A Appendix

A.1 Broader Impacts

Our work aims to develop a robust framework to address out-of-distribution (OOD) noise scenarios in autonomous driving (AD). To the best of our knowledge, RoboFusion is the first method that leverages the generalization capabilities of visual foundation models (VFMs) like SAM [Kirillov *et al.*, 2023], FastSAM [Zhao *et al.*, 2023], and MobileSAM [Zhang *et al.*, 2023] for multi-modal 3D object detection. Although existing multi-model 3D object detection methods achieve the state-of-the-art (SOTA) performance of 'clean' datasets, they overlook the robustness of real-world scenarios[Song *et al.*, 2024]. Therefore, we believe it is valuable to combine VFMs and multi-modal 3D object detection to mitigate the impact of OOD noise scenarios.

A.2 More Results

Roles of Different Modules in RoboFusion.

To assess the roles of different modules in RoboFusion, we conduct an ablation study on the original SAM rather than SAM-AD, as shown in Table 1, where a) is the results of the baseline [Chen *et al.*, 2022], b)-e) shows the performance of our RoboFusion-L under different modules. According to Table 1, SAM and AD-FPN modules significantly improve the performance in OOD noisy scenarios. It is worth noticing that DGWA module significantly improves the performance, especially in snow noisy scenarios. By Table 1, the impact of fog noise on point clouds is relatively minor. But, using A.F. (Adaptive Fusion) module to dynamically aggregate point cloud features and image features exhibits significant enhancements in fog-noise scenarios.

Table 1: Roles of RoboFusion modules on **KITTI-C validation** set for car class with AP of R_{40} at moderate difficulty. 'A.F.' denotes **Adaptive Fusion** module. 'S.L.' denotes Strong Sunlight.

Method	SAM	AD-FPN	DGWA	A.F.	Snow	Rain	Fog	S.L.
a)					34.77	41.30	44.55	80.97
b)	✓				57.43	54.27	68.81	82.07
c)	✓	✓			59.81	56.59	69.68	83.20
d)	✓	✓	✓		66.45	58.11	70.53	84.01
e)	✓	✓	✓	✓	68.47	59.07	74.38	84.07

More Results on the KITTI-C validation set.

Besides the experimental results mentioned in the main text, we test our RoboFusion on KITTI-C and nuScenes-C [Dong *et al.*, 2023] to extend our work to a wider range of noise scenarios, including Gaussian, Uniform, Impulse, Moving Object, Motion Blur, Local Density, Local Cutout, Local Gaussian, Local Uniform, and Local Impulse, as shown in Tables 2, 3, and 4. From these Tables, compared with LiDAR-only methods including SECOND [Yan *et al.*, 2018], PointPillars [Lang *et al.*, 2019], PointRCNN [Shi *et al.*, 2019] and PV-RCNN [Shi *et al.*, 2020], Camera-Only methods including Smoke [Liu *et al.*, 2020], ImVoxelNet [Rukhovich *et al.*, 2022], and multi-modal methods including EPNet [Huang *et al.*, 2020], Focals Conv [Chen *et al.*, 2022], and LoGoNet [Li *et al.*, 2023], our RoboFusion-L, RoboFusion-B, and RoboFusion-T consistently outperform across various noise scenarios and achieve the best overall performance. Overall, our RoboFusion demonstrates superior performance in weather-noisy (*i.e.* Snow, Rain, Fog, and Strong Sunlight) scenarios and exhibits better results across a broader range of scenarios, which shows remarkable robustness and generalizability.

Performance Comparison Analysis with the LoGoNet.

In addition, to provide a clearer analysis of performance across different noise scenarios, we present a more detailed comparative study of our RoboFusion-L and LoGoNet [Li *et al.*, 2023] on the KITTI-C validation dataset, as shown in Table 5. It is worth noting that LoGoNet is a SOTA multi-modal 3D detector known for its exceptional robustness and high accuracy. [Dong *et al.*, 2023] provides noise at varying levels, with the KITTI-C dataset including 5 severities. It is evident that our method demonstrates a high degree of robustness, exhibiting the most stable results with the variance of noise severities. For instance, when considering snow conditions, the performance of our RoboFusion-L shows a marginal variation from 86.69% to 83.67% across severities from 1 to 5. In contrast, LoGoNet's performance drops from 55.07% to 45.02% over the same severity range. Furthermore, in the presence of moving object noise, our method outperforms LoGoNet. In summary, our RoboFusion exhibits remarkable robustness and generalization capabilities, making it well-suited to diverse noise scenarios.

More Results on the nuScenes-C validation set.

As depicted in Table 6, compared with LiDAR-only methods including PointPillars [Lang *et al.*, 2019], and CenterPoint [Yin *et al.*, 2021], Camera-Only methods FCOS3D [Wang *et al.*, 2021], DETR3D [Wang *et al.*, 2022], and BEVFormer [Li *et al.*, 2022] and multi-modal methods including FUTR3D [Chen *et al.*, 2023], TransFusion [Bai *et al.*, 2022], BEV-Fusion [Liu *et al.*, 2023] and DeepInteraction [Yang *et al.*, 2022], our RoboFusion demonstrates superior performance across more noise scenarios in AD on average. For instance, our RoboFusion-L excels in 10 noise scenarios, including Weather (Snow, Rain, Fog, Strong Sunlight), Sensor (Density, Cutout, Crosstalk), Motion (Compensation, Motion Blur), and Object (Local Cutout), outperforming DeepInteraction [Yang *et al.*, 2022] which achieves the best performance only in 5 of these noise scenarios. Overall, our method exhibits not only exceptional robustness in weather-induced noise scenarios, but also shows remarkable resilience across a broader noise include sensor, motion and object noise.

A.3 Visualization

As shown in Fig. 1, we provide visualization results between our RoboFusion-L and LoGoNet on the KITTI-C dataset. Overall, compared to SOTA methods like LoGoNet [Li *et al.*, 2023], our method enhances the robustness of multi-modal 3D object detection by leveraging the generalization capability and robustness of VFMs to mitigate OOD noisy scenarios in AD.

Table 2: Comparison with SOTA methods on **KITTI-C validation** set. The results are evaluated based on the **car** class with AP of R_{40} at **moderate** difficulty. The best one is highlighted in **bold**.'S.L.' denote Strong Sunlight. 'RCE' denotes Relative Corruption Error from Ref.[Dong *et al.*, 2023].

Corruptions		LiDAR-Only				Camera-Only		LC Fusion					
		SECOND †	PointPillars †	PointRCNN †	PV-RCNN †	SMOKE †	ImVoxelNet †	EPNet †	Focals Conv †	LoGoNet *	RoboFusion **(Ours)** L	B	T
None(AP_{clean})		81.59	78.41	80.57	84.39	7.09	11.49	82.72	85.88	85.04	**88.04**	87.87	87.60
Weather	Snow	52.34	36.47	50.36	52.35	2.47	0.22	34.58	34.77	51.45	**85.29**	84.70	84.60
	Rain	52.55	36.18	51.27	51.58	3.94	1.24	36.27	41.30	55.80	**86.48**	85.54	84.79
	Fog	74.10	64.28	72.14	79.47	5.63	1.34	44.35	44.55	67.53	**85.53**	84.00	84.17
	S.L.	78.32	62.28	62.78	79.91	6.00	10.08	69.65	80.97	75.54	**85.50**	85.15	84.75
Sensor	Density	80.18	76.49	80.35	82.79	-	-	82.09	84.95	83.68	**85.71**	84.34	84.11
	Cutout	73.59	70.28	73.94	76.09	-	-	76.10	78.06	77.17	**83.17**	81.30	81.21
	Crosstalk	80.24	70.85	71.53	82.34	-	-	82.10	**85.82**	82.00	84.12	82.45	83.07
	Gaussian (L)	64.90	74.68	61.20	65.11	-	-	60.88	**82.14**	61.85	76.56	78.32	76.52
	Uniform (L)	79.18	77.31	76.39	81.16	-	-	79.24	**85.81**	82.94	85.05	83.04	84.11
	Impulse (L)	81.43	78.17	79.78	82.81	-	-	81.63	85.01	84.66	85.26	85.06	**85.46**
	Gaussian (C)	-	-	-	-	1.56	2.43	80.64	80.97	84.29	82.16	**84.63**	82.17
	Uniform (C)	-	-	-	-	2.67	4.85	81.61	83.38	84.45	83.30	**85.20**	83.30
	Impulse (C)	-	-	-	-	1.83	2.13	81.18	80.83	84.20	83.51	**84.55**	82.91
Motion	Moving Obj.	52.69	50.15	50.54	54.60	1.67	5.93	**55.78**	49.14	14.44	49.30	49.12	49.90
	Motion Blur	-	-	-	-	3.51	4.19	74.71	81.08	84.52	84.17	**84.56**	84.18
Object	Local Density	75.10	69.56	74.24	77.63	-	-	76.73	80.84	78.63	83.21	82.53	**83.22**
	Local Cutout	68.29	61.80	67.94	72.29	-	-	69.92	76.64	64.88	**77.22**	75.27	76.23
	Local Gaussian	72.31	76.58	69.82	70.44	-	-	75.76	**82.02**	55.66	79.02	78.32	78.33
	Local Uniform	80.17	78.04	77.67	82.09	-	-	81.71	84.69	79.94	**84.69**	83.70	84.37
	Local Impulse	81.56	78.43	80.26	84.03	-	-	82.21	**85.78**	84.29	85.26	85.08	85.06
Average(AP_{cor})		71.68	66.34	68.76	73.41	3.25	3.60	70.35	74.43	71.89	**81.72**	81.31	81.12
RCE (%) ↓		12.14	15.38	14.65	13.00	54.11	68.65	14.94	13.32	15.46	**7.17**	7.46	7.38

†: Results from Ref. [Dong *et al.*, 2023].
* denotes re-implement result.

Table 3: Comparison with SOTA methods on **KITTI-C validation** set. The results are evaluated based on the **car** class with AP of R_{40} at **easy** difficulty. The best one is highlighted in **bold**. 'S.L.' denotes Strong Sunlight. 'RCE' denotes Relative Corruption Error from Ref.[Dong *et al.*, 2023].

Corruptions		Lidar-Only				Camera-Only		LC Fusion					
		SECOND †	PointPillars †	PointRCNN †	PV-RCNN †	SMOKE †	ImVoxelNet †	EPNet †	Focals Conv †	LoGoNet *	RoboFusion **(Ours)** L	B	T
None(AP_{clean})		90.53	87.75	91.65	92.10	10.42	17.85	92.29	92.00	92.04	**93.30**	93.22	93.28
Weather	Snow	73.05	55.99	71.93	73.06	3.68	0.30	48.03	53.80	74.24	**88.77**	88.18	88.31
	Rain	73.31	55.17	70.79	72.37	5.66	1.77	50.93	61.44	75.96	88.12	**88.57**	87.75
	Fog	85.58	74.27	85.01	**89.21**	8.06	2.37	64.83	68.03	86.60	88.96	88.16	88.09
	S.L.	88.05	67.42	64.90	87.27	8.75	15.72	81.77	90.03	80.30	89.79	89.23	**90.36**
Sensor	Density	90.45	86.86	91.33	91.98	-	-	91.89	91.14	91.85	**92.90**	92.08	92.12
	Cutout	81.75	78.90	83.33	83.40	-	-	84.17	83.84	84.20	**85.94**	85.75	84.75
	Crosstalk	89.63	78.51	77.38	90.52	-	-	91.30	92.01	88.15	91.71	91.54	**92.07**
	Gaussian (L)	73.21	86.24	74.28	74.61	-	-	66.99	**88.56**	64.62	80.96	84.30	83.23
	Uniform (L)	89.50	87.49	89.48	90.65	-	-	89.70	91.77	90.75	**92.89**	91.28	91.63
	Impulse (L)	90.70	87.75	90.80	91.91	-	-	91.44	92.10	91.66	91.90	91.95	**92.30**
	Gaussian (C)	-	-	-	-	2.09	3.74	91.62	89.51	91.64	91.94	**92.08**	91.57
	Uniform (C)	-	-	-	-	3.81	7.66	91.95	91.20	91.84	92.01	92.14	**92.93**
	Impulse (C)	-	-	-	-	2.57	3.35	91.68	89.90	91.65	91.96	**92.04**	91.33
Motion	Moving Obj.	62.64	58.49	59.29	63.36	2.69	9.63	**66.32**	54.57	16.83	53.09	51.94	51.70
	Motion Blur	-	-	-	-	5.39	6.75	89.65	91.56	91.96	91.99	**92.09**	92.06
Object	Local Density	87.74	82.90	88.37	89.60	-	-	89.40	89.60	89.00	92.02	**92.42**	92.42
	Local Cutout	81.29	75.22	83.30	84.38	-	-	82.40	85.55	77.57	87.30	87.49	**87.79**
	Local Gaussian	82.05	87.69	82.44	77.89	-	-	85.72	**89.78**	60.03	89.56	89.41	89.62
	Local Uniform	90.11	87.83	89.30	90.63	-	-	91.32	91.88	88.51	91.59	91.53	**91.75**
	Local Impulse	90.58	87.84	90.60	91.91	-	-	91.67	92.02	91.34	**92.09**	91.97	90.69
Average(AP_{cor})		83.10	77.41	80.78	83.92	4.74	5.69	81.63	83.91	80.93	**88.27**	88.20	88.12
RCE(%)↓		8.20	11.78	11.85	8.87	54.46	68.07	11.54	8.78	12.07	**5.39**	**5.39**	5.53

†: Results from Ref. [Dong *et al.*, 2023].
* denotes re-implement result.

Table 4: Comparison with SOTA methods on **KITTI-C validation** set. The results are evaluated based on the **car** class with AP of R_{40} at **hard** difficulty. The best one is hightlighted in **bold**. 'S.L.' denotes Strong Sunlight. 'RCE' denotes Relative Corruption Error from Ref.[Dong *et al.*, 2023].

Corruptions		**Lidar-Only**				**Camera-Only**		**LC Fusion**					
		SECOND †	PointPillars †	PointRCNN †	PV-RCNN †	SMOKE †	ImVoxelNet †	EPNet †	Focals Conv †	LoGoNet *	RoboFusion **(Ours)** L	B	T
None(AP_{clean})		78.57	75.19	78.06	82.49	5.57	9.20	80.16	83.36	84.31	**85.27**	84.27	83.36
Weather	Snow	48.62	32.96	45.41	48.62	1.92	0.20	32.39	30.41	45.57	**64.26**	62.49	62.74
	Rain	48.79	32.65	45.78	48.20	3.16	0.99	34.69	35.71	50.12	**66.07**	64.89	63.18
	Fog	68.93	58.19	68.05	75.05	4.56	1.03	38.12	39.50	60.47	**80.03**	78.37	77.29
	S.L.	74.62	58.69	61.11	78.02	4.91	8.24	66.43	78.06	73.62	80.02	77.52	**81.61**
Sensor	Density	77.04	72.85	77.58	81.15	-	-	79.77	82.38	81.98	**83.06**	83.03	83.05
	Cutout	70.79	67.32	71.57	74.60	-	-	73.95	76.69	76.18	76.96	77.00	**77.38**
	Crosstalk	76.92	67.51	69.41	80.98	-	-	79.54	83.22	80.36	82.94	**83.22**	83.08
	Gaussian (L)	61.09	71.12	56.73	62.70	-	-	56.88	**77.15**	59.98	74.45	75.03	73.81
	Uniform (L)	75.61	74.09	72.25	78.93	-	-	75.92	81.62	80.68	81.74	81.79	**82.44**
	Impulse (L)	78.33	74.65	76.88	81.79	-	-	79.14	**83.28**	82.51	83.13	83.16	83.24
	Gaussian (C)	-	-	-	-	1.18	1.96	78.20	79.01	82.22	82.86	**83.05**	81.32
	Uniform (C)	-	-	-	-	2.19	3.90	79.14	81.39	82.37	**83.22**	83.03	82.06
	Impulse (C)	-	-	-	-	1.52	1.71	78.51	78.87	82.16	82.75	**83.00**	81.59
Motion	Moving Obj.	48.02	45.47	46.23	50.75	1.40	4.63	**50.97**	45.34	13.66	43.56	42.62	42.89
	Motion Blur	-	-	-	-	2.95	3.32	72.49	77.75	82.50	**83.12**	83.06	82.92
Object	Local Density	71.45	65.70	71.09	75.39	-	-	74.36	77.30	76.83	**81.71**	81.24	81.15
	Local Cutout	63.25	56.69	63.50	68.58	-	-	66.53	72.40	60.62	71.95	72.07	**73.78**
	Local Gaussian	68.16	73.11	65.65	68.03	-	-	72.71	**78.52**	54.02	76.38	76.41	76.26
	Local Uniform	76.67	74.68	74.37	80.17	-	-	78.85	81.99	77.44	82.04	82.06	**82.33**
	Local Impulse	78.47	75.18	77.38	82.33	-	-	79.79	**83.20**	82.21	82.99	83.16	82.99
Average(AP_{cor})		67.92	62.55	65.18	70.95	2.64	2.88	67.41	71.18	69.27	**77.16**	76.81	76.75
RCE(%)↓		13.55	16.80	16.49	13.98	52.54	68.62	15.89	14.59	17.83	9.51	9.71	**7.93**

†: Results from Ref. [Dong *et al.*, 2023].
* denotes re-implement result.

Table 5: Performance comparison of our RoboFusion-L with LoGoNet on KITTI-C with 5 noise severities. The results are reported based on the **car** with AP of R_{40} at **moderate** difficulty. 'S.L.' denotes Strong Sunlight. The better one is marked in **bold**.

Corruptions		Severity 1	2	3	4	5	AP_s
Weather	Snow	55.07 / **86.69**	52.98 / **86.55**	53.08 / **85.94**	51.14 / **83.61**	45.02 / **83.67**	51.45 / **85.29**
	Rain	57.29 / **87.84**	56.90 / **87.75**	56.76 / **86.49**	55.05 / **85.24**	53.01 / **85.07**	55.80 / **86.48**
	Fog	75.93 / **87.31**	69.69 / **86.58**	64.77 / **84.71**	64.69 / **84.56**	62.58 / **84.51**	67.53 / **85.53**
	S.L.	82.03 / **87.26**	80.53 / **86.53**	76.75 / **84.66**	71.12 / **84.61**	67.31 / **84.46**	75.54 / **85.50**
Sensor	Density	86.60 / **86.81**	84.59 / **86.59**	84.05 / **85.60**	82.74 / **85.27**	82.42 / **84.30**	83.68 / **85.71**
	Cutout	82.18 / **87.64**	80.02 / **86.21**	77.41 / **83.25**	74.66 / **80.81**	71.59 / **77.94**	77.17 / **83.17**
	Crosstalk	84.22 / **84.41**	83.38 / **84.38**	81.41 / **84.13**	80.78 / **83.79**	80.22 / **83.90**	82.00 / **84.12**
	Gaussian (L)	84.69 / **85.41**	82.52 / **84.66**	77.43 / **81.39**	47.28 / **73.58**	17.31 / **57.79**	61.85 / **76.56**
	Uniform (L)	84.77 / **85.77**	84.64 / **85.42**	84.39 / **85.47**	82.32 / **85.00**	78.59 / **83.59**	82.94 / **85.05**
	Impulse (L)	84.45 / **84.95**	**84.73** / 82.88	**84.92** / 82.20	**84.63** / 80.51	**84.56** / 80.29	**84.66** / 82.16
	Gaussian (C)	84.53 / **85.77**	84.47 / **85.42**	84.31 / **85.47**	84.18 / **85.32**	83.96 / **84.32**	84.29 / **85.26**
	Uniform (C)	84.74 / **85.57**	84.57 / **85.08**	**84.54** / 82.96	**84.36** / 82.53	**84.05** / 80.36	**84.45** / 83.30
	Impulse (C)	84.53 / **85.70**	**84.26** / 83.63	**84.38** / 83.54	**83.95** / 82.42	**83.86** / 82.28	**84.20** / 83.51
Motion	Moving Obj.	58.89 / **78.46**	12.78 / **67.86**	0.43 / **41.07**	0.06 / **36.28**	0.07 / **22.85**	14.44 / **49.30**
	Motion Blur	84.64 / **85.23**	84.53 / **84.98**	84.56 / **84.72**	**84.45** / 83.00	**84.43** / 82.96	**84.52** / 84.17
Object	Local Density	82.31 / **85.23**	81.66 / **84.87**	80.15 / **82.70**	76.53 / **82.08**	72.52 / **81.21**	78.63 / **83.21**
	Local Cutout	76.77 / **82.94**	72.46 / **81.31**	65.87 / **78.14**	59.14 / **74.12**	50.17 / **69.61**	64.88 / **77.22**
	Local Gaussian	84.45 / **86.81**	81.12 / **86.25**	67.13 / **82.72**	33.33 / **76.01**	12.27 / **63.31**	55.66 / **79.02**
	Local Uniform	84.51 / **85.91**	84.35 / **85.65**	81.95 / **85.23**	79.62 / **84.66**	69.25 / **81.99**	79.94 / **84.68**
	Local Impulse	84.53 / **85.65**	84.47 / **85.13**	84.32 / **85.18**	84.40 / **85.16**	83.72 / **85.16**	84.29 / **85.25**
AP_c		79.35 / **85.56**	75.73 / **84.38**	72.93 / **81.77**	68.22 / **79.92**	63.34 / **76.97**	71.81 / **81.72**
Clean							85.04 / **88.04**

Table 6: Comparison with SOTA methods on **nuScenes-C validation** set with **mAP**. 'D.I.' refers to DeepInteraction [Yang *et al.*, 2022]. The best one is highlighted in **bold**. 'S.L.' denotes Strong Sunlight. 'RCE' denotes Relative Corruption Error from Ref.[Dong *et al.*, 2023].

Corruptions		**Lidar-Only**		**Camera-Only**			**LC Fusion**						
		PointPillars†	CenterPoint†	FCOS3D†	DETR3D†	BEVFormer†	FUTR3D†	TransFusion†	BEVFusion†	D.I.*	RoboFusion **(Ours)** L	B	T
None(AP_{clean})		27.69	59.28	23.86	34.71	41.65	64.17	66.38	68.45	69.90	**69.91**	69.40	69.09
Weather	Snow	27.57	55.90	2.01	5.08	5.73	52.73	63.30	62.84	62.36	**67.12**	66.07	65.96
	Rain	27.71	56.08	13.00	20.39	24.97	58.40	65.35	66.13	66.48	**67.58**	67.01	66.45
	Fog	24.49	43.78	13.53	27.89	32.76	53.19	53.67	54.10	54.79	**67.01**	65.54	64.34
	S.L.	23.71	54.20	17.20	34.66	41.68	57.70	55.14	64.42	64.93	**67.24**	66.71	66.54
Sensor	Density	27.27	58.60	-	-	-	63.72	65.77	67.79	68.15	**69.48**	69.02	68.58
	Cutout	24.14	56.28	-	-	-	62.25	63.66	66.18	66.23	**69.18**	69.01	68.20
	Crosstalk	25.92	56.64	-	-	-	62.66	64.67	67.32	68.12	**68.68**	68.04	68.17
	FOV lost	8.87	20.84	-	-	-	26.32	24.63	27.17	**42.66**	39.48	39.30	39.43
	Gaussian (L)	19.41	45.79	-	-	-	58.94	55.10	**60.64**	57.46	57.77	57.07	56.00
	Uniform (L)	25.60	56.12	-	-	-	63.21	64.72	66.81	**67.42**	64.57	64.25	64.99
	Impulse (L)	26.44	57.67	-	-	-	63.43	65.51	**67.54**	67.41	65.64	65.45	65.44
	Gaussian (C)	-	-	3.96	14.86	15.04	54.96	64.52	64.44	66.52	66.73	**66.75**	66.53
	Uniform (C)	-	-	8.12	21.49	23.00	57.61	65.26	65.81	**65.90**	65.77	65.76	65.56
	Impulse (C)	-	-	3.55	14.32	13.99	55.16	64.37	64.30	**65.65**	64.82	64.75	64.56
Motion	Compensation	3.85	11.02	-	-	-	31.87	9.01	27.57	39.95	**41.88**	39.54	41.28
	Motion Blur	-	-	10.19	11.06	19.79	55.99	64.39	64.74	65.45	**67.21**	66.52	66.42
Obeject	Local Density	26.70	57.55	-	-	-	63.60	65.65	67.42	**67.71**	66.74	66.59	65.88
	Local Cutout	17.97	48.36	-	-	-	61.85	63.33	63.41	65.19	**66.82**	66.53	66.76
	Local Gaussian	25.93	51.13	-	-	-	62.94	63.76	64.34	64.75	65.08	**65.17**	64.77
	Local Uniform	27.69	57.87	-	-	-	64.09	66.20	**67.58**	66.44	66.71	66.19	65.40
	Local Impulse	27.67	58.49	-	-	-	64.02	66.29	**67.91**	67.86	66.53	66.87	66.67
Average(AP_{cor})		22.99	49.78	8.94	18.71	22.12	56.88	58.77	61.35	62.92	**63.90**	63.43	63.23
RCE (%) ↓		16.95	16.01	62.51	46.07	46.89	11.34	11.45	10.36	9.97	8.58	8.59	**8.47**

†: Results from Ref. [Dong *et al.*, 2023].
* denotes re-implement result.

A.4 More Limitations

Although we have mentioned the two main limitations in the 'Conclusions' section of the main text, our RoboFusion still has other limitations. Our method does not achieve the best performance in all noisy scenarios. For instance, as shown in Table 2, our method does not show the best in 'Moving Object' noisy scenarios. Furthermore, we conduct experiments only on the corruption datasets [Dong *et al.*, 2023] rather than real-world datasets. It is valuable to construct a real-world corruption dataset, but it must be an expensive work.

References

[Bai *et al.*, 2022] Xuyang Bai, Zeyu Hu, Xinge Zhu, Qingqiu Huang, Yilun Chen, Hongbo Fu, and Chiew-Lan Tai. Transfusion: Robust lidar-camera fusion for 3d object detection with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1090–1099, 2022.

[Chen *et al.*, 2022] Yukang Chen, Yanwei Li, Xiangyu Zhang, Jian Sun, and Jiaya Jia. Focal sparse convolutional networks for 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5428–5437, 2022.

[Chen *et al.*, 2023] Xuanyao Chen, Tianyuan Zhang, Yue Wang, Yilun Wang, and Hang Zhao. Futr3d: A unified sensor fusion framework for 3d detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 172–181, 2023.

[Dong *et al.*, 2023] Yinpeng Dong, Caixin Kang, Jinlai Zhang, Zijian Zhu, et al. Benchmarking robustness of 3d object detection to common corruptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1022–1032, 2023.

[Huang *et al.*, 2020] Tengteng Huang, Zhe Liu, Xiwu Chen, and Xiang Bai. Epnet: Enhancing point features with image semantics for 3d object detection. In *European Conference on Computer Vision*, pages 35–52. Springer, 2020.

[Kirillov *et al.*, 2023] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023.

[Lang *et al.*, 2019] Alex H Lang, Sourabh Vora, et al. Pointpillars: Fast encoders for object detection from point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12697–12705, 2019.

[Li *et al.*, 2022] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, et al. Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In *European conference on computer vision*, pages 1–18. Springer, 2022.

[Li *et al.*, 2023] Xin Li, Tao Ma, Yuenan Hou, Botian Shi, et al. Logonet: Towards accurate 3d object detection with local-to-global cross-modal fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17524–17534, 2023.

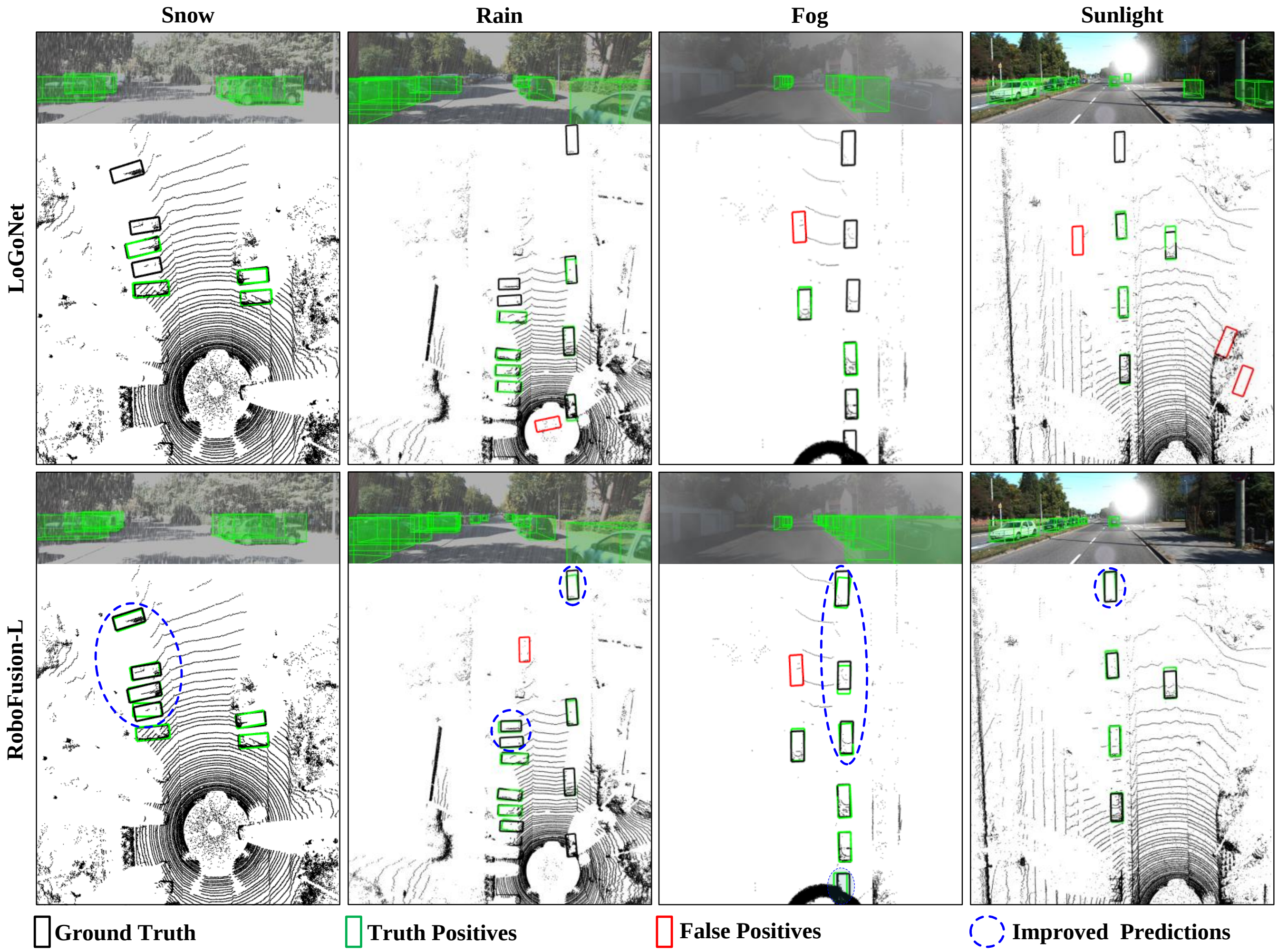

Figure 1: **Visualization Results of LoGoNet and our RoboFusion in KITTI-C dataset.** We use boxes in red to represent false positives, green boxes for truth positives, and black for the ground truth. We use blue dashed ovals to highlight the pronounced improvements in predictions.

[Liu *et al.*, 2020] Zechen Liu, Zizhang Wu, and Roland Tóth. Smoke: Single-stage monocular 3d object detection via keypoint estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 996–997, 2020.

[Liu *et al.*, 2023] Zhijian Liu, Haotian Tang, Alexander Amini, Xinyu Yang, Huizi Mao, Daniela L Rus, and Song Han. Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. In *2023 IEEE international conference on robotics and automation (ICRA)*, pages 2774–2781. IEEE, 2023.

[Rukhovich *et al.*, 2022] Danila Rukhovich, Anna Vorontsova, and Anton Konushin. Imvoxelnet: Image to voxels projection for monocular and multi-view general-purpose 3d object detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2397–2406, 2022.

[Shi *et al.*, 2019] Shaoshuai Shi, Xiaogang Wang, and Hongsheng Li. Pointrcnn: 3d object proposal generation and detection from point cloud. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 770–779, 2019.

[Shi *et al.*, 2020] Shaoshuai Shi, Chaoxu Guo, Li Jiang, Zhe Wang, Jianping Shi, Xiaogang Wang, and Hongsheng Li. Pv-rcnn: Point-voxel feature set abstraction for 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10529–10538, 2020.

[Song *et al.*, 2024] Ziying Song, Lin Liu, Feiyang Jia, et al. Robustness-aware 3d object detection in autonomous driving: A review and outlook. *arXiv preprint arXiv:2401.06542*, 2024.

[Wang *et al.*, 2021] Tai Wang, Xinge Zhu, Jiangmiao Pang, and Dahua Lin. Fcos3d: Fully convolutional one-stage monocular 3d object detection. In *Proceedings of the*

IEEE/CVF International Conference on Computer Vision, pages 913–922, 2021.

[Wang *et al.*, 2022] Yue Wang, Vitor Campagnolo Guizilini, Tianyuan Zhang, et al. Detr3d: 3d object detection from multi-view images via 3d-to-2d queries. In *Conference on Robot Learning*, pages 180–191. PMLR, 2022.

[Yan *et al.*, 2018] Yan Yan, Yuxing Mao, and Bo Li. Second: Sparsely embedded convolutional detection. *Sensors*, 18(10):3337, 2018.

[Yang *et al.*, 2022] Zeyu Yang, Jiaqi Chen, Zhenwei Miao, et al. Deepinteraction: 3d object detection via modality interaction. *Advances in Neural Information Processing Systems*, 35:1992–2005, 2022.

[Yin *et al.*, 2021] Tianwei Yin, Xingyi Zhou, and Philipp Krahenbuhl. Center-based 3d object detection and tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11784–11793, 2021.

[Zhang *et al.*, 2023] Chaoning Zhang, Dongshen Han, Yu Qiao, Jung Uk Kim, et al. Faster segment anything: Towards lightweight sam for mobile applications. *arXiv preprint arXiv:2306.14289*, 2023.

[Zhao *et al.*, 2023] Xu Zhao, Wenchao Ding, Yongqi An, Yinglong Du, et al. Fast segment anything, 2023.