A. Overview

This document provides additional technical details and extra analysis experiments to the main paper.

In Sec. B, we provide the information about the accuracy of mentioned depth maps while Sec. C shows the detection performance using stereo images and LiDAR point clouds. Then, Sec. D explains the correlation between 2D detector and resulting 3D detection performance. Finally, Sec. E presents detection results of *pedestrian* and *cyclist*.

B. Accuracy of Depth Maps

Tab. 10 and Tab. 11 show the accuracy of the monocular and stereo depth prediction methods listed in Tab. 7, respectively. Combined with Tab. 7, it is evident that 3D detection accuracy increases significantly when using much more accurate depth (or disparity). Note that the metrics for these two kinds of methods are different.

	Abs Rel	Sq Rel	RMSE	RMSE_{log}
MonoDepth	0.097	0.896	5.093	0.176
DORN	0.071	0.268	2.271	0.116

Table 10. Accuracy of depth prediction (monocular) on KITTI *validation* set. lower is better.

	D1-bg	D1-fg	D1-all
DispNet	4.32 %	4.41 %	4.34 %
PSMNet	1.86 %	4.62 %	2.32 %

Table 11. Accuracy of depth prediction (stereo) on KITTI *test* set. lower is better.

C. Extensions of Stereo and LiDAR

To further evaluate the proposed method, we extend it to stereo-based and LiDAR-based versions. We select some representational methods based on stereo images (or LiDAR point clouds) and report the comparative results in Table 12. The experimental results show that our method is able to give a competitive performance when using LiDAR point clouds or stereo images as input.

Note that the proposed method with LiDAR point cloud input outperforms F-PointNet [27] by **1.8** AP_{3D}, which proves that our RGB fusion module is equally effective for LiDAR-based methods.

D. 2D Detectors

Tab. 13 shows the correlation between the performance of 2D detectors and resulting 3D detection performance. We can see that improving the performance of 2D detector is an effective method to improve the overall detection accuracy. However, the huge gap between the performance of 2D detector and final 3D estimator reveals there is still a lot of room for improvement without modifying the 2D detector. The implementation details of the 2D detectors we used can be found in RRC [32] and F-PointNet [27].

Method	Data	Easy	Mod.	Hard
3DOP [4]	Stereo	6.55	5.07	4.10
Multi-Fusion [38]	Stereo	-	9.80	-
ours	Stereo	**45.85**	**26.03**	**23.16**
VoxelNet [43]	LiDAR	81.97	65.46	62.85
FPointNet [27]	LiDAR	83.26	69.28	62.56
ours	LiDAR	**84.53**	**71.07**	**63.49**

Table 12. $AP^{0.7}_{3D}(\%)$ of extended versions of proposed method and related works.

	AP_{2D}			AP_{3D}		
	Easy	Mod.	Hard	Easy	Mod.	Hard
[32]	88.4	86.7	76.6	31.1	20.0	16.8
[27]	90.5	89.9	80.7	32.2	21.1	17.3

Table 13. Comparisons of different 2D detectors. Metrics are AP_{2D} and AP_{3D} on KITTI *validation* set.

E. Pedestrian and Cyclist

Most of previous image-based 3D detection methods only focus on *Car* category as KITTI provides enough instances to train their models. Our model can also get a promising detection performance on *Pedestrian* and *Cyclist* categories because it is much easier and effective to do data augmentation for point clouds than depth maps used in previous methods. Table 14 shows their AP_{loc} and AP_{3D} on KITTI *validation* set.

Category	IoU	Task	Easy	Moderate	Hard
Pedestrian	0.25	Loc.	40.77	34.02	29.83
Pedestrian	0.25	Det.	40.17	33.45	29.28
Pedestrian	0.5	Loc.	14.30	11.26	9.23
Pedestrian	0.5	Det.	11.29	9.01	7.04
Cyclist	0.25	Loc.	28.15	17.79	16.57
Cyclist	0.25	Det.	24.80	15.66	15.11
Cyclist	0.5	Loc.	10.12	6.39	5.63
Cyclist	0.5	Det.	8.90	4.81	4.52

Table 14. Benchmarks for Pedestrian and Cyclist. 3D localization and detection AP(%) on KITTI *validation* set for *Pedestrian* and *Cyslist*. The IoU threshold is set to 0.25 and 0.5 for better comparison.