiang, yhtan}@hust.edu.cn

A. Cross-Dataset Evaluation

To demonstrate the generalizability of our proposed method, we conduct cross-dataset evaluations on KITTI [3] and nuScenes [2] datasets. Following [9], our model is trained on the KITTI training set (3712 images), and evaluated on KITTI (3769 images) and nuScenes frontal (6019 images) validation sets. We also provide the results of retraining MonoCon [6] using the official code but unrestricted from training on distant objects ($z > 65m$) as a fair comparison with others. To fit the model trained in KITTI, for the nuScenes dataset, we adjusted the image resolution to 384×672 and the ground plane equation prediction preset height to 1.562m (the ego car height in nuScenes [2]). Neither our method nor MonoCon uses normalized coordinates for the direct depth prediction branch and the images of KITTI and nuScenes have different focal lengths which the direct depth prediction relies on. Thus, following [4], we divide their direct predicted depth by 1.361.

The cross-dataset evaluation results are shown in Tab. 1, our method has lower prediction errors at different object depth ranges, which indicates the effectiveness of the proposed complementary depths in improving overall accuracy. In addition, our method outperforms other methods in most of the metrics on both datasets, which demonstrates the generalizability of our method.

Dataset	Method	Depth prediction MAE (meters)↓			
		0-20	20-40	40-∞	All
KITTI	M3D-RPN [1]	0.56	1.33	2.73	1.26
	MonoRCNN [9]	0.46	1.27	2.59	1.14
	GUPNet [7]	0.45	1.10	1.85	0.89
	MonoCon [6]	0.40	1.08	1.78	0.85
	MonoCD(Ours)	**0.37**	**1.04**	**1.72**	**0.83**
nuScenes	M3D-RPN [1]	0.94	3.06	10.36	2.67
	MonoRCNN [9]	0.94	2.84	8.65	2.39
	GUPNet [7]	0.82	1.70	6.20	1.45
	MonoCon [6]	0.78	1.65	6.02	1.40
	MonoCD(Ours)	**0.73**	**1.59**	**5.78**	**1.33**

Table 1. Cross-dataset evaluation on KITTI and nuScenes frontal validation with depth prediction MAE.

B. Discussion on multi-depth prediction methods

Tab. 2 shows some representative multi-depth prediction methods in recent years. The coupling between their multiple branches is shown in the third column of Tab. 2 in terms of Error Sign Opposite Proportions (ESOP). MonoFlex [10] contains 4 depth prediction branches including 1 directly predicted depth and 3 depths shown in the 2nd row of Tab. 2. MonoGround [8] and our method have 3 additional depth branches on top of them. Since the results of the public branches are similar, for MonoGround and our method, Tab. 2 only shows the results of unshared branches.

It can be observed that the error sign of the 3 depths from keypoint and height is similar to the error sign of the directly predicted depths. Benefiting from the wider range of dense depth supervision, the coupling phenomenon of depths from the ground added by MonoGround [8] is mitigated a bit, but it does not eliminate the coupling. Because its dense supervision comes from local sampled values around the bottom of the object. Although the code of MonoDDE [5] has not been released, a similar coupling phenomenon can be inferred based on the local information it uses. However, after our complementary design, the coupling phenomenon is significantly alleviated and the overall performance is further improved.

C. Additional Experiments on the Effect of Complementary Depths

This section supplements the part of Sec. 3.2 in the main paper that is not presented in detail due to space limits. With the analyses in this section, two experimental conclusions can be obtained:

(1) Existing multiple predicted depths suffer from a common problem of lacking complementarity.

Model	Branch dir&	ESOP (%)↑	Val, AP_{3D} Mod.↑
MonoFlex [10]	key0	4.08	17.51
	key1	5.22	
	key2	6.19	
MonoGround [8]	gro0	18.35	18.69
	gro1	20.72	
	gro2	14.73	
MonoCD (Ours)	comp0	38.19	19.37
	comp1	40.24	
	comp2	40.05	

Table 2. Comparison between multiple depth prediction methods. The second column in the table represents the branches used to calculate ESOP with the directly(dir) predicted depth of each model. Including depths from keypoint and height (key), depths from ground (gro), and depths for complementary (comp). Different suffix numbers are used to distinguish the specific branches. The accuracy in the last column is AP_{40} for the moderate Car category at 0.7 IoU threshold on KITTI.

Fipped Branch	Proportion of Flipped Samples 0%	25%	50%	75%	100%
dir	17.51	21.02	25.93	31.69	36.12
key0	17.51	21.06	25.78	31.26	35.87
key1	17.51	20.92	25.55	30.87	35.42
key2	17.51	20.85	25.33	29.76	34.92

Table 3. Perform flipping operation on different depth branches according to different sample proportions on KITTI dataset.

Model	Numbers of Flipped Branches	Val, AP_{3D} Mod.↑
MonoFlex [10]	0	17.51
	1	25.93
	2	35.79
	3	22.95
	4	15.55
MonoGround [8]	0	18.69
	1	20.59
	2	21.79
	3	24.24
	4	32.34
	5	32.75
	6	22.60
	7	17.12

Table 4. Evaluation results of two multi-depth prediction models with different numbers of flipped branches on KITTI dataset, where the proportion of flipped samples is fixed at 50%.

(2) To maximize the complementary effect, it is beneficial to keep prediction branches symmetrical in number.

Setting	Combined Depth prediction MAE (meters) ↓					
	y_{glo} MAE (meters) ↓					overall
	0-0.1	0.1-0.2	0.2-0.3	0.3-0.4	0.4-∞	
	Proportion of samples (%)					
	54.09	27.37	9.61	4.77	4.15	
Baseline	0.90	1.17	1.72	1.84	**2.78**	1.18
MonoCD(Ours)	**0.85**	**1.13**	**1.66**	**1.82**	3.02	**1.14**

Table 5. The system robustness evaluation in KITTI val set, which contains five levels based on the MAE of y_{glo}. The larger the value, the worse the conditions the system faces. The percentage under each level represents the proportion of samples.

C.1. Flip on Different Branch

As shown in Tab. 3, we perform flipping on different branches of MonoFlex [10] according to different flipped sample proportions. The first row of results in the table is presented to the left of Fig. 3 in the main paper. It can be observed that the results of selecting different branches for flipping are similar, which indicates that the coupling between multiple-depth branches is relatively similar and lacking complementarity is common.

C.2. Flip with Different Numbers of Branches

To maximize the complementary effects, we additionally conducted an analytical study on two multi-depth prediction models with different numbers of flipped branches. The results in Tab. 4 show that realizing branch flips with different numbers is effective in improving performance except in the case where all branches are flipped. This is because although the accuracy of the depth prediction does not change with flipping, the depth values will be completely flipped to the other side of the ground truth. According to Eq. (1) in the main paper, it introduces additional error to the predicted x and y, resulting in a decrease in the accuracy of the predicted 3D bounding box.

Furthermore, it is worth noting that both models perform best when the number of flipped branches and the number of unflipped branches are close to the same. This indicates that for multiple depth prediction branches with complementary effects, maintaining a certain level of symmetry in number is preferable to maximize their effectiveness. This is why we follow the number of z_{key} and design three symmetrical z_{comp} in the main paper.

D. System Robustness Evaluation

As we discussed in the limitations of the main paper, the performance of our method is affected by the estimation of the ground plane equation and keypoints. Thus, we conduct a system robustness evaluation to check the performance of our method in severe conditions as shown in Tab. 5. For our added complementary depths, the effect of inaccuracies in ground plane estimation or keypoint detection is directly reflected in the prediction error of y_{glo}. Therefore, we divide

the samples into five levels according to the MAE of y_{glo} and count the mean absolute error of the combined depth at each level. It can be observed that our method outperforms the baseline in most cases, and in a few severe conditions (less than 5%), the performance of our method degrades. This problem will be alleviated by enhancing the understanding of road scenes in the future.

References

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

[2] 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. In *CVPR*, pages 11621–11631, 2020. 1

[3] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, pages 3354–3361. IEEE, 2012. 1

[4] Abhinav Kumar, Garrick Brazil, Enrique Corona, Armin Parchami, and Xiaoming Liu. Deviant: Depth equivariant network for monocular 3d object detection. In *ECCV*, pages 664–683. Springer, 2022. 1

[5] 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*, pages 2791–2800, 2022. 1

[6] Xianpeng Liu, Nan Xue, and Tianfu Wu. Learning auxiliary monocular contexts helps monocular 3d object detection. In *AAAI*, pages 1810–1818, 2022. 1

[7] 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*, pages 3111–3121, 2021. 1

[8] Zequn Qin and Xi Li. Monoground: Detecting monocular 3d objects from the ground. In *CVPR*, pages 3793–3802, 2022. 1, 2

[9] 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*, pages 15172–15181, 2021. 1

[10] Yunpeng Zhang, Jiwen Lu, and Jie Zhou. Objects are different: Flexible monocular 3d object detection. In *CVPR*, pages 3289–3298, 2021. 1, 2