Collaborative Semantic Occupancy Prediction with Hybrid Feature Fusion in Connected Automated Vehicles Supplementary Material

Rui Song[1,2 *], Chenwei Liang[1], Hu Cao[2], Zhiran Yan[3], Walter Zimmer[2], Markus Gross[1], Andreas Festag[1,3], Alois Knoll[2]

[1]Fraunhofer IVI [2]Technical University of Munich [3]Technische Hochschule Ingolstadt

https://rruisong.github.io/publications/CoHFF

In this supplementary material, we provide more details of Semantic-OPV2V in Sec. 1. To showcase the robustness of the proposed CoHFF approach, we give extended results for its performance in relation to the communication budget and additionally assess its robustness in the presence of GPS noise in Sec. 2. We also present a range of visual results illustrating the effectiveness of CoHFF in diverse scenarios in Sec. 3. Note that we consistently use the same color scheme for each semantic class, as illustrated in the first column in Tab. 1.

1. Semantic-OPV2V dataset

We first equip each Connected and Automated Vehicle (CAV) in the CARLA simulation [1] with a semantic LiDAR at the position of each camera. This setup aims to capture the road environment within the Field of View (FoV) of the cameras as comprehensively as possible. Fig. 1 illustrates the semantically labeled point clouds generated by these semantic LiDARs.

Additionally, we outfit the surroundings of each CAV with a system comprising 18 semantic LiDARs to collect data on the road environment, including semantic occupancy space with occluded objects, as shown in Fig. 2. Specifically, we choose 9 positions surrounding each CAV, with each adjacent position spaced 30 meters apart. At each of these positions, we install two semantic LiDARs: one set at an vertical FoV ranging from -20 to -90 degrees, and the other ranging from -20 to 0 degrees.

By replaying the OPV2V dataset in CARLA-based OpenCDA [5], we collect semantically-labeled point clouds with 4 and 18 semantic LiDARs for each frame in the dataset. These point clouds are saved in PCD-format for further processing into semantic voxel data, useful for supervision or evaluation purposes.

*Corresponding author, email address: rui.song@ivi.fraunhofer.de

Table 1. CoHFF achieves robust IoU and mIoU performance, when the communication volume (CV) is reduced by setting various sparsification rates (Spar. Rate). The mask used for sparsification is learned under collaborative supervision.

Spar. Rate	0.00	0.50	0.80	0.95	0.99
CV (MB) (↓)	16.53	8.27	3.31	0.83	0.17
IoU (↑)	50.46	49.56	49.53	48.52	48.02
mIoU (↑)	34.16	32.97	32.70	30.13	29.48
Building	25.72	17.77	16.79	13.08	12.12
Fence	27.83	29.61	29.12	25.25	22.76
Terrain	48.30	47.98	47.60	44.42	44.77
Pole	42.74	37.73	37.69	35.65	35.83
Road	61.77	59.47	60.15	59.42	59.86
Side walk	39.62	42.03	41.36	40.81	39.11
Vegetation	20.59	21.36	20.18	13.35	14.74
Vehicles	63.28	60.25	60.33	60.14	59.98
Wall	58.27	52.68	53.41	51.94	51.20
Guard rail	1.94	3.86	3.51	1.66	1.55
Traffic signs	16.33	19.50	19.09	13.13	10.74
Bridge	3.53	3.39	3.11	2.67	1.11

Moreover, to train the Depth Net, we gather corresponding depth labels for the RGB cameras in the training dataset, as shown in Fig. 3. For a visual evaluation, we transform and visualize the results of depth estimation in the 3D voxel space. Fig. 4 compares these results with voxels based on raw LiDAR and collaborative semantic voxel labels.

2. Robustness

2.1. Low communication budget

We present additional results of the CoHFF performance in reducing the communication budget in Tab. 1. This in-

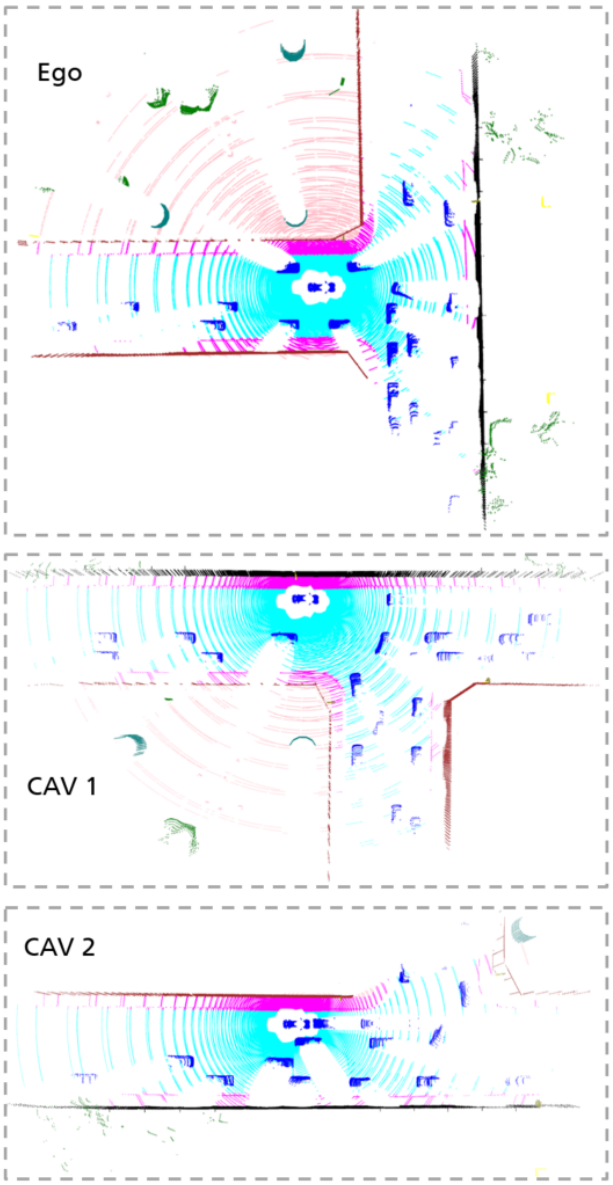

Figure 1. Visualization of semantic point clouds from 4 semantic LiDARs in Ego vehicle and CAVs.

cludes an assessment of the robust performance for overall Intersection over Union (IoU) as well as individual IoU for each class.

2.2. GPS noise

In our paper, we assess the performance of CoHFF using accurate GPS information. This section extends the experiment to include scenarios with varying GPS noise levels in Fig. 5, specifically Gaussian noise with a standard deviation ranging from 0 m to 0.6 m, which aligns with methodologies used in previous work, such as [2–4] for evaluating

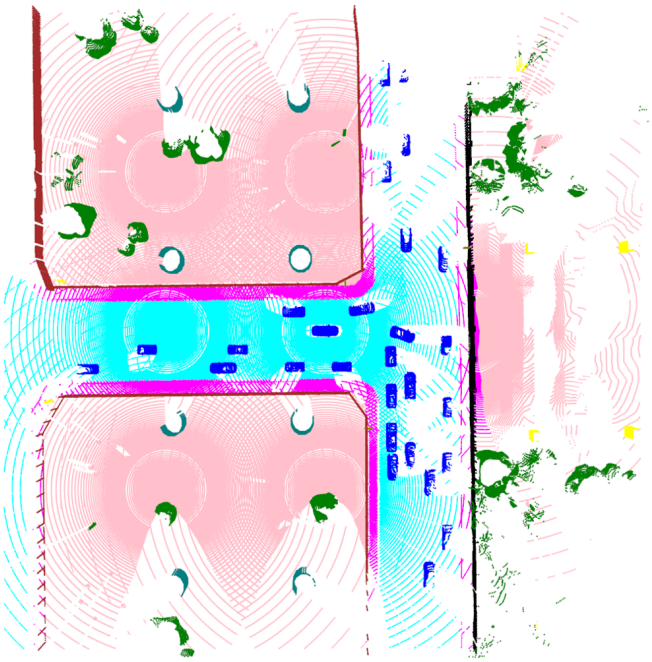

Figure 2. Visualization of semantic point clouds from 18 semantic LiDARs.

collaborative perception.

3. Further visual results

We provide a further visual comparison of CoHFF prediction results with collaborative and ego ground truth (GT) in an urban lane-change scenario in Fig. 6, an urban junction scenario in Fig. 7 and a highway scenario in Fig. 8. Our results demonstrate that the collaborative semantic occupancy prediction using CoHFF can achieve more complete perception than the ground truth in ego GT.

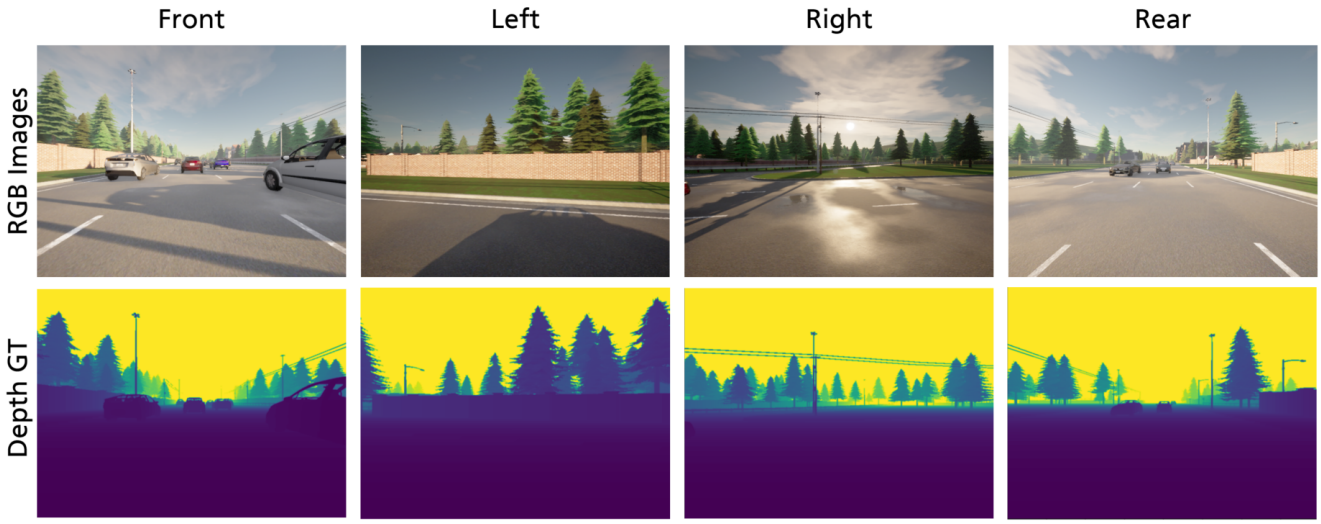

Figure 3. Corresponding depth labels gathered for the RGB cameras in the training dataset.

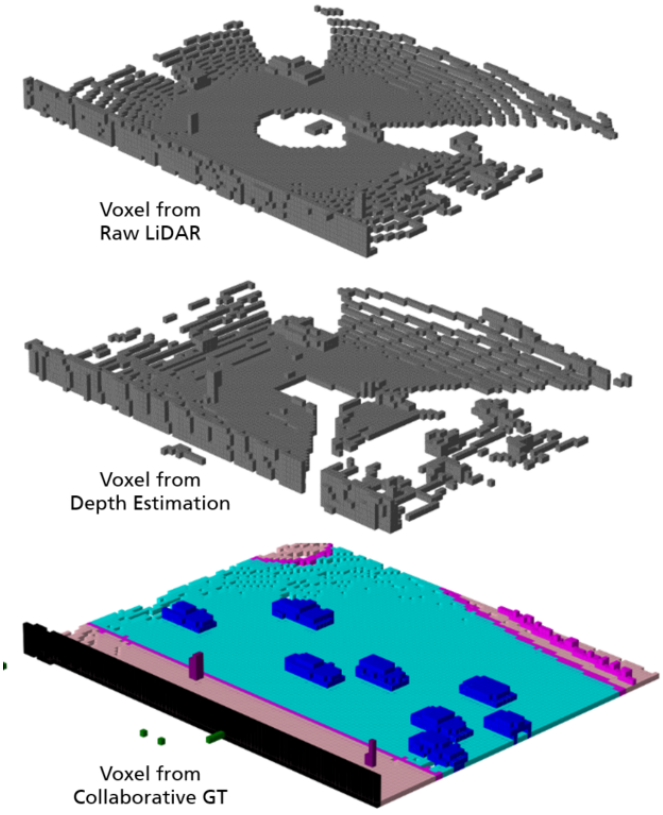

Figure 4. Visual comparison of occupied voxels derived from depth estimation, raw LiDAR, and collaborative semantic voxel labels. The gray color represents occupied voxels with an unknown semantic label.

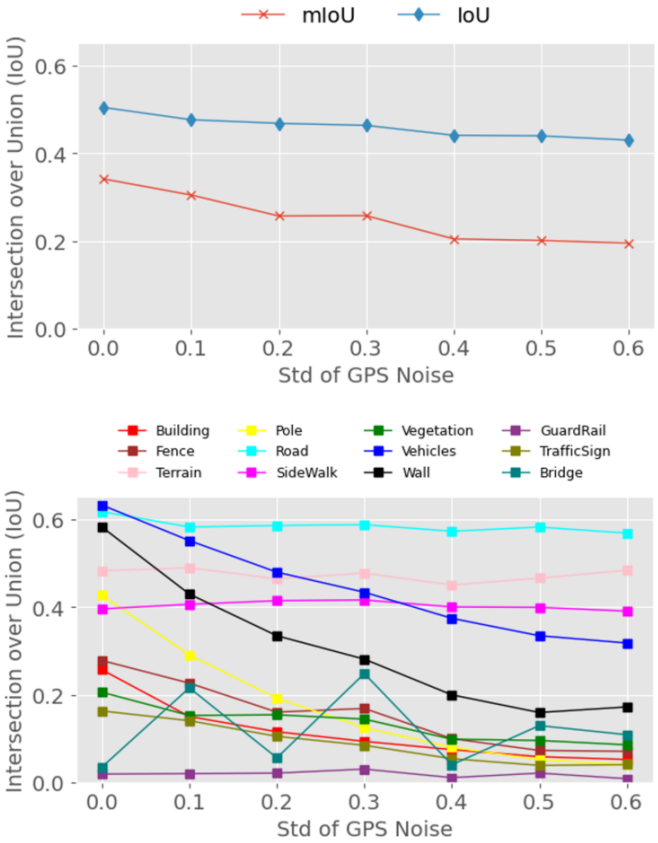

Figure 5. Our CoHFF model demonstrates robust performance in terms of overall IoU and mIoU stability. However, the IoU for each class exhibits individual variations, reflecting the unique impact of GPS noise on different categories.

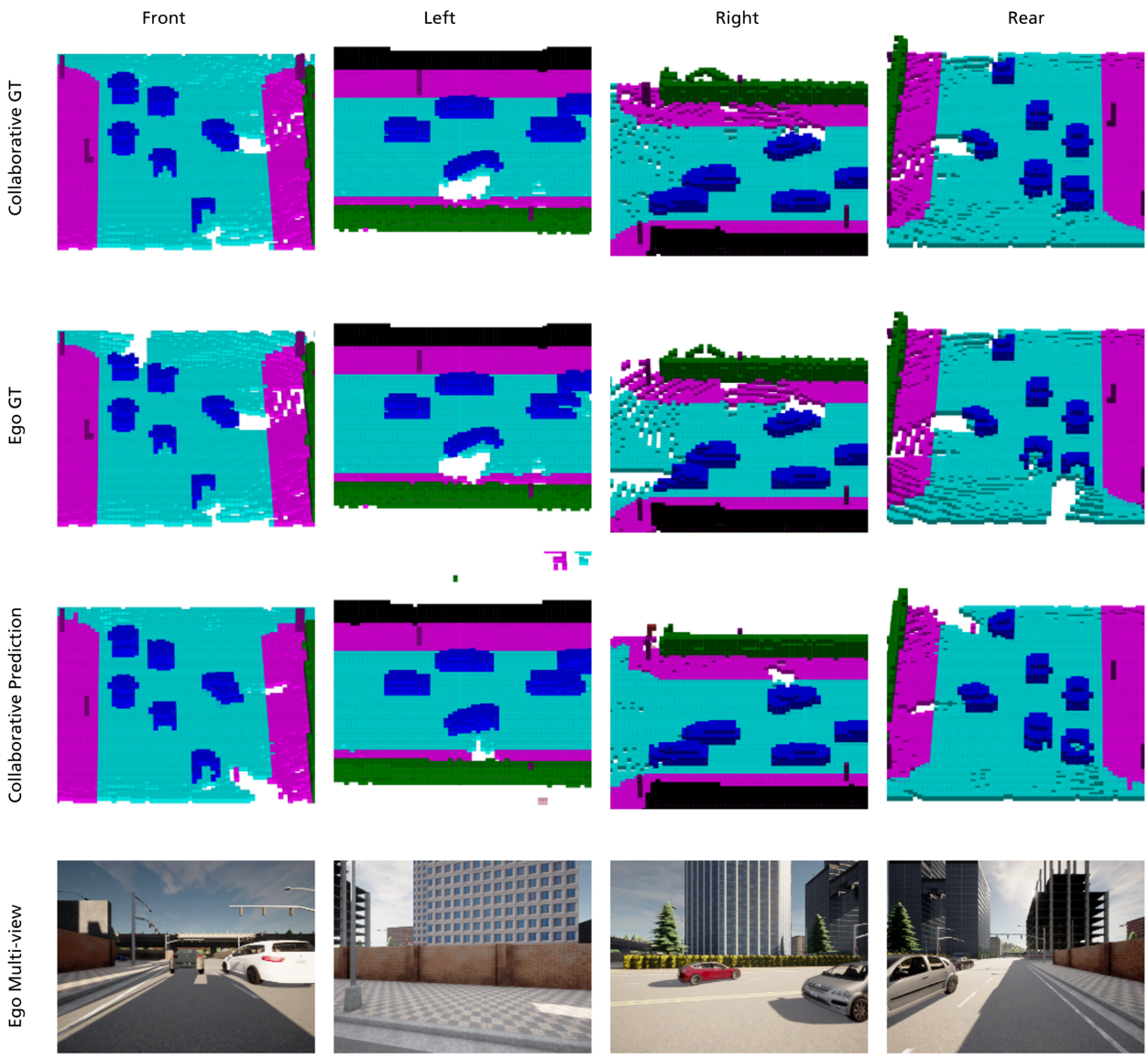

Figure 6. Visual comparison of CoHFF prediction results with collaborative and ego GT in an urban lane-change scenario.

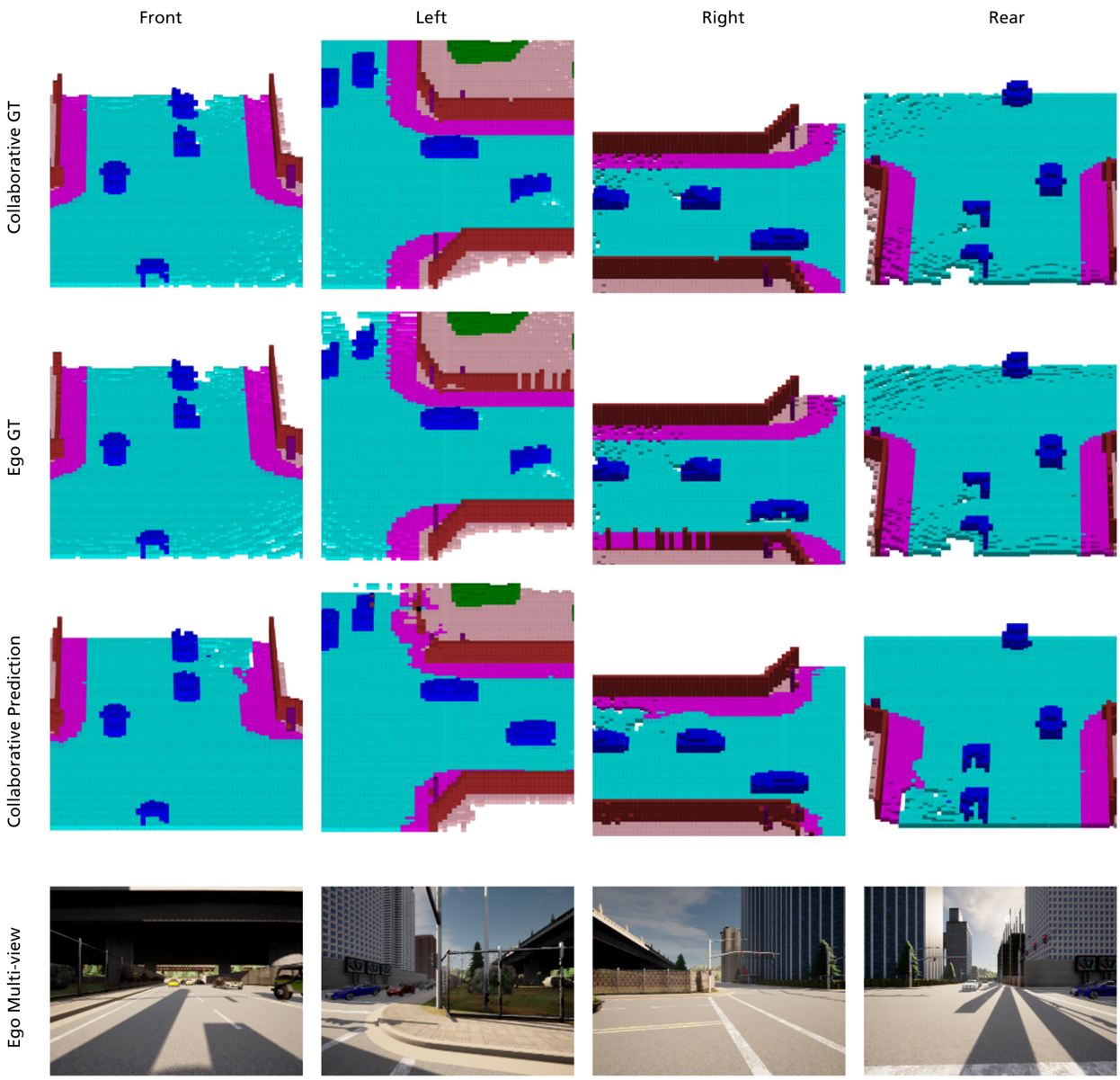

Figure 7. Visual comparison of CoHFF prediction results with collaborative and ego GT in an urban junction scenario.

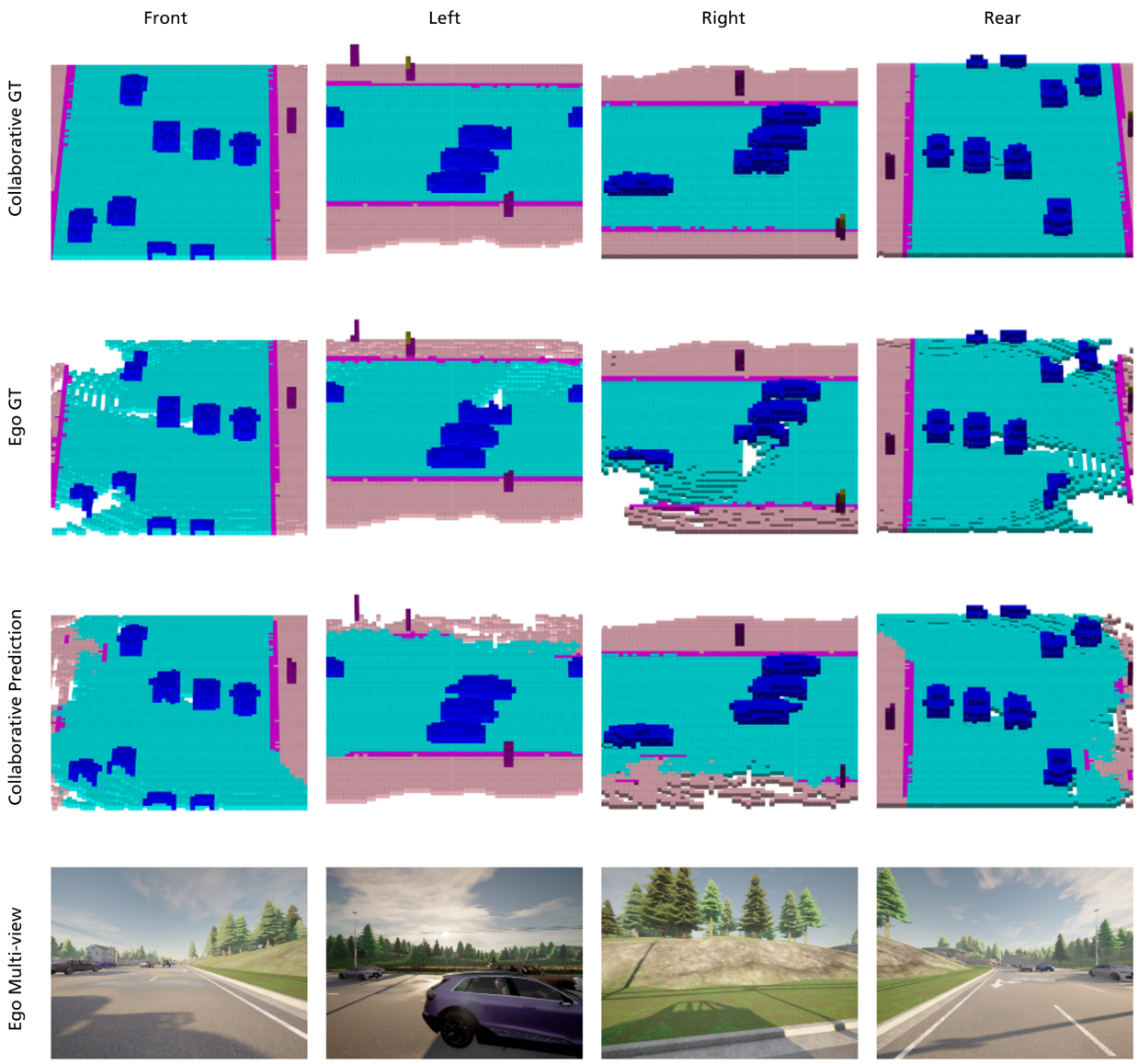

Figure 8. Visual comparison of CoHFF prediction results with collaborative and ego GT on a highway scenario.

References

[1] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. Carla: An open urban driving simulator. In *Conference on Robot Learning*, pages 1–16. PMLR, 2017. 1

[2] Yue Hu, Shaoheng Fang, Zixing Lei, Yiqi Zhong, and Siheng Chen. Where2comm: Communication-efficient collaborative perception via spatial confidence maps. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:4874–4886, 2022. 2

[3] Yue Hu et al. Collaboration helps camera overtake LiDAR in 3D detection. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9243–9252. IEEE, 2023. DOI: 10.1109/CVPR52729.2023.00892.

[4] Runsheng Xu et al. V2X-VIT: Vehicle-to-everything cooperative perception with vision transformer. In *2022 European Conference on Computer Vision (ECCV)*, pages 107–124. Springer, 2022. DOI: 10.1007/978-3-031-19842-7_7. 2

[5] Runsheng Xu et al. The OpenCDA open-source ecosystem for cooperative driving automation research. *IEEE Transactions on Intelligent Vehicles*, 8(4):2698–2711, 2023. DOI: 10.1109/TIV.2023.3244948. 1