Monocular 3D Object Detection with Bounding Box Denoising in 3D by Perceiver
Supplementary Materials

Overview

In this supplementary material, we provide more details on the following aspects that are not presented in the main paper due to space limit:

- *Results on the KITTI Validation Set* are provided in Sec. A.1.
- *Supplementary ablation studies* are provided in Sec. A.2.
- *Results and analysis on other category* are provided in Sec. A.3.
- *Qualitative failure case study on the KITTI Validation Set* are provided in Sec. B.
- *Qualitative failure case study on the Waymo Validation Set* are provided in Sec. C.
- *Implementation details* are provided in Sec. D.
- *Details of KITTI validation results using 5 random seeds* are provided in Sec. E.

A. Supplementary Results

A.1. Results on the KITTI Validation Set

Comparison with SOTA Methods. We report quantitative results of car category on the KITTI validation set as shown in Tab. 1. It shows that our MonoXiver obtains significant improvements with different backbone detectors including SMOKE [11] and MonoCon [3].

Median/Average Results on KITTI Validation Set. KITTI dataset is known to exist performance fluctuations on its validation set. Therefore, we report median and average results across 5 different runs to avoid randomness following [7]'s settings. The result is shown in Tab. 2. We observe consistent improvement under various evaluation metrics. The details of 5 different runs are provided in E.

AP$_{3D}$ at Different Depth Ranges on KITTI Validation Set. We report results in Tab. 3. Our MonoXiver consistently improves baseline at different depth ranges.

Methods	Extra	$AP_{3D\|R40}$			$AP_{BEV\|R40}$		
		Easy	Mod.	Hard	Easy	Mod.	Hard
Kinematic3D [2]	Temporal	19.76	14.10	10.47	27.83	19.72	15.10
DFM [16]	Temporal & Lidar	<u>29.27</u>	<u>20.22</u>	<u>17.46</u>	<u>38.60</u>	<u>27.13</u>	<u>24.05</u>
DID-M3D [13]	Lidar	22.98	16.12	14.03	31.10	22.76	19.50
CaDDN [14]		23.57	16.31	13.84	-	-	-
MonoJSG [9]		26.40	18.30	15.40	-	-	-
MonoDistill [4]		24.31	18.47	15.76	33.09	25.40	22.16
MonoDTR [6]		24.52	18.57	15.51	33.33	25.35	21.68
MonoFlex [17]	None	23.64	17.51	14.83	-	-	-
GUPNet [12]		22.76	16.46	13.72	31.07	22.94	19.75
DEVIANT [7]		24.63	16.54	14.52	32.60	23.04	19.99
Homography [5]		23.04	16.89	14.90	31.04	22.99	19.84
SMOKE [11]	None	10.43	7.09	5.57	17.62	12.02	10.07
Ours + SMOKE		11.58	9.40	7.75	18.07	14.47	12.01
MonoCon [10]		26.33	19.01	15.98	34.65	25.39	21.93
Ours + MonoCon		**30.48**	**22.40**	**19.13**	**38.77**	**28.67**	**24.89**

Table 1: Quantitative performance of the **Car** category on the KITTI *validation* set. Method are ranked by moderate settings based on 3D detection performance following KITTI leaderboard within each group. We highlight the best results in **bold** and the second place in <u>underline</u>.

Methods	$IoU_{3D} \geq 0.7$			$IoU_{3D} \geq 0.5$		
	Easy	Mod.	Hard	Easy	Mod.	Hard
MonoCon (MonoCon paper)	26.33	19.01	15.98	64.53	47.35	42.49
+ MonoXiver (our main paper)	**30.48** (**+4.15**)	22.40 (+3.39)	19.13 (+3.15)	65.37 (+0.84)	47.12 (-0.23)	41.33 (-1.16)
MonoCon (reproduced, Med.)	25.99	18.98	16.13	65.36	48.33	43.63
+ MonoXiver (Med.)	29.67 (+3.68)	22.40 (3.42)	19.41 (+3.28)	**67.00** (+1.64)	**50.40** (+2.07)	**45.00** (+1.37)
MonoCon (reproduced, Ave.)	25.86	18.92	16.00	64.15	47.49	42.56
+ MonoXiver (Ave.)	29.48 (+3.62)	**22.44** (**+3.52**)	**19.49** (**+3.49**)	66.46 (**+2.31**)	49.60 (**+2.11**)	44.12 (**+1.56**)

Table 2: Median/Average Car category AP_{3D} results on the KITTI *validation* set.

Methods	Easy, $IoU_{3D} \geq 0.7$			Moderate, $IoU_{3D} \geq 0.7$			Hard, $IoU_{3D} \geq 0.7$		
	0−15	15−30	30−45	0−15	15−30	30−45	0−15	15−30	30−45
MonoCon (MonoCon paper)	42.18	7.69	-	43.80	9.01	0.50	36.73	7.97	0.54
MonoXiver (our main paper)	**42.30**	**10.49**	-	**45.47**	**11.28**	**1.04**	**37.92**	**11.03**	**0.96**

Table 3: Car category AP_{3D} result at different depth ranges on the KITTI *validation* set.

Extra Experiments for the Effectiveness of the Perciver We run four more experiments to confirm the effectiveness of the Perciver and the results are shown in Tab. 4.

A.2. Supplementary Ablation Studies

Choice of Top-down Generated Proposal Anchors. We report the result of generating a different number of top-

With Perciver	$IoU_{3D} \geq 0.7$ Easy	 Mod.	 Hard
- (our main paper)	29.33	21.67	18.46
✓(our main paper)	**30.48** (+1.15)	22.40 (+0.73)	19.13 (+0.67)
- (Med.)	28.47	21.71	18.80
✓(Med.)	29.67 (**+1.20**)	22.40 (+0.69)	19.41 (+0.61)
- (Ave.)	28.55	21.70	18.80
✓(Ave.)	29.48 (+0.93)	**22.44** (**+0.74**)	**19.49** (**+0.69**)

Table 4: Median/Average Car category AP_{3D} results on the KITTI *validation* set.

down proposals in Tab. 5. When we reduce the number of proposals or the range of generated proposals (Tab. 5 b. v.s. Tab. 5 c. and Tab. 5 d.), the performance will drop a lot. This is because the detection performance is dependent on the quality of anchors (recall rate), which aligns with our empirical upper-bound analysis presented in Sec. 3 of the main paper well. When we increase the number of the top-down proposals (Tab. 5 a.), the performance also drops. The reason might be that the densely generated proposals will heavily overlap with each other in 2D feature maps (as discussed in the introduction of our main paper). This will make the model confused, and lead to the difficulty of optimization during training.

	Range	Stride	#Bboxes	Easy/Mod./Hard
MonoCon [10]	-	-	-	26.33/19.01/15.98
a.	1.5	0.5	49	29.34/21.18/17.82
b.	1.5	0.75	25	**30.48/22.40/19.13**
c.	1.5	1.5	9	27.32/19.72/16.67
d.	1.0	0.5	25	28.80/20.97/17.61

Table 5: Ablation studies on the top-down proposal generation. Setting b. is used in our main experiments in the main paper.

	Cls	2D Box	3D Box	Easy/Mod./Hard
MonoCon [10]	-	-	-	26.33/19.01/15.98
a.	✓	✓	-	22.54/15.73/12.99
b.	✓	-	✓	30.09/22.08/18.92
c.	✓	✓	✓	**30.48/22.40/19.13**
d.	-	-	-	10.18/8.41/7.23

Table 6: Ablation studies on the bounding box assigner. The setting c. is used in our main experiments in the main paper.

Design of Ground Truth Assignment. We use set-prediction formulation during training. The bipartite matching consists of four costs: 1) classification cost, 2) 2D bounding box L_1 cost, 3) 2D IoU cost, 4) 3D IoU cost. We report detailed ablations in Tab. 6. Tab. 6 a. shows that the performance will drop a lot if we only use the 2D bounding boxes for assignment. This implies that the quality of the 2D box cannot ensure the prediction quality in 3D. Tab. 6 b. shows that only using the 3D box as an assignment basis will also lead to a performance drop compared with Tab. 6 c. This is because there are many cases in which predicted bottom-up proposals have no overlap with ground truth in 3D space. In these cases, the 2D box terms will serve as an auxiliary criterion during training to help the model select highly related 2D regions in the feature map to predict 3D boxes. We also try to use max IoU-based assignment criterion following Faster-RCNN [15], whose result is shown in Tab. 6 d. It shows that the performance will drop by a large margin compared with Tab. 6 c., which is because our proposed denoising process requires deleting over-generated bounding boxes in 3D. The max-IoU based assignment treats all qualified proposals as positive. Therefore the predicted score of max-IoU based models cannot be used as removing unnecessary boxes.

Method	GrooMeD-NMS NMS	 Assignment	$AP_{IoU\geq 0.7}$ Easy	 Mod.	 Hard
MonoCon (reproduced, Ave.)	-	-	25.86	18.92	16.00
+ MonoXiver (Ave.)	-	-	**29.48**	22.44	19.49
I. (Ave.)	-	✓	28.13	21.00	17.78
II. (Ave.)	✓	✓	28.29	21.25	17.85
III. (Ave.)	-	gIoU$_{3D}$	29.06	**22.56**	**19.77**

Table 7: Ablation studies by adopting Groomed-NMS's design [8].

Study of GrooMeD-NMS [8]. Groomed-NMS introduces a novel differentiable NMS approach to achieve fully end-to-end monocular 3D object detection, aiming to fulfill a similar objective to ours – eliminating redundant boxes. A key differentiator in their method involves utilizing the product of IoU2D and gIoU3D as metrics for NMS and ground truth assignment, in contrast to our employment of Hungarian assignment metrics. Our investigation into substituting our denoising module/ground truth assignment with Groomed-NMS is detailed in Table 7. Additionally, we replaced IoU3D with gIoU3D for our Hungarian assignment in the final set of experiments. The results indicate that our Hungarian assignment approach outperforms Groomed-NMS. One possible explanation is the adoption of the image-wise AP loss in their work, which might have complementary effect for their NMS design. Furthermore, our findings demonstrate a marginal performance enhancement when replacing IoU3D with gIoU3D in the ground truth assignment. We extend our gratitude to an anonymous ICCV'23 reviewer for bringing this to our attention.

Study of Different Intra-Proposal Attention in MonoXiver Structure. In the main paper, we first fuse the projection-point appearance encoding $\mathbf{f}^{pt}_{9\times N\times d}$ and the geometric encoding $\mathbf{f}^{geo}_{g\times N\times d}$, and then we decode the RoI appearance encoding $\mathbf{f}^{roi}_{1\times N\times d}$ with the 9-token geometryaware projection-point encoding in another cross-attention module. For convenience, we denote the first fusion stage as the encoder stage and the second stage as the decoder stage. We report more study results on fusion structures by changing query inputs at the encoder and decoder stages. The result is shown in Tab. 8. The result shows that using the appearance feature as decoder query input achieves better performance for easy and moderate instances. The reason might be that using queries as input will keep more RoI information in the cross-attention calculation process.

	Appearance	Geometry	Easy/Mod./Hard
a.	Decoder Q	Encoder K,V	**30.48**/**22.40**/19.13
b.	Decoder Q	Encoder Q	30.00/22.07/18.78
c.	Encoder Q	Decoder Q	29.33/21.87/19.02
d.	Encoder K,V	Decoder Q	29.42/22.03/**19.18**

Table 8: Ablation studies on different MonoXiver structures. The setting a. is used in our main experiments in the main paper.

A.3. Detection Performance on Other Categories

KITTI has limited samples of other categories (pedestrians and cyclists). Their performance is empirically unstable, which is reported in [12, 16]. Therefore, in the main paper, we mainly focus on the detection performance of car category. Here, we also discuss related empirical upper-bound analysis and experiment results in Tab. 9 and Tab. 10 for reference.

Range	Stride	*Val, AP_{R40}, Ped.*			*Val, AP_{R40}, Cyc.*		
		Easy	Moderate	Hard	Easy	Moderate	Hard
MonoCon [10]		1.46	1.31	0.99	7.60	4.35	3.55
± 1.5	0.2	33.75 $_{\uparrow}$32.29	27.06 $_{\uparrow}$25.75	23.09 $_{\uparrow}$22.10	39.09 $_{\uparrow}$31.49	21.70 $_{\uparrow}$17.35	20.20 $_{\uparrow}$16.65
± 1.5	0.3	21.61 $_{\uparrow}$20.15	17.59 $_{\uparrow}$16.28	14.46 $_{\uparrow}$13.47	34.02 $_{\uparrow}$26.42	19.00 $_{\uparrow}$14.65	17.48 $_{\uparrow}$13.93
± 1.5	0.5	6.56 $_{\uparrow}$5.10	6.07 $_{\uparrow}$4.76	4.75 $_{\uparrow}$3.76	21.31 $_{\uparrow}$13.71	12.12 $_{\uparrow}$7.77	10.92 $_{\uparrow}$7.37
± 1.5	0.75	4.85 $_{\uparrow}$3.39	3.72 $_{\uparrow}$2.41	3.11 $_{\uparrow}$2.12	12.81 $_{\uparrow}$5.21	6.99 $_{\uparrow}$2.64	6.31 $_{\uparrow}$2.76

Table 9: The empirical upper bounds of performance in Pedestrian and Cyclist on KITTI validation set based on the bottom-up anchor proposals computed by the MonoCon [10].

Empirical Upperbound Analysis. As shown in Table 1 of the main paper, the empirical performance upper bound is subject to the search range and stride. Table 9 shows the empirical upper bound for Pedestrians and Cyclists on the KITTI dataset. It shows that if we use a large stride and range the same as the setting used in the car category, the improvement potential is relatively small. If we use a small stride and large range, the potential improvement can be also very high.

Experiment Results. We report the detection performance on the Pedestrian and Cyclist category in Tab. 10. It shows that the MonoXiver is able to improve the detection performance on the Pedestrian category by a large margin. It has little improvement in the Cyclist category. The possible reason might be that the Cyclist category does not have enough data for MonoXiver to learn denoising over-generated Cyclist bounding boxes.

Range	Stride	*Val, AP_{R40}, Ped.*			*Val, AP_{R40}, Cyc.*		
		Easy	Moderate	Hard	Easy	Moderate	Hard
MonoCon [10]		1.46	1.31	0.99	7.60	4.35	3.55
± 1.5	0.5	5.59	4.57	3.64	6.48	3.45	2.99
± 1.5	0.75	7.95	5.49	4.62	8.04	4.42	3.91
± 1.5	1.5	3.57	2.79	1.90	7.60	3.87	3.35

Table 10: The detection performance on Pedestrian and Cyclist on KITTI validation set.

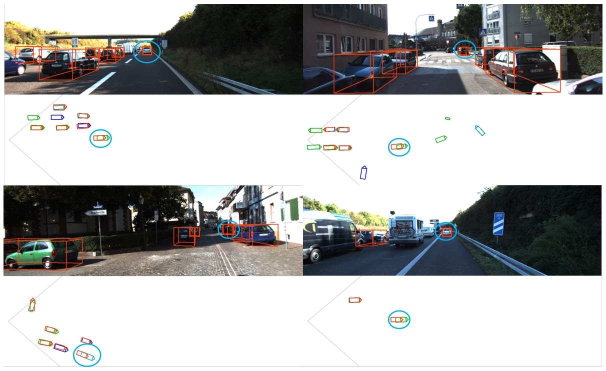

Figure 1: Qualitative results our MonoXiver with MonoCon [10] on KITTI *validation* set [3]. The ground truth is shown in green and blue. The prediction result is shown in red. We use top-1 prediction results for visualization.

B. Failure Case Study on KITTI

In Figure 1, we present an analysis of failure cases using the baseline method MonoCon [10]. The results indicate that MonoXiver faces challenges in accurately classifying top-down proposals for instances that are located far away from the camera or that are directly in front of the camera. As we have discussed in the introduction of our main paper, these instances are considered to be extremely difficult negatives due to their high overlap with the ground truth, which in turn, presents an inherent challenge of depth ambiguity in monocular 3D object detection. We believe that incorporating temporal cues in our approach could be an effective solution to address this challenge, which we intend to explore in future work.

C. Failure Case Study on Waymo

In Figure 2, we present an analysis of failure cases using the baseline method GUPNet [12]. The results indicate that

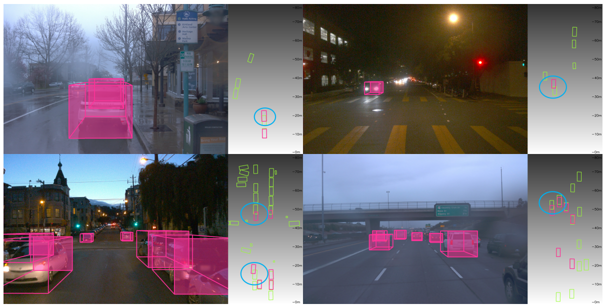

Figure 2: Qualitative results our MonoXiver with GUP-Net [12] on Waymo *validation* set [3]. The ground truth is shown in green. The prediction result is shown in pink. We use top-1 prediction results for visualization.

MonoXiver faces challenges in accurately classifying top-down proposals for instances that are highly occluded, truncated and that are located far from the camera. We believe that enhancing semantic cues (e.g. using spatial attention modules, larger/more powerful pretrained feature extraction backbone networks, etc.) will help resolve the occlusion and truncation issues.

D. Detailed Network Architecture

Embedding MLP. We use MLP to encode geometric features, projection point features, and RoI features. The structure is a stack of FC + LN [1] + ReLU blocks. We use one block to keep the structure simple. We use $C = 256$ for embedding dimensions.
Multi-head Attention layer. We use PyTorch built-in multi-head attention for implementing intra-proposal attention and inter-proposal attention. We use 8 heads for dividing the channels. We use 2 layers of MLP (with residual connection) for projecting the attended queries. We use GELU as activate function in the MLP layer.
Refinement Head. We append two blocks of stacked MLP (FC + LN + ReLU) to the encoded queries for predicting classification scores, 3D location residuals, and 3D dimension residuals separately. We use a linear layer for prediction after the two stacked MLPs.

E. Detailed Results of different runs

The detailed results of different runs is shown in Tab. 11.

References

[1] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 4

[2] Garrick Brazil, Gerard Pons-Moll, Xiaoming Liu, and Bernt Schiele. Kinematic 3d object detection in monocular video. In *ECCV*, pages 135–152. Springer, 2020. 1

Methods	$IoU_{3D} \geq 0.7$			$IoU_{3D} \geq 0.5$		
	Easy	Mod.	Hard	Easy	Mod.	Hard
MonoCon	25.99	18.98	16.13	65.36	48.33	43.63
	25.86	18.78	16.00	64.00	47.16	42.59
	25.21	18.74	15.87	64.04	47.38	42.76
	25.92	19.08	16.03	62.83	47.25	41.34
	26.33	19.01	15.98	64.53	47.35	42.49
Median	25.99	18.98	16.13	65.36	48.33	43.63
Average	25.86	18.92	16.00	64.15	47.49	42.56
MonoXiver	29.67	22.78	20.11	67.00	50.88	45.55
	30.48	22.40	19.13	65.37	47.12	41.33
	28.62	22.31	19.41	67.08	50.50	45.16
	28.61	22.07	19.00	67.08	50.40	45.00
	30.00	22.62	19.81	65.77	49.10	43.54
Median	29.67 (+3.68)	22.40 (3.42)	19.41 (+3.28)	**67.00** (+1.64)	**50.40** (+2.07)	**45.00** (+1.37)
Average	29.48 (+3.62)	**22.44** (**+3.52**)	**19.49** (**+3.49**)	66.46 (**+2.31**)	49.60 (**+2.11**)	44.12 (**+1.56**)

Table 11: Five Different Runs Car category AP_{3D} results on the KITTI *validation* set.

[3] Xiaozhi Chen, Kaustav Kundu, Ziyu Zhang, Huimin Ma, Sanja Fidler, and Raquel Urtasun. Monocular 3d object detection for autonomous driving. In *CVPR*, pages 2147–2156, 2016. 1, 3, 4

[4] 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

[5] Jiaqi Gu, Bojian Wu, Lubin Fan, Jianqiang Huang, Shen Cao, Zhiyu Xiang, and Xian-Sheng Hua. Homography loss for monocular 3d object detection. In *CVPR*, 2022. 1

[6] 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

[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

[8] Abhinav Kumar, Garrick Brazil, and Xiaoming Liu. Groomed-nms: Grouped mathematically differentiable nms for monocular 3d object detection. In *CVPR*, pages 8973–8983, 2021. 2

[9] Qing Lian, Peiliang Li, and Xiaozhi Chen. Monojsg: Joint semantic and geometric cost volume for monocular 3d object detection. In *CVPR*, 2022. 1

[10] Xianpeng Liu, Nan Xue, and Tianfu Wu. Learning auxiliary monocular contexts helps monocular 3d object detection. In *AAAI*, 2022. 1, 2, 3

[11] Zechen Liu, Zizhang Wu, and Roland Tóth. Smoke: Single-stage monocular 3d object detection via keypoint estimation. In *CVPR Workshops*, pages 996–997, 2020. 1

[12] 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:2107.13774*, 2021. 1, 3, 4

[13] 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

[14] Cody Reading, Ali Harakeh, Julia Chae, and Steven L Waslander. Categorical depth distribution network for monocular 3d object detection. In *CVPR*, pages 8555–8564, 2021. 1

[15] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 2

[16] Tai Wang, Pang Jiangmiao, and Lin Dahua. Monocular 3d object detection with depth from motion. In *ECCV*, 2022. 1, 3

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