FogFusion: robust camera-LiDAR fusion for fog-resilient 3D perception in pedestrian-vehicle interaction scenarios

Yifan Wang[1,2], Bingli Zhang[1,2], Yixin Wang[1,2,*], Chengbiao Zhang[1,2], Xingyu Wang[1,2], Junzhao Jiang[1,2], Xiang Luo[1,2], Shuqing Zhao[1,2]

***Abstract*— Robust pedestrian perception in highly interactive traffic environments is a critical challenge for Autonomous Vehicles (AVs). This challenge is drastically exacerbated in out-of-distribution (OOD) scenarios like foggy weather, where severe sensor attenuation disrupts upstream 3D perception, depriving downstream prediction pipelines of reliable spatial states. To restore reliable spatial observations of pedestrians for safe Pedestrian-Vehicle Interaction (PVI) in such conditions, we introduce FogFusion, a synergistic camera-LiDAR fusion network designed to recover high-fidelity 3D representations through fog. By integrating Depth Completion with Fog Convolution (DCFC) and Flexible Cylindrical Voxel (FCV) encoding, FogFusion effectively mitigates back-scattering noise and reconstructs sparse geometric features. Evaluated on the KITTI and KITTI-C benchmarks, our framework improves overall 3D detection robustness by at least 3.32% under foggy conditions. By providing stable 3D spatial states of traffic agents as a necessary precondition, FogFusion bridges the perception-prediction gap and lays a resilient foundation for downstream pedestrian intent forecasting and safe vehicle-pedestrian interaction.**

I. Introduction

The safety of Pedestrian-Vehicle Interaction (PVI) in urban environments depends on an AV's ability to accurately perceive the spatial states of surrounding traffic participants, including pedestrians. downstream behavior prediction suffers from "tracklet fragmentation,"

In adverse weather such as dense fog, this assumption breaks down. Fog severely attenuates LiDAR signals, introduces back-scattering, and degrades camera visibility. As upstream perception can no longer reliably capture the spatial structure of traffic participants, downstream behavior prediction suffers from "tracklet fragmentation," making reliable intent inference and interaction modeling difficult. Therefore, reconstructing robust 3D representations from degraded multi-modal signals is essential for resilient PBP. To this end, we propose FogFusion, an adaptive camera-LiDAR fusion paradigm that restores the structural integrity of dynamic obstacles in fog and enables safer prediction in OOD environments.

II. Method

Overall Structure

To supply downstream predictors with reliable spatial sequences, FogFusion addresses the missing data problem at the sensory and feature levels, as shown in **Fig. 1**. The framework focuses on mitigating physical sensor degradation through the following three core components:

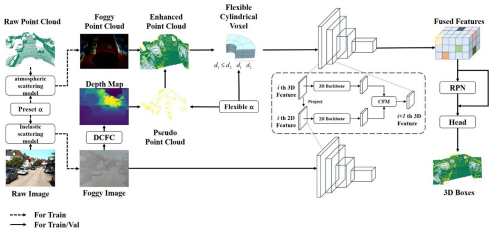

Fig. 1. Structure of FogFusion network

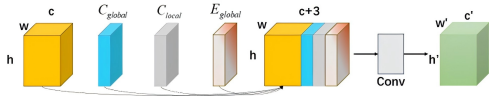

Fig. 2. Fog convolution layer

Depth Completion with Fog Convolution (DCFC): As shown in **Fig. 2**, to counteract LiDAR signal loss, we introduce the DCFC module. It leverages dense visual semantics to guide the depth completion of sparse point clouds. Crucially, the specialized fog convolution dynamically filters out back-scattering noise, generating a dense virtual point cloud that recovers the structural boundaries of occluded agents.

Flexible Cylindrical Voxel (FCV) Encoding: To handle the varying density of the generated virtual points without losing fine-grained geometric details, we propose the FCV encoding method. It provides a highly adaptive spatial discretization, which is essential for capturing the precise pose and orientation of dynamic entities.

Cylindrical Fusion Module (CFM): As shown in **Fig. 3**, to achieve a unified representation, the CFM thoroughly integrates the FCV-encoded geometric features with 2D camera context. This ensures that the generated 3D bounding boxes are informed by both robust structure and rich visual cues.

III. Experimental Validation

A. *Experimental Setup*

We evaluated our framework on the KITTI and KITTI-C benchmarks using the Car category as a representative proxy for traffic agent detection, stress-testing perceptual resilience under clear and foggy conditions. To better reflect real-world signal degradation, we adopted Hahner's attenuation models

[1]School of Automotive and Transportation Engineering, Hefei University of Technology, Hefei, 230009, China.
[2]Anhui Engineering Laboratory of Intelligent Automobile, Hefei, 230009, China.
*Correspondence: 2021110938@mail.hfut.edu.cn.

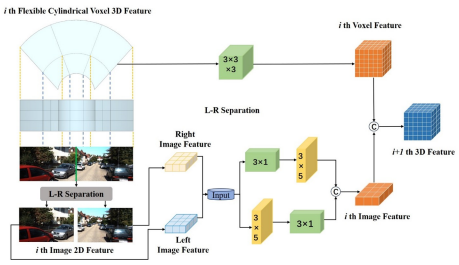

Fig. 3. Cylindrical Fusion Module

for LiDAR and opacity masks for cameras, avoiding the limitations of simplistic simulators such as Robo3D. The network is trained in PyTorch on an RTX A6000 GPU for 50 epochs. Finally, to verify FogFusion's ability to maintain continuous spatial tracklets of dynamic traffic participants—a key prerequisite for intent forecasting—we compared it with state-of-the-art single-modal and multi-modal fusion architectures.

B. *Comparison with state-of-the-art methods*

To evaluate FogFusion on foggy datasets, the relative corruption error (RCE) is used to measure the percentage of performance drop, where AP_{clean} and AP_{fog} denote the AP values on the clean-weather and fog simulation datasets, respectively:

$$RCE = \frac{AP_{clean} - AP_{fog}}{AP_{clean}}$$

As shown in Table I, FogFusion achieves the highest AP_{clean} (88.94%) and AP_{fog} (87.74%) among all listed methods. Compared to the strongest multi-modal baseline LoGoNet, FogFusion reduces RCE by 3.32% (from 1.81 to 1.35), while its overall robustness ranks second only to VFF which sacrifices 3.44% clean-weather AP.

TABLE I
COMPARISON WITH SOTA METHODS ON KITTI-C DATASET

Modality	Network	Year	AP_{clean} (%)	AP_{fog} (%)	RCE (%)
LiDAR	PV-RCNN[1]	2020	84.39	79.47	5.83
	TED[2]	2024	88.92	87.65	1.43
Camera	PGD[3]	2022	8.10	0.87	89.26
	ImVoxelNet[4]	2022	11.49	1.34	88.34
Multi-modal	VFF[5]	2022	85.50	84.48	**1.25**
	LoGoNet[6]	2023	87.13	85.52	1.81
	FogFusion (Ours)	-	**88.94**	**87.74**	**1.35**

C. *Qualitative results*

To further demonstrate the effectiveness of the proposed model, we visualized detection results for qualitative analysis. FogFusion shows two main advantages: (1) It avoids false detections caused by backscattered point clouds. As shown in Fig. 4(a), the baseline mistakenly recognizes the false point cloud as a car (yellow box), whereas FogFusion in Fig. 4(b) removes this interference and produces no false detection boxes. (2) It compensates for the increased sparsity caused by attenuation. In Fig. 4(c), the baseline generates a false detection (yellow box) in long-range perception because the sparse point cloud lacks sufficient semantic and geometric

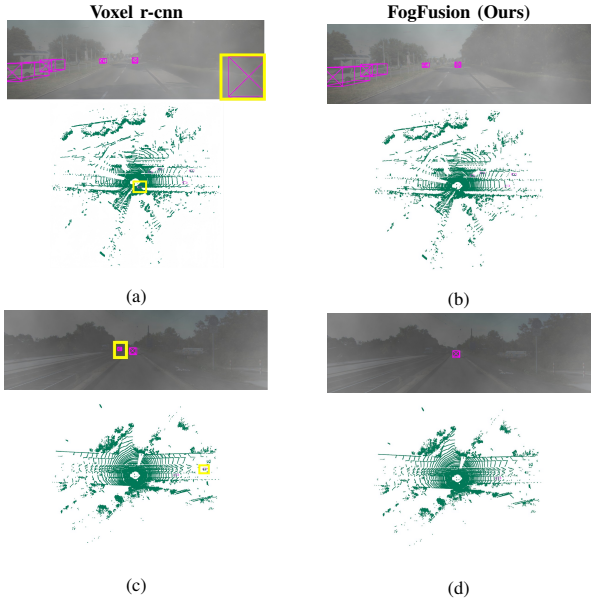

Fig. 4. Qualitative results of FogFusion.

information. In contrast, Fig. 4(d) shows that the proposed method overcomes this issue and yields accurate detection.

IV. CONCLUSION

Unpredictable weather should not compromise interaction safety. In this extended abstract, we present FogFusion, a camera-LiDAR fusion network for mitigating sensor degradation in foggy environments. By recovering reliable 3D spatial states from attenuated signals, FogFusion addresses a key challenge for autonomous systems in real-world settings. Our evaluations show that it maintains robust detection of traffic agents under severe visibility degradation, providing reliable spatial inputs for downstream behavior prediction. Future work will integrate this perception backbone with pedestrian behavior datasets to build an end-to-end prediction pipeline robust to adverse weather.

REFERENCES

[1] S. Shi, C. Guo, L. Jiang, Z. Wang, J. Shi, X. Wang, and H. Li, "PV-RCNN: Point-Voxel Feature Set Abstraction for 3D Object Detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020.

[2] H. Wu, C. Wen, W. Li, X. Li, R. Yang, and C. Wang, "Transformation-equivariant 3D object detection for autonomous driving," in *Proc. AAAI Conf. Artif. Intell.*, vol. 37, no. 3, pp. 2795–2802, 2023.

[3] T. Wang, X. Zhu, J. Pang, and D. Lin, "Probabilistic and Geometric Depth: Detecting Objects in Perspective," in *Proc. 5th Conf. Robot Learn.*, A. Faust, D. Hsu, and G. Neumann, Eds., vol. 164 of *Proc. Mach. Learn. Res.*, pp. 1475–1485, 2022.

[4] D. Rukhovich, A. Vorontsova, and A. Konushin, "ImVoxelNet: Image to Voxels Projection for Monocular and Multi-View General-Purpose 3D Object Detection," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, 2022, pp. 2397–2406.

[5] Y. Li, X. Qi, Y. Chen, L. Wang, Z. Li, J. Sun, and J. Jia, "Voxel Field Fusion for 3D Object Detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 1120–1129.

[6] X. Li, T. Ma, Y. Hou, B. Shi, Y. Yang, Y. Liu, X. Wu, Q. Chen, Y. Li, Y. Qiao, and L. He, "LoGoNet: Towards Accurate 3D Object Detection With Local-to-Global Cross-Modal Fusion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 17524–17534.