OPV2V: An Open Benchmark Dataset and Fusion Pipeline for Perception with Vehicle-to-Vehicle Communication

Runsheng Xu[1*], Hao Xiang[1*], Xin Xia[1], Xu Han[1], Jinlong Li[2], Jiaqi Ma[1]

***Abstract*— Employing Vehicle-to-Vehicle communication to enhance perception performance in self-driving technology has attracted considerable attention recently; however, the absence of a suitable open dataset for benchmarking algorithms has made it difficult to develop and assess cooperative perception technologies. To this end, we present the first large-scale open simulated dataset for Vehicle-to-Vehicle perception. It contains over 70 interesting scenes, 11,464 frames, and 232,913 annotated 3D vehicle bounding boxes, collected from 8 towns in CARLA and a digital town of Culver City, Los Angeles. We then construct a comprehensive benchmark with a total of 16 implemented models to evaluate several information fusion strategies (i.e. early, late, and intermediate fusion) with state-of-the-art LiDAR detection algorithms. Moreover, we propose a new Attentive Intermediate Fusion pipeline to aggregate information from multiple connected vehicles. Our experiments show that the proposed pipeline can be easily integrated with existing 3D LiDAR detectors and achieve outstanding performance even with large compression rates. To encourage more researchers to investigate Vehicle-to-Vehicle perception, we will release the dataset, benchmark methods, and all related codes in https://mobility-lab.seas.ucla.edu/opv2v/.**

I. INTRODUCTION

Perceiving the dynamic environment accurately is critical for robust intelligent driving. With recent advancements in robotic sensing and machine learning, the reliability of perception has been significantly improved [1], [2], [3], and 3D object detection algorithms have achieved outstanding performance either with LiDAR point clouds [4], [5], [6], [7] or multi-sensor data [8], [9].

Despite the recent breakthroughs in the perception field, challenges remain. When the objects are heavily occluded or have small scales, the detection performance will dramatically drop. Such problems can lead to catastrophic accidents and are difficult to solve by any algorithms since the sensor observations are too sparse. An example is revealed in Fig. 1a. Such circumstances are very common but dangerous in real-world scenarios, and these blind spot issues are extremely tough to handle by a single self-driving car.

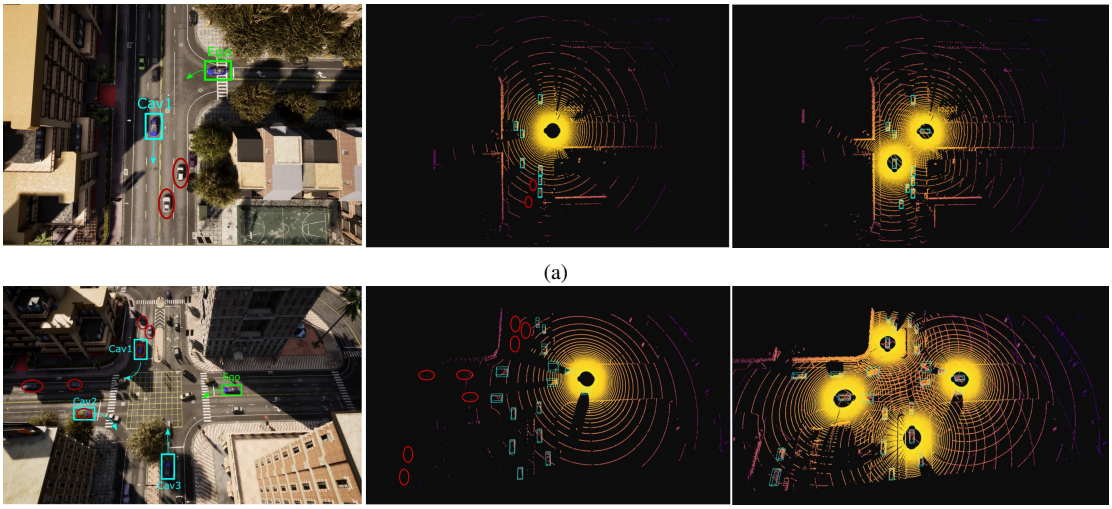
To this end, researchers started recently investigating dynamic agent detection in a cooperative fashion, such as USDOT CARMA [10] and Cooper [11]. By leveraging the Vehicle-to-Vehicle (V2V) communication technology, different Connected Automated Vehicles (CAVs) can share their sensing information and thus provide multiple viewpoints for the same obstacle to compensate each other. The shared information could be raw data, intermediate features, single CAV's detection output, and metadata e.g., timestamps and poses. Despite the big potential in this field, it is still in its infancy. One of the major barriers is the lack of a large open-source dataset. Unlike the single vehicle's perception area where multiple large-scale public datasets exist [12], [13], [14], most of the current V2V perception algorithms conduct experiments based on their customized data [15], [16], [17]. These datasets are either too small in scale and variance or they are not publicly available. Consequently, there is no large-scale dataset suitable for benchmarking distinct V2V perception algorithms, and such deficiency will preclude further progress in this research field.

To address this gap, we present OPV2V, the first large-scale **O**pen Dataset for **P**erception with **V2V** communication. By utilizing a cooperative driving co-simulation framework named OpenCDA [18] and CARLA simulator [19], we collect 73 divergent scenes with a various number of connected vehicles to cover challenging driving situations like severe occlusions. To narrow down the gap between the simulation and real-world traffic, we further build a digital town of Culver City, Los Angeles with the same road topology and spawn dynamic agents that mimic the realistic traffic flow on it. Data samples are shown in Fig. 1 and Fig. 4. We benchmark several state-of-the-art 3D object detection algorithms combined with different multi-vehicle fusion strategies. On top of that, we propose an Attentive Intermediate Fusion pipeline to better capture interactions between connected agents within the network. Our experiments show that the proposed pipeline can efficiently reduce the bandwidth requirements while achieving state-of-the-art performance.

II. RELATED WORK

Vehicle-to-Vehicle Perception: V2V perception methods can be divided into three categories: early fusion, late fusion, and intermediate fusion. Early fusion methods [11] share raw data with CAVs within the communication range, and the ego vehicle will predict the objects based on the aggregated data. These methods preserve the complete sensor measurements but require large bandwidth and are hard to operate in real time [15]. In contrast, late fusion methods transmit the detection outputs and fuse received proposals into a consistent prediction. Following this idea, Rauch et al. [20] propose a Car2X-based perception module to jointly align the shared bounding box proposals spatially and temporally via an EKF. In [21], a machine learning-based method is utilized to fuse proposals generated by different connected agents. This stream of work requires less bandwidth, but

*Equal contribution
[1]University of California, Los Angeles, Mobility Lab. {rxx3386, haxiang, x35xia, hanxu417, jiaqima}@ucla.edu
[2]Cleveland State University, Cleveland Vision and AI Lab, j.li56@vikes.csuohio.edu



(a)

(b)

Fig. 1: Two examples from our dataset. *Left*: Screenshot of the constructed scenarios in CARLA. *Middle*: The LiDAR point cloud collected by the ego vehicle. *Right*: The aggregated point clouds from all surrounding CAVs. The red circles represent the cars that are invisible to the ego vehicle due to the occlusion but can be seen by other connected vehicles. (a): The ego vehicle plans to turn left in a T-intersection and the roadside vehicles block its sight to the incoming traffic. (b): Ego-vehicle's LiDAR has no measurements on several cars because of the occlusion caused by the dense traffic.

Sensors	Details
4x Camera	RGB, 800×600 resolution, $110°$ FOV
1x LiDAR	64 channels, 1.3 M points per second, 120 m capturaing range, $-25°$ to $5°$ vertical FOV, ± 2 cm error
GPS & IMU	20 mm positional error, $2°$ heading error

TABLE I: Sensor specifications.

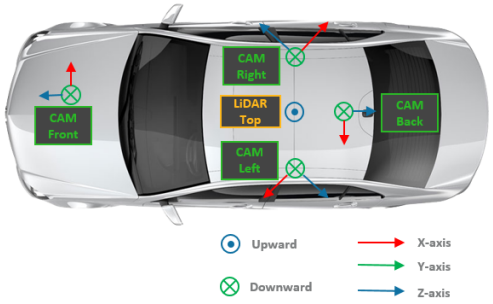


Fig. 2: Sensor setup for each CAV in OPV2V.

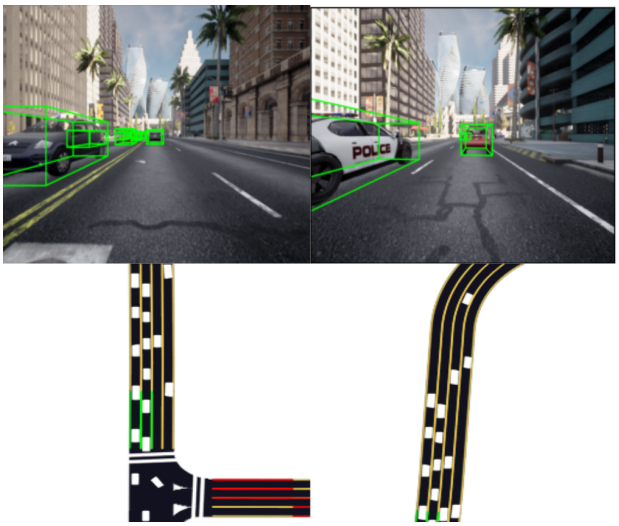


Fig. 3: Examples of the front camera data and BEV map of two CAVs in OPV2V. The yellow, green, red, and white lanes in the BEV map represent the lanes without traffic light control, under green light control, under red light control, and crosswalks.

the performance of the model is highly dependent on each agent's performance within the vehicular network. To meet requirements of both bandwidth and detection accuracy, intermediate fusion [22], [15] has been investigated, where intermediate features are shared among connected vehicles and fused to infer the surrounding objects. F-Cooper [22] utilizes max pooling to aggregate shared Voxel features, and V2VNet [15] jointly reason the bounding boxes and trajectories based on shared messages.

Vehicle-to-Vehicle Dataset: To the best of our knowledge, there is no large-scale open-source dataset for V2V perception in the literature. Some work [11], [22] adapts KITTI [14] to emulate V2V settings by regarding the ego vehicle at different timestamps as multiple CAVs. Such synthetic procedure is unrealistic and not appropriate for V2V tasks since the dynamic agents will appear at different locations,