# V2X-Real: a Large-Scale Dataset for Vehicle-to-Everything Cooperative Perception

Hao Xiang, Zhaoliang Zheng, Xin Xia, Runsheng Xu, Letian Gao, Zewei Zhou, Xu Han, Xinkai Ji, Mingxi Li, Zonglin Meng, Li Jin, Mingyue Lei, Zhaoyang Ma, Zihang He, Haoxuan Ma, Yunshuang Yuan, Yingqian Zhao, Jiaqi Ma⋆

University of California, Los Angeles

**Abstract.** Recent advancements in Vehicle-to-Everything (V2X) technologies have enabled autonomous vehicles to share sensing information to see through occlusions, greatly boosting the perception capability. However, there are no real-world datasets to facilitate the real V2X cooperative perception research – existing datasets either only support Vehicle-to-Infrastructure cooperation or Vehicle-to-Vehicle cooperation. In this paper, we present V2X-Real, a large-scale dataset that includes a mixture of multiple vehicles and smart infrastructure to facilitate the V2X cooperative perception development with multi-modality sensing data. Our V2X-Real is collected using two connected automated vehicles and two smart infrastructure, which are all equipped with multi-modal sensors including LiDAR sensors and multi-view cameras. The whole dataset contains 33K LiDAR frames and 171K camera data with over 1.2M annotated bounding boxes of 10 categories in very challenging urban scenarios. According to the collaboration mode and ego perspective, we derive four types of datasets for Vehicle-Centric, Infrastructure-Centric, Vehicle-to-Vehicle, and Infrastructure-to-Infrastructure cooperative perception. Comprehensive multi-class multi-agent benchmarks of SOTA cooperative perception methods are provided. The V2X-Real dataset and codebase are available at https://mobility-lab.seas.ucla.edu/v2x-real.

**Keywords:** V2X Dataset · Cooperative Perception · Autonomous Driving

## 1 Introduction

The evolution of autonomous driving technology has been significantly accelerated by advancements in deep learning, particularly in perception tasks [17, 23, 27, 9, 38] crucial for safe navigation [11, 28, 8, 39] and decision-making [24]. While progress in various perception tasks has been notable, the single-vehicle vision systems still suffer from occlusions and limited perception range. The bottlenecks mainly stem from the fact that each vehicle can only assess its surroundings from a single view point, leading to a partial understanding of the

---

⋆ Corresponding author. Work done at UCLA Mobility Lab.

environment. To address these issues, recent works [22, 35, 36, 12, 26, 16] have investigated Vehicle-to-Everything (V2X) Cooperative Perception, where each connected agent within the communication range can share the sensing information (*e.g.*, raw point clouds, detection results and intermediate neural features) with each other and leverage neighboring agent's sensing information for visual reasoning to enlarge the perception range and see through occlusions.

Despite the great potential, there is no open V2X dataset from real-world scenarios for researchers to work with. The majority of available V2X datasets such as OPV2V [35], V2X-Sim [12] and V2X-Set [34] are generated through simulators [3] with simulated traffic dynamics and sensor renderings. This approach introduces a significant simulation-to-reality gap, thereby impeding the effective evaluation and deployment of V2X algorithms in the real world [32]. Recently, DAIR-V2X [36] and V2V4Real [32] present the pioneering real-world datasets for Vehicle-to-Infrastructure (V2I) and Vehicle-to-Vehicle (V2V) collaborations. However, these datasets are limited by a single collaboration mode (*i.e.* either V2V or V2I), involving at most two agents within the same spatial vicinity. Nevertheless, in real world, V2X collaborations can dynamically adapt to multifaceted interactions inherent in real-world traffic scenarios, including V2V, V2I, I2I and thus truly V2X. Developing algorithms based on single collaboration mode will greatly constrain the scope of V2X collaborations and hinder deploying the algorithms in complex V2X scenarios. In addition, existing datasets are vehicle-centric by assuming the vehicle as the ego agent, ignoring the pivotal role of infrastructure; however, the infrastructure perception is vital for both autonomous driving and intelligent transportation systems. On the one hand, it can augment the perception systems of self-driving cars by sharing sensing information with nearby agents. On the other hand, infrastructure perception can provide a reliable basis for various intelligent transportation tasks such as traffic management, traffic monitoring and signal control, which will greatly benefit from the V2I and I2I collaborations. Moreover, DAIR-V2X [36] and V2V4Real [32] are mostly collected in suburban and rural areas with an average of 11-12 objects per scene. This relatively low traffic density limits their effectiveness in testing V2X cooperative perception in more challenging scenarios and restricts the scalability of V2X collaborations.

To this end, we introduce the V2X-Real, a large-scale multi-modal multi-view dataset tailored for V2X research, including 33K LiDAR frames, 171K camera images and over 1.2M 3D bounding box annotations. The proposed dataset is collected via two smart infrastructure and two autonomous vehicles equipped with full sensor suits (Fig. 1). To ensure the data diversity, we gather the data in two types of scenarios *i.e.*, V2X smart intersection and V2V corridors. With over 75 hours of collected driving logs, we carefully curate and select 68 representative scenarios to form the final dataset. Unlike previous datasets, V2X-Real is primarily collected in urban environments, featuring an average of 36 objects per scene. Such environments offer a high density of traffic and a significant presence of vulnerable road users, providing rich challenging scenarios for V2X cooperative perception research. To innovate and support a wide range of research interests
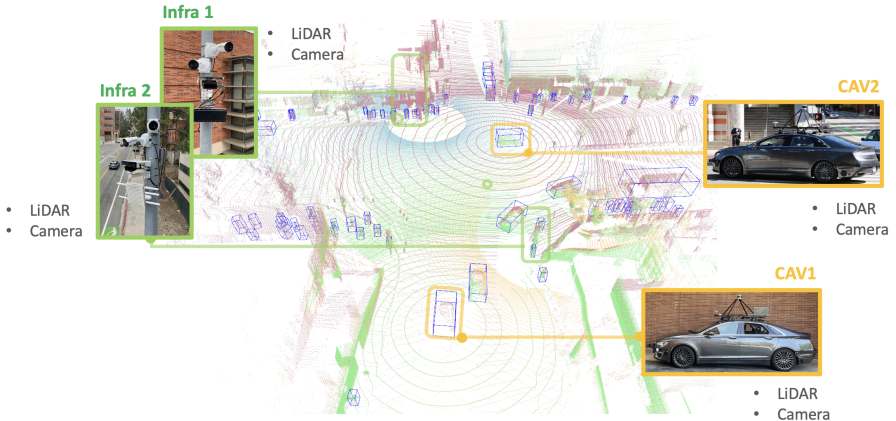
Fig. 1: Demonstration of smart intersection and overall data acquisition systems. Smart infrastructure (Infra) units and connected and automated vehicles (CAVs) are depicted in their combined LiDAR point clouds.

in V2X cooperative perception, the dataset is divided into four specialized sub-datasets based on the type of ego agent and their collaboration modes: V2X-Real-VC for Vehicle-Centric, V2X-Real-IC for Infrastructure-Centric, V2X-Real-V2V for Vehicle-to-Vehicle, and V2X-Real-I2I for Infrastructure-to-Infrastructure cooperative perception.

The key contributions can be summarized as follows:

- We build V2X-Real, the first open large-scale real-world dataset designed for V2X cooperative perception. According to ego agent's type and its collaboration mode, we derive 4 sub-datasets tailored for vehicle-centric, infrastructure-centric, V2V and I2I cooperative perception.
- We provide over 1.2M annotated bounding boxes of 10 object categories with 33K LiDAR frames and 171K multi-view camera data.
- We conduct comprehensive benchmarks for multi-class multi-agent V2X cooperative perception for V2X-Real-VC, V2X-Real-IC, V2X-Real-V2V, and V2X-Real-I2I.

## 2   Related Work

### 2.1   Self-driving datasets

Over the past few decades, a number of autonomous driving datasets have been released, greatly promoting the research and development of self-driving technologies. Among these datasets, KITTI is a pioneering autonomous driving dataset featured with full sensor suits [5]. Following KITTI, the NuScenes [1] and Waymo Open datasets [20] have expanded the scale of data significantly, offering both LiDAR point clouds and multi-view cameras that cover a 360-degree

Table 1: Comparisons of our dataset and existing autonomous driving and V2X dataset. V2V: Vehicle-to-Vehicle. V2I: Vehicle-to-Infrastructure. I2I: Infrastructure-to-Infrastructure. VC: Vehicle-Centric. IC: Infrastructure-Centric.

| Dataset | Year | Type | V2V | V2I | I2I | Image (360°) | Agent Number | RGB images | LiDAR frames | 3D boxes | Types |
|---|---|---|---|---|---|---|---|---|---|---|---|
| KITTI [5] | 2012 | Real | | | | | 1 | 15k | 15k | 200k | 8 |
| nuScenes [1] | 2019 | Real | | | | ✓ | 1 | 1.4M | 400k | 1.4M | 23 |
| Waymo Open [20] | 2019 | Real | | | | ✓ | 1 | 1M | 200k | 12M | 4 |
| OPV2V [35] | 2022 | Sim | ✓ | | | ✓ | 2.89 | 44k | 11k | 230k | 1 |
| V2X-Sim [12] | 2022 | Sim | ✓ | ✓ | | ✓ | 10 | 0 | 10k | 26.6k | 1 |
| V2XSet [34] | 2022 | Sim | ✓ | ✓ | | ✓ | 2-7 | 44k | 11k | 230k | 1 |
| DAIR-V2X [36] | 2022 | Real | | ✓ | | | 2 | 39K | 39K | 464K | 10 |
| V2V4Real [32] | 2023 | Real | ✓ | | | | 2 | 40K | 20K | 240K | 5 |
| V2X-Real (ours) | 2024 | Real | ✓ | ✓ | ✓ | ✓ | **4** | **171K** | 33K | **1.2M** | **10** |
| – V2X-Real-VC | 2024 | Real | ✓ | ✓ | | ✓ | 4 | 145K | 30K | 1.1M | 10 |
| – V2X-Real-IC | 2024 | Real | | ✓ | ✓ | ✓ | 4 | 145K | 30K | 1.1M | 10 |
| – V2X-Real-V2V | 2024 | Real | ✓ | | | ✓ | 2 | 140K | 17K | 719K | 10 |
| – V2X-Real-I2I | 2024 | Real | | | ✓ | | 2 | 31K | 15K | 470K | 10 |

field of view, thereby enhancing the depth and breadth of research possibilities in the field. However, these datasets are all collected by a single vehicle. OPV2V introduces the first V2V cooperative perception dataset, created using CARLA simulator [3] and OpenCDA [30, 33, 37]. V2XSet and V2X-Sim [12] further extended to V2X scenarios in the CARLA simulator. In contrast to these simulated datasets, DAIR-V2X [36] is the first real-world dataset for V2I cooperative perception while V2V4Real [32] presents the first real-world V2V dataset collected by two connected vehicles.

Nevertheless, existing V2X datasets all focus on a single collaboration mode. Furthermore, the benchmarks and evaluations within these datasets are vehicle-centric, treating infrastructure merely as an auxiliary collaborator. Infrastructure-centric collaborations such as V2I and I2I are rarely discussed, highlighting a gap in the comprehensive evaluation of V2X interactions. A concurrent work, RCooper [6], introduces an I2I dataset for roadside cooperative perception. Compared with these datasets, our dataset features with all three V2X collaboration modes *i.e.* V2V, V2I, and I2I, and investigates and benchmarks both Vehicle-Centric and Infrastructure Centric V2X cooperative perception. Moreover, it is uniquely collected in congested urban environments with dense traffic and pedestrian flow, resulting in a record number of annotated 3D boxes compared to existing datasets. Furthermore, to facilitate multi-view camera-based perception, we will also provide multi-view stereo cameras covering 360 field-of-view for the vehicle agent and high-resolution network camera data for the infrastructure. Additionally, existing V2X datasets only benchmark the detection perfor-

mance based on a single vehicle category, ignoring the perception performance for vulnerable road users. In contrast, we provide benchmarks for all the annotation types and will release our codebase for multi-class cooperative perception to facilitate V2X research on vulnerable road users, which is an important yet under-explored research area.

Table 2: Key Sensor Specifications in V2X-Real Dataset. Infra means infrastructure.

| Agent | Sensor | Sensor Model | Details |
|-------|--------|--------------|---------|
| Infra | LiDAR | Ouster OS1-128/64 (×1) | 128/64 beams, 10Hz capture frequency, 360° horizontal FOV, $-16.6°$ to $+16.6°(64)$, $-22.5°$ to $+22.5°(128)$ vertical FOV, $\leq 40$m range |
| | Camera | Axis-P14555 (×2) | RGB, 30Hz capture frequency, 1920 × 1080 resolution, 76° horizontal FOV, 48° vertical FOV, JPEG compressed |
| | GPS | Garmin 18x LVC GPS (×1) | 5Hz update rate |
| Vehicle | LiDAR | RoboSense Ruby Plus (×1) | 128 beams, 10Hz capture frequency, 360° horizontal FOV, $-25°$ to $+15°$ vertical FOV, $\leq 200$m range |
| | Camera | ZED 2i (×4) | Stereo RGBD, 10Hz capture frequency, 1920 × 1080 resolution, 120° horizontal FOV, JPEG compressed |
| | GPS/IMU | Gongji GPS System (×1) | 1000Hz update rate, Double Precision |

## 2.2 V2X Cooperative Perception

Cooperative perception aims to enhance the perception capability of the V2X system by leveraging the shared information among connected agents. According to its fusion strategies, it can be broadly classified into three categories: 1) Early Fusion where the raw LiDAR point clouds are transmitted and the ego agent will aggregate the raw data to predict objects, 2) Late Fusion [18] where the detection outputs are circulated and then fused into a consistent prediction, 3) Intermediate Fusion [22, 34, 35, 7, 13, 19, 29]where the intermediate neural features are shared and fused. Though early Fusion preserves the raw information and thus can potentially reach higher accuracy, the high bandwidth requirement of transmitting raw data makes it infeasible in real-world scenarios. On the other hand, Late Fusion requires minimal bandwidth but loses the context information during transmission, resulting in inferior performance with compound errors. To strike a good balance between bandwidth requirements and accuracy, Intermediate Fusion has been gaining increasing attention. Att-Fuse [35] presents an attention-based fusion strategy to refine multi-agent's BEV

features. V2VNet [22] employs a graph neural network architecture for joint perception and prediction. F-Cooper [2] presents a pooling mechanism to identify salient features. V2X-ViT [34] and CoBEVT [31] employ a sparse vision transformer for V2X collaborations. HM-ViT [26] presents a 3D heterogeneous graph transformer for camera and LiDAR fusion.

## 3 V2X-Real datasets

To facilitate the V2X research, we present V2X-Real, the first large-scale real-world V2X dataset for cooperative perception. In this section, we first detail the data acquisition in Sec. 3.1, then present the data annotation and processing pipeline in Sec. 3.2, and finally discuss the diverse data distribution and dataset statistics in Sec. 3.3

### 3.1 Data Acquisition

**Sensor Specifications:** The dataset was gathered using two connected and automated vehicles and two smart infrastructure. Each of the vehicles is outfitted with a 128-beam LiDAR, four ZED 2i Stereo Cameras, and a GPS/IMU integration system. The four cameras are strategically positioned on the front, rear, left and right sides of the vehicle to achieve a full 360-degree field-of-view (FoV). Similarly, each infrastructure unit is equipped with a 128(64)-beam LiDAR, two Axis high-resolution network cameras and a GPS to ensure precise time synchronization among the distributed devices. Detailed specifications of these sensors are provided in Tab. 2.

**Data Collection:** To ensure the diversity of our V2X data, We collected our data in two types of scenarios – V2X intersection and V2V corridor. In the intersection scenarios, two smart infrastructure units and two autonomous vehicles are operated simultaneously. Our autonomous vehicles will follow predefined routes, covering different V2X interactions. Through a combination of basic maneuvers of each autonomous vehicle at the intersection – such as left turn, right turn, waiting for a traffic signal, and going straight – we designed a total of 50 distinct V2X interactions in our dataset including car following, two cars driving towards each other, two cars making left (right) turn together *et al*. As the V2X data is primarily collected in a single smart intersection and its surrounding streets for diverse V2X interactions, to enlarge the diversity of our dataset, we further collected separate V2V scenarios in urban corridor. In these V2V data, we designed 10 types of V2V interactions, including car following, lane change, and overtaking between two autonomous vehicles *et al*.

Our data collection efforts span 3 months in diverse illumination conditions, with a total of over 75 hours of collected driving logs. We carefully curate the data and select 68 representative scenarios for annotation with a total of 33K LiDAR frames and 171K Camera images, providing a rich resource for V2X research.

(a) Vehicle calibration results   (b) Infrastructure calibration results
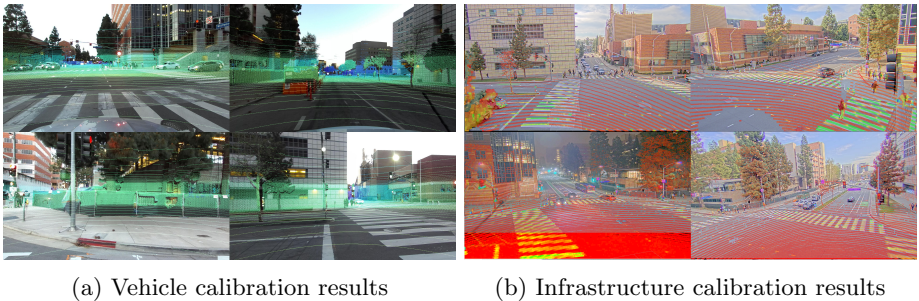
Fig. 2: Visualization of calibration results for (a) vehicle and (b) infrastructure (better viewed in color and zoomed in). The LiDAR points are projected onto the camera plane using the camera intrinsics and camera-LiDAR extrinsics.
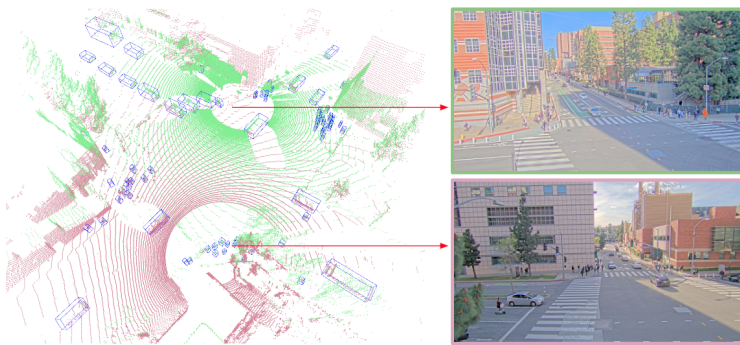


Fig. 3: Demonstration of combined LiDAR annotations for infrastructure. LiDAR points from different agents (green and pink) are aggregated for the same object, reducing half of the infrastructure annotation workload and enhancing annotators' ability to identify objects accurately and efficiently.

**Coordinate system:** There are three types of coordinate systems in our dataset, namely the LiDAR coordinate, the camera coordinate, and the map coordinate. Each agent maintains a separate LiDAR and camera coordinate frame. We built a dense LiDAR point cloud map [25] covering all the driving routes and leverage map matching combined with GPS/IMU positional inputs [4] to identify the transformation between each agent's LiDAR coordinate frame and map coordinate. For 3D-2D associations, intrinsic and extrinsic calibrations between the camera and LiDAR are conducted. The visualization of the calibration result is shown in Fig. 2.

**Synchronization:** For the multi-agent datasets, it is crucial to synchronize the time clock of all agents. We synchronize all the computer's time clock with GPS time to ensure consistent timestamp across sensing messages and further lock the LiDAR phase with GPS signals to ensure the temporal synchrony of multi-agent sensing observation. As the ZED2i stereo camera and Axis network

camera cannot be hardware-triggered with GPS signals for synchronization, we post-process the collected data to identify the closest camera message as per LiDAR frame to reach minimal time difference. Messages with a large time gap are filtered to ensure suitable synchronization between multi-agent multi-modal data.

## 3.2 Data Annotation and Processing

We leverage the open-sourced labeling tool called SUSTechPOINTS [10] to annotate 3D bounding boxes. Expert annotators are employed to make high-quality annotations with a stringent review and revision process in place. On average, each scenario goes through 5-6 rounds of review and revision to guarantee accuracy and quality. In our dataset, there are in total 10 categories of annotated objects including pedestrian, scooter, motorcycle, bicycle, car, truck, van, box truck, bus and barrier on the road. For each object, we annotate its 9-dimensional cuboid containing x, y, z for the cuboid center, width, length, height for the cuboid extent dimension, and roll, yaw, pitch for the cuboid orientation.



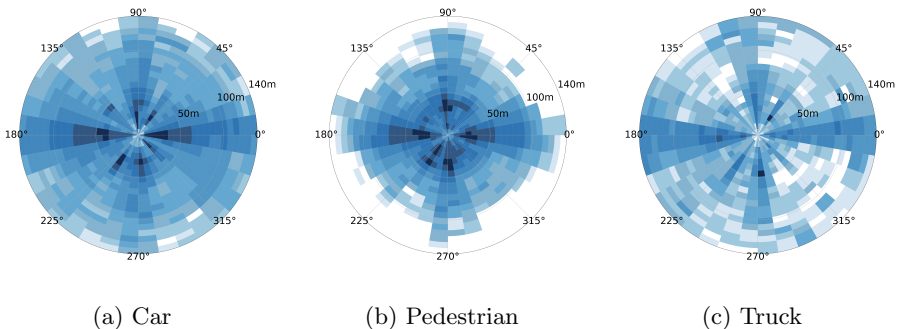|        (a) Car        |     (b) Pedestrian     |       (c) Truck       |

Fig. 4: Polar density map of annotated bounding boxes for (a) Car, (b) Pedestrian, and (c) Truck. The polar and radial axes represent the distance and angle of the bounding boxes respectively in the ego coordinate frame. The color indicates the number of boxes in the bin and the darker color corresponds to a larger density.

Additionally, we have implemented several strategies to enhance the efficiency and accuracy of the annotation process. We combine two infrastructure's LiDAR point clouds to form a holistic view for annotation (Fig. 3). The combined LiDAR point clouds can efficiently reduce the annotation workload and provide dense sensing measurements for objects of interest, greatly increasing the annotation efficiency and accuracy. After annotating the 3D boxes in combined LiDAR point clouds, we will project the boxes into each infrastructure's coordinate frame and for each box, we will count the number of LiDAR points within the box. If the

number is greater than a threshold (*i.e.*, 5 in our dataset), we will add the box to this infrastructure's annotation result, otherwise, the box is ignored for this agent due to its sparse sensing measurements. Afterwords, we manually check the result and refine the annotations for each of the agents. For vehicles, their poses vary from scene to scene with different levels of localization error, thus separate annotations for each of the vehicles are conducted to ensure accurate annotations. Additionally, for the static objects (*e.g.*, vehicles that are parked and those awaiting the change of a traffic signal), we provide annotators with a script to automatically interpolate their poses in one frame to a sequence of frames. This greatly increased the annotation accuracy and efficiency for static objects.
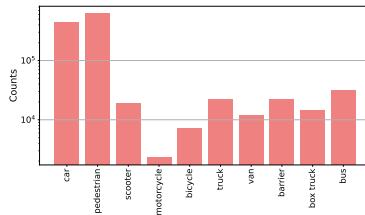


Fig. 5: Number of annotations per category in log scale.

As the objects are annotated separately in different agents' coordinate frames, identical objects may be assigned with different IDs. To ensure consistent object ID, we project all the annotated boxes from different agents to the same map coordinate and calculate their IoU overlap. If their overlap is greater than a certain threshold, they are considered the same object and thus a unique ID is assigned.

### 3.3 Dataset Analysis

Our dataset has an average of 36 objects per scene with a maximum of 115 and a minimum of 9 objects, ensuring a high density of traffic in challenging urban scenarios. The number of annotations for each of the 10 categories is shown in Fig. 5. The dominant objects in the dataset are pedestrians and cars. Compared with the existing V2X dataset (Tab. 1), V2X-Real is featured with a wide variety and quantity of vulnerable road users including pedestrian, scooter, motorcycle, and bicycle. In our benchmarks, we group the 10 categories, based on their box sizes, into three super-classes *i.e.* pedestrian, car, and truck. In Fig. 4, we reveal the polar density map of annotated bounding boxes for these three classes, where each bin in the polar map represents a spatial area with a certain distance and angle range in the ego agent's coordinate frame. From the figure, we can observe a diverse distribution of annotations and each category is denser in the front, left, right, and rare sides of the ego vehicle. We argue this is due to the fact that
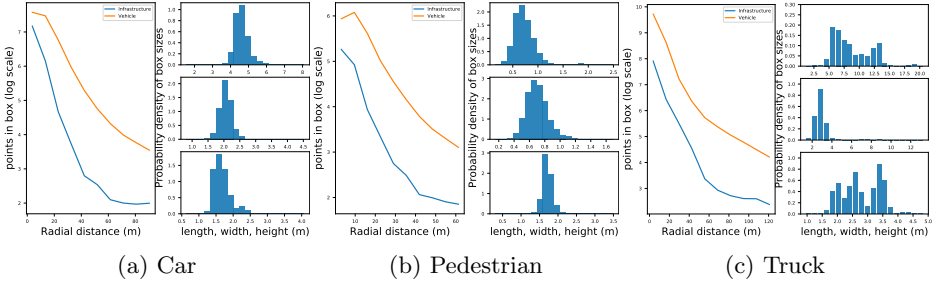
Fig. 6: Bounding box LiDAR point density distribution and box size distribution for (a) Car, (b) Pedestrian, and (c) Truck. In each category's figure, the left figure depicts the number of lidar points within bounding boxes as per the radial distance from the data collection vehicle while the right figure demonstrates the bounding box size distribution.

most of our data is collected adjacent to the intersection and the surrounding vehicles tend to have these four relative poses with respect to the data collection vehicle. Compared with car and truck categories, pedestrians have fewer annotations above 100 meters due to the sparse LiDAR measurements of the small objects in distant areas. Fig. 6 depicts the point clouds density distribution within bounding boxes and the bounding box size distribution. From the figure, we can see the diverse bounding box size distribution and each category will have a size distribution and could serve as a prior for object detection tasks. Additionally, the number of points within the box will decrease for longer radial distances while collaboratively viewing the objects can potentially increase the sensing observation and thus provide better visual cues.

## 4   Tasks

In this section, we first present the V2X cooperative 3D object detection task with derived datasets as per the collaboration mode and ego perspective in Sec. 4.1, and then we introduce the evaluation metrics in Sec. 4.2. Afterwords, in Sec. 4.3, we introduce various benchmark methods and fusion strategies supported in our dataset.

### 4.1   V2X Cooperative 3D object detection

**Scope:** The V2X Cooperative 3D object detection aims to leverage multi-agent's LiDAR point clouds to collaboratively perform 3D object detection. In our dataset, there are four collaborators including two infrastructure and two autonomous vehicles. Based on the V2X collaboration mode and ego agent type, we present four types of datasets: V2X-Real-VC, V2X-Real-IC, V2X-Real-V2V, and V2X-Real-I2I.

**V2X-Real-VC:** In V2X-Real-VC, the ego agent is fixed as the autonomous vehicle while the collaborators include both the infrastructure and vehicle. This setup is specifically designed for research of vehicle-centric V2X collaborations.
**V2X-Real-IC:** The infrastructure is chosen as the ego agent and the neighboring vehicles and infrastructure can collaborate with the ego infrastructure via sharing sensing observations. The final evaluation is conducted in the ego infrastructure side. This dataset is tailored for infrastructure-centric cooperative perception study.
**V2X-Real-V2V:** Only Vehicle-to-Vehicle collaboration is considered in this dataset. Additional scenarios collected in V2V corridors are used to supplement intersection scenarios with more diverse data distribution.
**V2X-Real-I2I:** In this dataset, Infrastructure-to-Infrastructure collaborations are studied, which is critical for intelligent transportation systems.

Contrary to existing works tailored for a single collaboration mode, the presented datasets have been meticulously crafted to cater to the diverse research interests of the broader V2X community.

### 4.2 Metrics

**Ground truth:** During the training stage, a random agent will be selected as the ego agent while during the inference, the ego is fixed as per dataset categories. For example, in the V2X-Real-VC, the ego is fixed as one of the data collection vehicles while in the V2X-Real-IC, the ego is infrastructure. When the bounding boxes are not consistent across agents, we choose the ego agent's annotation as the ground truth.
**Evaluation:** The evaluation is conducted in the range of -100m to 100m in x direction and -40m to 40m in y direction of the ego coordinate frame. We group the annotated 10 categories into three classes *i.e.* car, pedestrian and truck as per their bounding box size distribution. We calculate the Average Precision(AP) for each class based on the specified Intersection-over-Union (IoU) threshold, and a final mean Average Precision (mAP) is evaluated based on each category's AP:

$$mAP = \frac{1}{C} \sum_{i=1}^{C} AP_i \tag{1}$$

where $C$ is the class number. As our dataset contains a substantial number of pedestrians and large buses and the bounding box sizes can vary from less than 0.5 meters (pedestrian) to over 20 meters (bus), accurately predicting objects' poses and dimensions is particularly challenging. Thus similar as KITTI-360 dataset [15], we adopt smaller IoU thresholds (IoU=0.3, 0.5) for the AP calculation to faithfully reflect model's perception capability.

### 4.3 Benchmark methods

We provide benchmarks for all three fusion strategies in cooperative perception with SOTA fusion methods. Note that different from existing V2X 3D

Table 3: Benchmark results of SOTA cooperative perception methods for V2X-Real-VC, V2X-Real-IC, V2X-Real-V2V, and V2X-Real-I2I. All numbers in the table are represented in percentage format. For detection accuracy, Average Precision (AP) and mean Average Precision (mAP) for cars, pedestrians, and trucks are reported under IoU thresholds of 0.3 and 0.5.

| Dataset | Models | $AP_{car}$@IoU | | $AP_{ped.}$@IoU | | $AP_{truck}$@IoU | | mAP@IoU | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| | | 0.3 | 0.5 | 0.3 | 0.5 | 0.3 | 0.5 | 0.3 | 0.5 |
| V2X-Real-VC | No Fusion | 38.7 | 35.9 | 25.5 | 13.1 | 20.2 | 14.5 | 28.2 | 21.2 |
| | Late Fusion | 46.1 | 43.1 | 28.3 | 11.8 | 19.8 | 14.0 | 31.4 | 23.0 |
| | Early Fusion | 51.1 | 47.6 | 31.6 | 16.0 | 32.5 | 23.6 | 38.4 | 29.1 |
| | F-Cooper [2] | 57.3 | 54.2 | 30.0 | 14.1 | 27.0 | 21.2 | 38.1 | 29.8 |
| | AttFuse [35] | 62.6 | 59.4 | 32.2 | 15.5 | 32.6 | 26.6 | 42.5 | 33.8 |
| | V2X-ViT [34] | 62.7 | 60.3 | 36.7 | 18.6 | 35.1 | 28.3 | 44.8 | 35.8 |
| V2X-Real-IC | No Fusion | 43.3 | 35.6 | 25.6 | 12.7 | 21.2 | 20.2 | 30.0 | 22.8 |
| | Late Fusion | 65.2 | 60.4 | 33.4 | 16.0 | 30.8 | 24.5 | 43.1 | 33.6 |
| | Early Fusion | 64.2 | 59.8 | 32.3 | 15.0 | 33.3 | 27.4 | 43.3 | 34.1 |
| | F-Cooper [2] | 49.7 | 43.0 | 31.0 | 15.3 | 31.9 | 21.1 | 37.5 | 26.5 |
| | AttFuse [35] | 70.4 | 67.3 | 37.5 | 17.0 | 42.9 | 31.3 | 50.3 | 38.5 |
| | V2X-ViT [34] | 64.2 | 56.5 | 37.8 | 19.2 | 35.6 | 28.8 | 45.9 | 34.8 |
| V2X-Real-V2V | No Fusion | 41.7 | 39.4 | 26.9 | 14.4 | 21.6 | 13.7 | 30.1 | 22.5 |
| | Late Fusion | 47.4 | 44.4 | 29.2 | 14.9 | 18.7 | 9.1 | 31.8 | 22.8 |
| | Early Fusion | 54.0 | 49.8 | 31.9 | 17.1 | 28.6 | 18.6 | 38.1 | 28.5 |
| | F-Cooper [2] | 42.7 | 40.3 | 27.7 | 14.0 | 25.6 | 18.6 | 32.0 | 24.3 |
| | AttFuse [35] | 58.6 | 55.3 | 30.1 | 15.4 | 28.9 | 21.7 | 39.2 | 30.8 |
| | V2X-ViT [34] | 59.0 | 56.3 | 37.4 | 20.7 | 42.9 | 35.0 | 46.5 | 37.3 |
| V2X-Real-I2I | No Fusion | 48.6 | 40.0 | 30.9 | 15.8 | 23.5 | 22.4 | 34.3 | 26.1 |
| | Late Fusion | 67.2 | 63.3 | 41.1 | 23.1 | 48.4 | 39.1 | 52.2 | 41.8 |
| | Early Fusion | 60.9 | 57.2 | 41.2 | 21.9 | 38.5 | 30.8 | 46.9 | 36.6 |
| | F-Cooper [2] | 71.5 | 65.5 | 49.3 | 27.5 | 50.0 | 40.7 | 56.9 | 44.6 |
| | AttFuse [35] | 73.5 | 69.5 | 42.8 | 20.5 | 53.2 | 39.4 | 56.5 | 43.1 |
| | V2X-ViT [34] | 77.3 | 68.8 | 54.4 | 30.5 | 56.2 | 51.5 | 62.7 | 50.3 |

object detection benchmarks [36, 32] where only single vehicle categories are detected and evaluated, we provide comprehensive benchmarks for all the annotated classes *i.e.* pedestrian, car, and truck, providing the first open cooperative perception framework for multi-class cooperative 3D object detection. Here we list the benchmarked fusion strategies and methods:

- **No Fusion:** Only ego agent's LiDAR will be used to produce the 3D bounding boxes. It serves as the baseline.
- **Late Fusion:** Each agent first generates its own predictions and then shares the detection results with neighboring agents. The ego agent will receive these proposals and leverage non-maximum suppression (NMS) to produce consistent predictions.
- **Early Fusion:** The raw LiDAR point clouds will be shared for agents within the communication range. After receiving the neighboring agent's LiDAR point clouds, the ego agent will transform LiDAR point clouds into the ego agent's coordinate frame and aggregate all the LiDAR point clouds to produce a holistic view for visual reasoning.
- **Intermediate Fusion:** Each agent will first project its LiDAR to the ego agent's coordinate frame and leverage the LiDAR backbone to extract Birds' Eye View (BEV) features, which are then compressed and shared with the ego agent. The ego agent will leverage received multi-agent features to refine its own feature representations. In this work, we provide benchmarks for AttFuse [35], F-Cooper [2], and V2X-ViT [34].

## 5 Experiments

### 5.1 Implementation details

We split the dataset into train/val/test with 23379, 2770, and 6850 frames respectively. For the detection models, we adopt PointPillar [9] as the LiDAR backbone. The cooperative perception framework is built upon OpenCOOD [35]. To enable multi-class cooperative perception, we extend the original single-class head to multi-class heads following OpenPCDet's design [21]. All the models are trained with 80 epochs with a batch size of 2 and a learning rate of 0.001. Adam optimizer (decay weight $\lambda = 10^{-4}$) and multi-step learning rate scheduler ($\gamma = 0.1$, milestones=[10,50]) are employed for optimization.

### 5.2 Benchmark results

Tab. 3 summarizes the benchmark results of all six models for the four datasets. For all the datasets, cooperative methods outperform No Fusion baseline, demonstrating the great potential of V2X collaboration in enhancing the perception capabilities of individual agents. Among these cooperative methods, Intermediate Fusion methods typically yield more precise predictions compared to Late

and Early Fusion strategies. Notably, across all three object categories, the models generally perform best in the car category, whereas the performance is comparatively lower for the pedestrian and truck categories. We argue this is due to the fact that truck type has a diverse dimension distribution (Fig. 6c) and pedestrian bounding boxes usually only contain a few sparse LiDAR points, posing challenges for the learning. Generally, intermediate fusion ranks first among three fusion strategies, and late fusion methods outperform No Fusion baselines while early fusion shows superior performance than late fusion.

## 6    Conclusion

In this work, we present the V2X-Real, the first large-scale real-world dataset for V2X cooperative perception. It encompasses 33K LiDAR frames, 171K RGB images, and an unprecedented 1.2 million annotated 3D bounding boxes, making it one of the largest V2X datasets of its kind with dense traffic flows. To facilitate a wide range of V2X collaboration research, we provide four sub-datasets including V2X-Real-VC, V2X-Real-IC, V2X-Real-V2V, and V2X-Real-I2I. We also present a comprehensive multi-class benchmark framework for cooperative 3D object detection and from the best of our knowledge, this framework is also the first open multi-class cooperative 3D object detection framework in the literature, which enables future multi-class cooperative perception research like cooperative perception for vulnerable road users. We will release our data and benchmark codes and hope our open-source efforts can inspire more researchers to investigate this new yet important field.

**Future work and limitation:** The current dataset comprises only LiDAR and camera data, lacking support for radar, which is crucial for developing low-cost, scalable autonomous driving solutions. In the future, we plan to further extend our sensor suits to low-cost sensors such as radar and low-beam LiDAR for studying scalable low-cost V2X solutions.

## Acknowledgements

# Appendix

In this appendix, we first present additional visualizations for the V2X-Real dataset (Sec. A). Then, we detail the model configurations for LiDAR-based and Camera-based Cooperative Perception methods (Sec. B). Afterwords, additional benchmark results, and qualitative visualizations are shown in Sec. C.



(a) A V2V cityroad



(b) A V2V four-way intersection

Fig. 7: Sample scenarios from V2V corridors. *Top:* The aggregated LiDAR point clouds. *Bottom:* The front cameras of two data collection vehicles.

## A    Dataset Visualization

V2X-Real is collected in two types of scenarios *i.e.* V2X smart intersection and V2V corridors. Visualizations of the smart intersection are presented in Fig. 1-3 of the main paper. In this supplementary material, additional visualizations are provided for the V2V corridors (Fig. 7). Each figure displays three images: the top image showcases the aggregated LiDAR point clouds from both vehicles, while the bottom images present views from the associated front-view cameras. Further, Fig. 8 illustrates the stereo image pairs from the front camera. Notably, each vehicle is equipped with four stereo cameras, yielding eight images per frame, as illustrated in Fig. 9. These cameras, each offering a 120-degree horizontal field of view (FoV), are mounted in four perpendicular directions to achieve a complete 360-degree FoV around the vehicle. In Fig. 9, ground-truth 3D bounding boxes are projected onto the image plane via the intrinsics and extrinsics and each category *i.e.*, pedestrian, car, and truck, are colored differently.



(a) Left image                    (b) Right image

Fig. 8: Stereo Camera Image Pairs: (a) Image from the left camera. (b) Image from the right camera.

## B    Model Details

In this section, we detail the model configurations for the LiDAR-based cooperative perception (Sec. B.1) and Camera-based cooperative perception (Sec. B.2).

### B.1    LiDAR-based Cooperative Perception

**LiDAR backbone:** We adopt PointPillar [9] as the LiDAR backbone with a voxel size of 0.4 meters in the x and y directions. Following [34, 35, 26], each agent first project its LiDAR to the ego agent's coordinate frame. Then, the raw points are converted into pillar features and scattered into a 2D Bird's Eye View (BEV) feature map, which is shared with neighboring connected agents. In this work, the number of points per voxel is set to 32 while the maximum

(a) Front camera

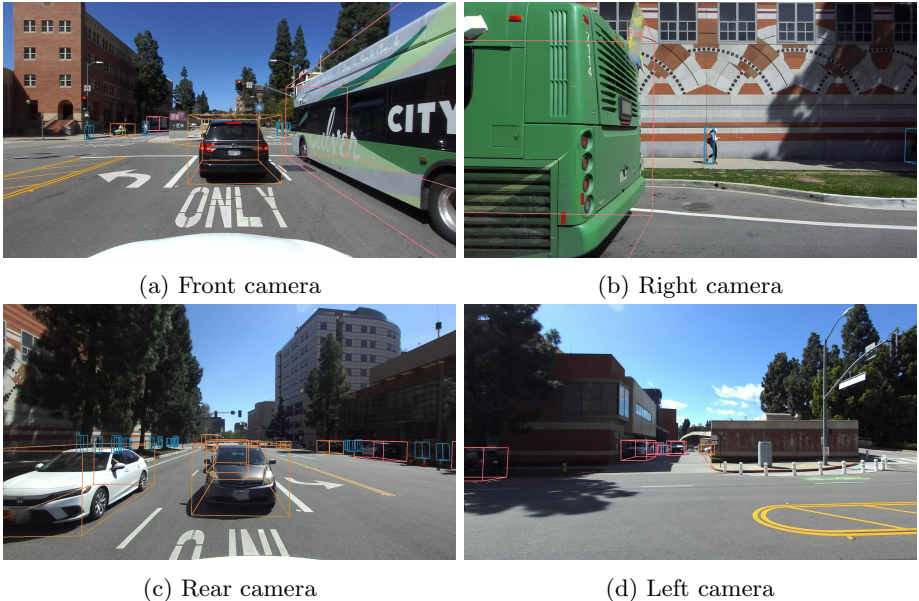(b) Right camera

(c) Rear camera

(d) Left camera

Fig. 9: Visualization of four stereo cameras of the vehicle. (a)-(d) plot the images from the front, right, rear, and left cameras respectively. 3D bounding boxes are projected onto a 2D image plane and are colored differently as per category. Best viewed in color with zoom for enhanced detail.

voxel number is 64000.

**Fusion method:** The ego agent receives all the collaborators' features and produces refined feature representation via fusing all the received feature maps and its own extracted features. We provide benchmarks for three intermediate fusion methods *i.e.*, F-Cooper [2], AttFuse [35], V2X-ViT [34]. F-Cooper leverages max-pooling to aggregate features. AttFuse adopts per-location attention for feature refinement. V2X-ViT employs multi-scale window attention and heterogeneous self-attention to jointly learn the inter-agent and intra-agent interactions.

**Head:** We adopt two 1x1 convolution layers for box regression and classification. The regression branch outputs $(x, y, z, w, h, l, \theta)$, denoting the position $(x, y, z)$, box dimensions $(w, h, l)$ and yaw angle $(\theta)$ with respect to the predefined anchor box of each category. For the classification branch, the head outputs the confidence score of each category *i.e.*, pedestrian, car, truck. To capture the dimension prior of various types of objects, we define a set of fixed-size anchor boxes for each category as per its box size distribution and the network only learns the displacement with respect to these predefined anchor boxes. In this way, the model can efficiently identify multi-scale objects (Fig. 10).

### B.2 Camera-based Cooperative Perception

For saving the computation, only the left camera from each stereo camera pair is used, leading to a total of 4 images (Fig. 9) per vehicle. To extract salient
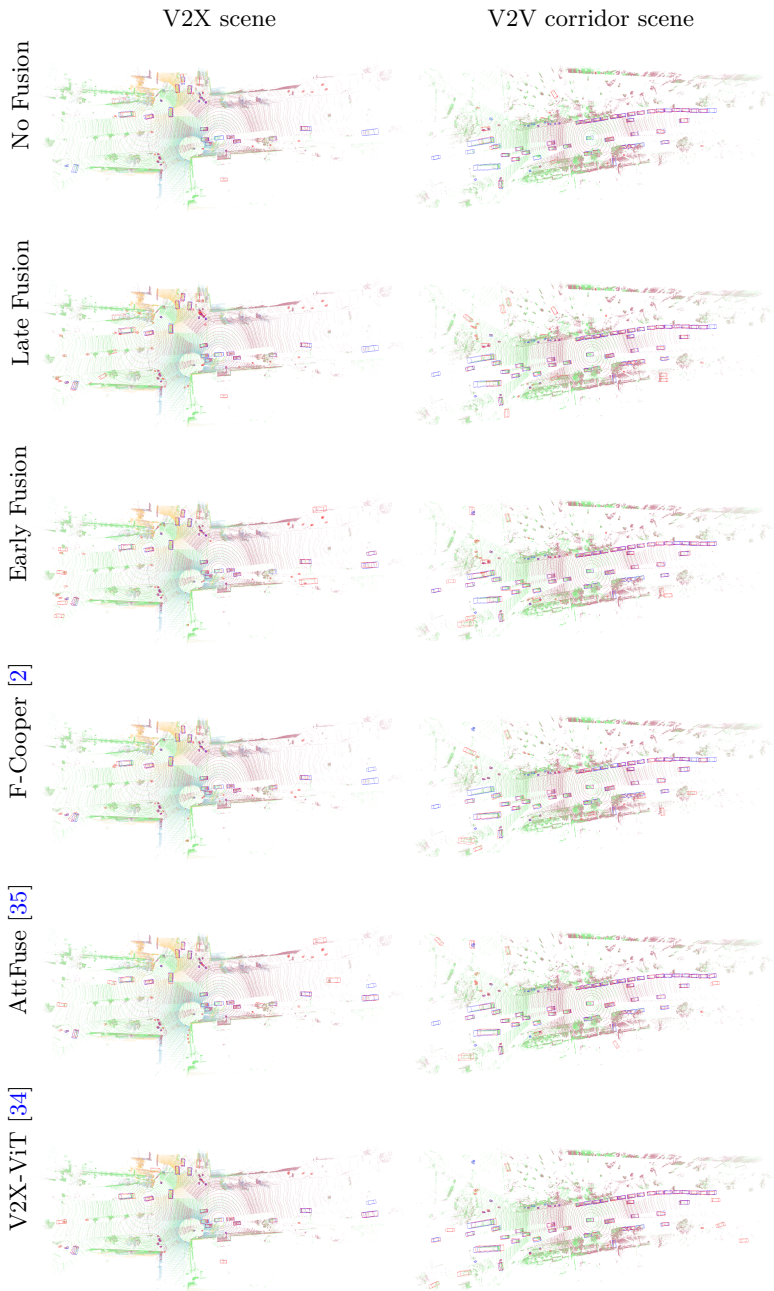
Fig. 10: **Qualitative visualization of cooperative perception in a V2X smart intersection scene and a V2V corridor scene.** Blue and red 3D bounding boxes correspond to the ground-truth and detection outputs, respectively.

features from 2D images, we employ an architecture similar to BEVFormer [14] with no temporal information. Following [26], for faster running speed, we adopt ResNet50 to extract 2D image features and leverage anchor-based head to generate the final detection outputs. In camera-based cooperative perception, we provide benchmarks for No Fusion and Late Fusion methods.

**No Fusion:** This approach utilizes solely the camera data from a single agent to generate 3D bounding boxes, serving as the baseline that operates without any collaborative input.

**Late Fusion:** In this method, each agent independently generates 3D bounding box proposals using its own camera data, along with associated confidence scores. These proposals and associated confidence scores are then transmitted to the ego agent, which employs non-maximum suppression (NMS) to produce consistent results from the received proposals.

The evaluation for the camera perception is conducted in the range of 50 meters in the x and y directions of the ego coordinate frame.

## C  More experimental results

### C.1  LiDAR-based Cooperative Perception

Fig. 10 offers qualitative visualizations for all the benchmarked methods, confirming the overarching trend observed in the main paper—fusion methods excel over the No Fusion baseline, with Intermediate Fusion methods showing superior precision. Notably, while Late Fusion successfully identifies more true positives compared with No Fusion, it also leads to a rise in false negative predictions. We argue this is due to the loss of context during the late fusion strategy where only the detection results are shared and fused. Consequently, the uncertainty inherent in collaborator messages may be amplified, resulting in an increased rate of false negatives. On the other hand, intermediate fusion methods produce much fewer false positives. We contend that this is due to the end-to-end multi-agent fusion, where the shared/fused features preserve the raw context information and thus can better help refine the ego agent's feature representation, leading to boosted performance.

### C.2  Camera-based Cooperative Perception

In this work, we provide benchmarks for vehicle-side multi-view camera-based cooperative perception and leave infrastructure-side camera solution for future works as the infrastructure camera has a fixed topology with roads, which could serve as a prior for detection task and thus require a dedicated new algorithm for efficient Infrastructure-to-Everything cooperative perception. Tab. 4 demonstrates the benchmark performance for camera-based cooperative perception

Table 4: Benchmark results of camera-based cooperative perception methods. AP is measured under IoU thresholds of 0.3 and 0.5.

| Dataset | Methods | Backbone | Collaboration | AP@0.3 | AP@0.5 |
|---|---|---|---|---|---|
| V2X-Real-VC | No Fusion | BEVFormer [14] | | 14.7 | 13.5 |
| | Late Fusion | BEVFormer [14] | ✓ | **16.9** | **15.0** |
| V2X-Real-V2V | No Fusion | BEVFormer [14] | | 7.3 | 6.8 |
| | Late Fusion | BEVFormer [14] | ✓ | **8.2** | **7.2** |

on V2X-Real-VC (excluding infrastructure collaborator) and V2X-Real-V2V. The results indicate that Late Fusion outperforms No Fusion baseline for both datasets. Moreover, both models perform higher in the V2X-Real-VC dataset than the V2X-Real-V2V, which contains V2V corridor scenarios. We argue this is due to the fact that the V2V corridors contain more diverse scenarios and roads with high slope changes, posing challenges for BEV-based camera solutions in accurately determining 3D poses. We encourage further exploration into these important yet under-explored areas by the research community.

# References

1. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuscenes: A multimodal dataset for autonomous driving. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11621–11631 (2020) 3, 4
2. Chen, Q., Ma, X., Tang, S., Guo, J., Yang, Q., Fu, S.: F-cooper: Feature based cooperative perception for autonomous vehicle edge computing system using 3d point clouds. In: Proceedings of the 4th ACM/IEEE Symposium on Edge Computing. pp. 88–100 (2019) 6, 12, 13, 17, 18
3. Dosovitskiy, A., Ros, G., Codevilla, F., Lopez, A., Koltun, V.: Carla: An open urban driving simulator. In: Conference on robot learning. pp. 1–16. PMLR (2017) 2, 4
4. Gao, L., Xia, X., Zheng, Z., Ma, J.: Gnss/imu/lidar fusion for vehicle localization in urban driving environments within a consensus framework. Mechanical Systems and Signal Processing **205**, 110862 (2023) 7
5. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: 2012 IEEE conference on computer vision and pattern recognition. pp. 3354–3361. IEEE (2012) 3, 4
6. Hao, R., Fan, S., Dai, Y., Zhang, Z., Li, C., Wang, Y., Yu, H., Yang, W., Yuan, J., Nie, Z.: Rcooper: A real-world large-scale dataset for roadside cooperative perception. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22347–22357 (2024) 4
7. Hu, Y., Fang, S., Lei, Z., Zhong, Y., Chen, S.: Where2comm: Communication-efficient collaborative perception via spatial confidence maps. Advances in neural information processing systems **35**, 4874–4886 (2022) 5
8. Huang, Y., Du, J., Yang, Z., Zhou, Z., Zhang, L., Chen, H.: A survey on trajectory-prediction methods for autonomous driving. IEEE Transactions on Intelligent Vehicles **7**(3), 652–674 (2022) 1

9. Lang, A.H., Vora, S., Caesar, H., Zhou, L., Yang, J., Beijbom, O.: Pointpillars: Fast encoders for object detection from point clouds. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12697–12705 (2019) 1, 13, 16

10. Li, E., Wang, S., Li, C., Li, D., Wu, X., Hao, Q.: Sustech points: A portable 3d point cloud interactive annotation platform system. In: 2020 IEEE Intelligent Vehicles Symposium (IV). pp. 1108–1115. IEEE (2020) 8

11. Li, Y., Zhao, S.Z., Xu, C., Tang, C., Li, C., Ding, M., Tomizuka, M., Zhan, W.: Pre-training on synthetic driving data for trajectory prediction (2023) 1

12. Li, Y., Ma, D., An, Z., Wang, Z., Zhong, Y., Chen, S., Feng, C.: V2x-sim: Multi-agent collaborative perception dataset and benchmark for autonomous driving. IEEE Robotics and Automation Letters **7**(4), 10914–10921 (2022) 2, 4

13. Li, Y., Ren, S., Wu, P., Chen, S., Feng, C., Zhang, W.: Learning distilled collaboration graph for multi-agent perception. Advances in Neural Information Processing Systems **34**, 29541–29552 (2021) 5

14. Li, Z., Wang, W., Li, H., Xie, E., Sima, C., Lu, T., Qiao, Y., Dai, J.: Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In: European conference on computer vision. pp. 1–18. Springer (2022) 19, 20

15. Liao, Y., Xie, J., Geiger, A.: Kitti-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(3), 3292–3310 (2022) 11

16. Lu, Y., Hu, Y., Zhong, Y., Wang, D., Chen, S., Wang, Y.: An extensible framework for open heterogeneous collaborative perception. arXiv preprint arXiv:2401.13964 (2024) 2

17. Lu, Y., Li, Q., Liu, B., Dianati, M., Feng, C., Chen, S., Wang, Y.: Robust collaborative 3d object detection in presence of pose errors. In: 2023 IEEE International Conference on Robotics and Automation (ICRA). pp. 4812–4818. IEEE (2023) 1

18. Rawashdeh, Z.Y., Wang, Z.: Collaborative automated driving: A machine learning-based method to enhance the accuracy of shared information. In: 2018 21st International Conference on Intelligent Transportation Systems (ITSC). pp. 3961–3966. IEEE (2018) 5

19. Su, S., Li, Y., He, S., Han, S., Feng, C., Ding, C., Miao, F.: Uncertainty quantification of collaborative detection for self-driving. In: 2023 IEEE International Conference on Robotics and Automation (ICRA). pp. 5588–5594. IEEE (2023) 5

20. Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B., et al.: Scalability in perception for autonomous driving: Waymo open dataset. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2446–2454 (2020) 3, 4

21. Team, O.D.: Openpcdet: An open-source toolbox for 3d object detection from point clouds. https://github.com/open-mmlab/OpenPCDet (2020) 13

22. Wang, T.H., Manivasagam, S., Liang, M., Yang, B., Zeng, W., Urtasun, R.: V2vnet: Vehicle-to-vehicle communication for joint perception and prediction. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16. pp. 605–621. Springer (2020) 2, 5, 6

23. Wei, S., Wei, Y., Hu, Y., Lu, Y., Zhong, Y., Chen, S., Zhang, Y.: Asynchrony-robust collaborative perception via bird's eye view flow. In: Advances in Neural Information Processing Systems (2023) 1

24. Wu, Z., Wang, Y., Ma, H., Li, Z., Qiu, H., Li, J.: Cmp: Cooperative motion prediction with multi-agent communication. arXiv preprint arXiv:2403.17916 (2024) 1

25. Xia, X., Meng, Z., Han, X., Li, H., Tsukiji, T., Xu, R., Zheng, Z., Ma, J.: An automated driving systems data acquisition and analytics platform. Transportation research part C: emerging technologies **151**, 104120 (2023) 7

26. Xiang, H., Xu, R., Ma, J.: Hm-vit: Hetero-modal vehicle-to-vehicle cooperative perception with vision transformer. arXiv preprint arXiv:2304.10628 (2023) 2, 6, 16, 19

27. Xiang, H., Xu, R., Xia, X., Zheng, Z., Zhou, B., Ma, J.: V2xp-asg: Generating adversarial scenes for vehicle-to-everything perception. In: 2023 IEEE International Conference on Robotics and Automation (ICRA). pp. 3584–3591. IEEE (2023) 1

28. Xu, C., Li, T., Tang, C., Sun, L., Keutzer, K., Tomizuka, M., Fathi, A., Zhan, W.: Pretram: Self-supervised pre-training via connecting trajectory and map. arXiv preprint arXiv:2204.10435 (2022) 1

29. Xu, R., Chen, W., Xiang, H., Xia, X., Liu, L., Ma, J.: Model-agnostic multi-agent perception framework. In: 2023 IEEE International Conference on Robotics and Automation (ICRA). pp. 1471–1478. IEEE (2023) 5

30. Xu, R., Guo, Y., Han, X., Xia, X., Xiang, H., Ma, J.: Opencda: an open cooperative driving automation framework integrated with co-simulation. In: 2021 IEEE International Intelligent Transportation Systems Conference (ITSC). pp. 1155–1162. IEEE (2021) 4

31. Xu, R., Tu, Z., Xiang, H., Shao, W., Zhou, B., Ma, J.: Cobevt: Cooperative bird's eye view semantic segmentation with sparse transformers. arXiv preprint arXiv:2207.02202 (2022) 6

32. Xu, R., Xia, X., Li, J., Li, H., Zhang, S., Tu, Z., Meng, Z., Xiang, H., Dong, X., Song, R., et al.: V2v4real: A real-world large-scale dataset for vehicle-to-vehicle cooperative perception. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13712–13722 (2023) 2, 4, 13

33. Xu, R., Xiang, H., Han, X., Xia, X., Meng, Z., Chen, C.J., Correa-Jullian, C., Ma, J.: The opencda open-source ecosystem for cooperative driving automation research. IEEE Transactions on Intelligent Vehicles (2023) 4

34. Xu, R., Xiang, H., Tu, Z., Xia, X., Yang, M.H., Ma, J.: V2x-vit: Vehicle-to-everything cooperative perception with vision transformer. In: European conference on computer vision. pp. 107–124. Springer (2022) 2, 4, 5, 6, 12, 13, 16, 17, 18

35. Xu, R., Xiang, H., Xia, X., Han, X., Li, J., Ma, J.: Opv2v: An open benchmark dataset and fusion pipeline for perception with vehicle-to-vehicle communication. In: 2022 International Conference on Robotics and Automation (ICRA). pp. 2583–2589. IEEE (2022) 2, 4, 5, 12, 13, 16, 17, 18

36. Yu, H., Luo, Y., Shu, M., Huo, Y., Yang, Z., Shi, Y., Guo, Z., Li, H., Hu, X., Yuan, J., et al.: Dair-v2x: A large-scale dataset for vehicle-infrastructure cooperative 3d object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21361–21370 (2022) 2, 4, 13

37. Zheng, Z., Han, X., Xia, X., Gao, L., Xiang, H., Ma, J.: Opencda-ros: Enabling seamless integration of simulation and real-world cooperative driving automation. IEEE Transactions on Intelligent Vehicles (2023) 4

38. Zhou, Y., Tuzel, O.: VoxelNet: End-to-End Learning for Point Cloud Based 3D Object Detection. pp. 4490–4499 (2018) 1

39. Zhou, Z., Yang, Z., Zhang, Y., Huang, Y., Chen, H., Yu, Z.: A comprehensive study of speed prediction in transportation system: From vehicle to traffic. Iscience **25**(3) (2022) 1