# H-V2X: A Large Scale Highway Dataset for BEV Perception

Chang Liu, Mingxu Zhu, and Cong Ma

ADLab, Tencent
{changeliu,mingxuzhu,codyma}@tencent.com

**Abstract.** Vehicle-to-everything (V2X) technology has become an area of interest in research due to the availability of roadside infrastructure perception datasets. However, these datasets primarily focus on urban intersections and lack data on highway scenarios. Additionally, the perception tasks in the datasets are mainly MONO 3D due to limited synchronized data across multiple sensors. To bridge this gap, we propose Highway-V2X (H-V2X), the first large-scale highway Bird's-Eye-View (BEV) perception dataset captured by sensors in the real world. The dataset covers over 100 kilometers of highway, with a diverse range of road and weather conditions. H-V2X consists of over 1.9 million fine-grained categorized samples in BEV space, captured by multiple synchronized cameras, with vector map provided. We performed joint 2D-3D calibrations to ensure correct projection and human labor was involved to ensure data quality. Furthermore, we propose three highly relevant tasks to the highway scenario: BEV detection, BEV tracking, and trajectory prediction. We conducted benchmarks for each task, and innovative methods incorporating vector map information were proposed. We hope that H-V2X and benchmark methods will facilitate highway BEV perception research direction. The dataset is available at https://pan.quark.cn/s/86d19da10d18

**Keywords:** V2X, Dataset, BEV Detection, Tracking, Prediction

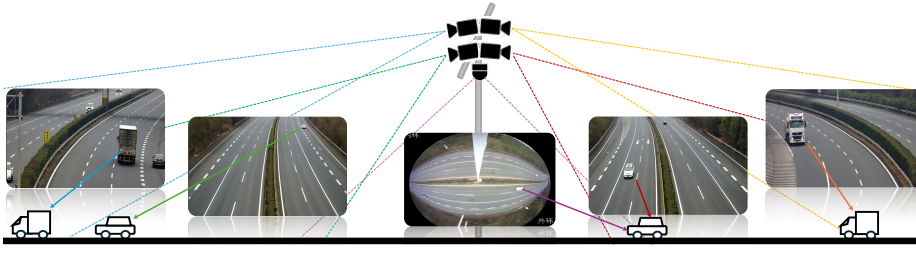

**Fig. 1:** We introduce H-V2X, the first large scale highway dataset captured from real-world senario and sensors. H-V2X provides synchroized sensor data, vector map, projection parameters and 3D ground truth, enabling end-to-end highway BEV perception.

## 1 Introduction

The highway is pivotal in transportation, but persistent safety concerns remain unresolved. While the development of intelligent vehicles presents a potential

and practical solution to address human-driver-related issues such as fatigue, distraction, and unsafe driving practices, the adoption of vehicle intelligence is still in its nascent stages. It is noteworthy that the majority of vehicles on the highway lack advanced driver assistance systems.

Conversely, there is substantial promise in the emerging research focus on roadside infrastructure, particularly within vehicle-to-everything (V2X) communication. V2X has the potential to significantly enhance driving safety by leveraging sensors installed along the roadside to perceive the surrounding environment of the ego-car and subsequently communicate with it through network connections such as 5G or Road-Side-Unit (RSU). There have been several works that focus on city roadside perception by constructing roadside infrastructure datasets (sec2.1), and successive works that focus on 3D perception given the data (sec2.2). However, when considering highway roadside perception, the existing datasets are not directly applicable for training neural networks due to differences in scenarios, sensors and tasks.

To address the disparities and advance the study of highway roadside perception, we are introducing Highway-V2X (H-V2X), a large-scale 3D perception dataset presented in Bird's Eye View (BEV) space. This dataset was captured in real world highway scenarios, with sensors strategically mounted on masts. Encompassing over 100 kilometers of highway, the dataset leverages cameras and radars, complemented by a local vector map converted from High-Definition Map (HDMap). In comparison to previous V2X datasets, H-V2X offers the following distinct characteristics:

1. Emphasis on Highway Scenarios: The dataset is tailored specifically for highway scenarios, showcasing a data distribution that significantly differs from urban datasets. As depicted in the accompanying figure, large vehicles, in particular, feature prominently within this dataset (sec3.3).

2. Ground Truth Construction in BEV Space: Our dataset facilitates end-to-end perception across multiple sensors by establishing ground truths in BEV space. Furthermore, the ground truth data is sequentially organized, enabling sequential learning.

3. Sensor Utilization: H-V2X employs radars and cameras, including fisheye cameras to address blind spots beneath the mast, as illustrated in the figure1. Notably, lidars are not included due to impractical installation on highway mast. Consequently, constructing 3D ground truth presents a significant challenge, a facet we will delve into extensively in subsequent sections.

4. Unified ID Tracking: The dataset emphasizes unified ID tracking across multiple sensors within the actual trajectories of vehicles, rendering multi-object tracking (MOT) an even more formidable task. In essence, vehicles are tracked throughout their entire lifespan.

To the best of our knowledge, H-V2X stands as the first large-scale dataset situated in a real-world highway scenario, incorporating BEV space attributes, sequential data, and multimodal sensing. We aim for this dataset to make a substantial contribution to the advancement of highway perception research. Leveraging the H-V2X dataset, we have also introduced three tasks accompanied

by ground truths and metrics to facilitate research in highway perception, along with novel bencmark methods.

To summarise, our contributions are outlined as follows:

1. We release the H-V2X dataset, the first large-scale highway roadside infrastructure perception dataset collected using real-world data. In addition, ground truths are constructed in BEV space across multiple sensors, making end-to-end BEV learning possible.

2. Based on H-V2X, we introduce three tasks across multiple sensors, i.e., BEV detection, BEV tracking, and trajectory prediction. We comprehensively provide time and space aligned original sensor data, along with ground truths constructed by algorithms and human labor.

3. We present benchmark methods for each task, where novel neural nets are proposed incorporating HDMap information, designed to address perception across multiple sensors in BEV space.

## 2 Related Works

### 2.1 Roadside Infrastructure Datasets

In this work, we focus on datasets that belong to the roadside rather than the vehicle side due to the significant variations in sensor perspectives and associated subtasks. For datasets relevant to autonomous vehicles from the vehicle side, we recommend referring to the autonomous driving datasets survey in [22, 39]. To enable autonomous driving vehicles to perceive long-range and blind-spot areas roadside perception (detailed in next section) and V2X datasets [6, 11, 14, 15, 18, 27, 35, 36, 38] have emerged as a new research direction. Among these works, roadside sensor data are provided in either simulated or real form, with scenarios focused on either urban or highway. A comprehensive comparison of these datasets can be found in table 1. Regarding simulated datasets, OPV2V [27] introduced a vehicle-to-vehicle simulated perception dataset featuring an average of 3 vehicles in the CARLA [7] simulator. Similarly, V2X-Sim [18] utilized CARLA and SUMO [16] to create a vehicle-to-RSU collaborative perception dataset. Additionally, Roadside-Opt [14] provided a roadside LiDARs dataset that focuses on sensor placement optimization, also using CARLA. These simulated datasets predominantly center on urban scenes, particularly intersections, with relatively limited representation of highway scenarios. On the other hand, roadside datasets captured by real sensors such as those proposed in [11, 35, 36, 38] use cameras or LiDARs in infrastructure-only or infrastructure-vehicle cooperative scenarios in urban areas. These datasets primarily support Mono3D detection tasks [11, 35, 36], as well as tracking and prediction tasks [38]. For highway scenarios, HighD [15] released a camera-based aerial perspective (top-down view) highway dataset using drones. A9-Dataset [6] released a 3km-long roadside Mono3D detection dataset using cameras and LiDARs. DAIR-V2X-V [36] released a highway Mono3D dataset but from the vehicle side. However, there is still a lack of a large-scale highway roadside dataset collected from real-world sensors, where vehicles pass across multiple roadside sensors, thus forming a BEV perception task. We believe that the H-V2X dataset can bridge this gap.

**Table 1:** Roadside Perception Dataset Comparison. C:Camera, L:LiDAR, R:Radar

| Dataset | Year | Senario | Sim Real | Num Samples | Num Classes | With Map | Provided Sensors | Sequential Trajectories | Supported Tasks | Range \| Coverage |
|---|---|---|---|---|---|---|---|---|---|---|
| WIBAM [11] | 2021 | Urban | Real | 33092 | 1 | X | C | X | Mono3D | 1 intersection |
| Rope3d [35] | 2022 | Urban | Real | 50009 | 12 | X | C | X | Mono3D | 200m |
| DAIR-V2X-C [36] | 2022 | Urban | Real | 12424 | 10 | X | C&L | X | Mono3D | 20km |
| DAIR-V2X-I [36] | 2022 | Urban | Real | 7058 | 10 | X | C&L | X | Mono3D | 20km |
| V2X-Seq [38] | 2023 | Urban | Real | 11275 | 9 | VectorMap | C&L | ✓ | Tracking 3D & Pred | 28 intersections |
| A9-I [46] | 2023 | Urban | Real | 4800 | 10 | X | C&L | ✓ | Mono3D | 1 intersection |
| INT2 [29] | 2023 | Urban | Real | 106.8M | 1 | VectorMap | C&L | ✓ | Trajectory Prediction | 16 intersections |
| HighD [15] | 2018 | Highway Drone | Real | 1.48M | 2 | X | C | X | 2DBbox | 420m |
| DAIR-V2X-V [36] | 2022 | Urban Highway | Real | 15627 | 10 | X | C&L | X | Mono3D | 20km |
| A9-Dataset [6] | 2022 | Highway Ramp | Real | 1098 | 9 | X | C&L | ✓ | Mono3D | 3km |
| OPV2V [27] | 2022 | Urban Rural | Sim | 11464 | 1 | X | C&L | X | BEV | 70 Scenes |
| V2X-Sim [18] | 2022 | Urban | Sim | 57200 | 1 | X | C&L | X | BEV & 3D Tracking | 100Scenes |
| Roadside-Opt [14] | 2023 | Urban | Sim | 37641 | 1 | X | L | X | BEV | 10 Scenes |
| H-V2X**(ours)** | 2024 | Highway | Real | 1.94 Million | 4 | VectorMap | C&R | ✓ | BEV Det MOT Tracking Trajectory Prediction | $>$ 100km |

## 2.2 Roadside Perception Methods

Roadside perception refers to the use of roadside sensors to provide collaborative perception information for autonomous driving vehicles. M3D-RPN [2], Kinematic3D [3], MonoDLE [21], MonoFlex [40] and Imvoxelnet [23] were reimplemented for roadside camera-based monocular 3D detection on dataset Rope3D [35] and DAIR-V2X [36] as baseline methods. PointPillars [17], MVX-Net [24], Second [28], PIXOR [30] and VoxelNet [43] were reimplemented for LiDAR-based monocular 3D detection baseline methods in DAIR-V2X-C/I [36] and OPV2V [27] dataset. Additionally, TNT [42] and HiVT [44] are utilized as trajectory prediction baseline methods in V2X-Seq [38]. Beyond these baseline methods, which were adopted from recent works on autonomous detection tasks, InfraDet3D [45] proposed a multi-modal 3D detector that fuses LiDAR point clouds in an early fusion manner and incorporates camera detections in a late fusion manner. TransIFFTransIFF [5] and Quest [8] proposed instance-level feature fusion frameworks based on transformers by fusing infrastructure and vehicle side inputs. FFNet [37] proposed a flow-based feature fusion network that transmits feature flow instead of static features. BEVHeight [34] and BEVHeight++ [32] tackled roadside monocular 3D detection by predicting categorical height distribution per pixel to regress the height of vehicles to the ground. MonoGAE [33] proposed a ground-aware embedding by integrating implicit roadside ground information with a pixel-level ground plane equation map encoder. Furthermore, there is a line of work that addresses roadside perception from the perspective of multi-agent cooperative perception by treating roadside input as one of the agents, as seen in Where2comm [12], SCOPE [31], CORE [26], etc. For a more comprehensive understanding of cooperative perception, we recommend referring to the V2X survey paper [20]. It is important to note that these roadside perception methods are limited in their ability to perform an across-sensors BEV perception task, as the most commonly used datasets (such as DAIR-V2X, Rope3D, etc.) do not provide synchronized sensor data across multiple sensors. Moreover, these methods currently lack the capability to handle fisheye cameras and incorporate HDMap information, which is essential for highway perception scenarios.

# 3 H-V2X Dataset

## 3.1 Setup

**Basic Information** The roadside infrastructure dataset presented in this paper is derived from real-world highways, with sensors strategically mounted on masts situated in the midst of the road. This network of masts spans over 100 kilometers, ensuring near-complete coverage of the highway with minimal blind spots (except for occlusions caused by bridges, signs, etc.). The dataset encompasses diverse highway scenarios, encompassing long straight roads, sharp curves, elevated bridges, ramps, and more, presenting a comprehensive portrayal of all road conditions typically encountered in highway settings. Each mast has multi-

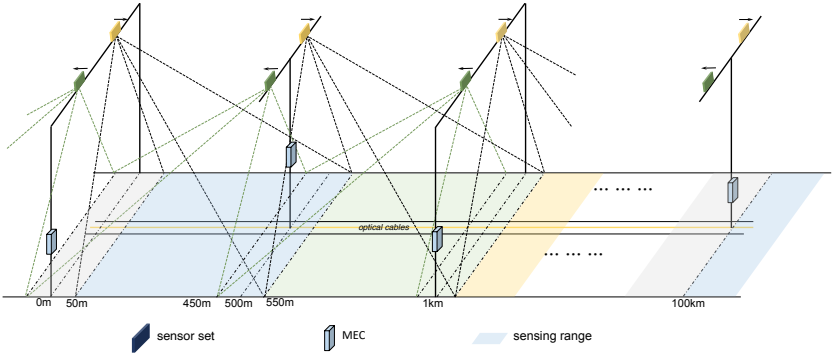


**Fig. 2:** Sensor Deployment Illustration. Long-range cameras, short-range cameras, fisheye cameras are mounted on the mast, covering over 1km, eight lanes of the road. Synchronized sensor data are collected, and shared observation area are ensured across sensors. Radars are also mounted on each side of the mast, and are omitted in the figure for simplification.

ple sensors, including two long-range (>800m) radars, two long-range cameras, two short-range cameras, and a fisheye camera. A mobile edge computer (MEC)is installed on each mast for data collection. All hardware is synchronized using the NTP service deployed in the local cloud machine, and the time difference between sensors within the same mast is less than 50ms. A topology of the setup can be referred to as Fig. 2, with sensor specifications in table 2.

**Table 2:** Sensors Specifications and Configurations

| Sensor | Focus | Frequency | Resolution | Format | FOV | BandWidth |
|---|---|---|---|---|---|---|
| Short-range Camera | 12mm | 10hz | 1920x1080 | JPEG | 49.78 | 4Mbps |
| Long-range Camera | 70mm | 10hz | 1920x1080 | JPEG | 9.1 | 4Mbps |
| Fisheye Camera | 1.27mm | 10hz | 1280x1280 | JPEG | 180 | 4Mbps |

**Coordinates and Calibration** There are three coordinate systems in the H-V2X dataset: image coordinate, camera coordinate, and world coordinate. The primary purpose of the H-V2X dataset is to project objects from the image coordinate system to the world coordinate system, which is a local vector map converted from High-Definition Map (HDMap). We propose two calibration methods: a human calibration tool and an automated calibration algorithm.

- **Calibration Tool**: Prior to installation on the masts, the intrinsics of the cameras undergo calibration using a chessboard calibration algorithm [41]. The calibration tool incorporates the intrinsics, vector map, and the camera video streams. Subsequently, a human operator can fine-tune the extrinsic parameters (x, y, z, yaw, pitch, roll) to ensure that the projection of the vector map aligns seamlessly with the visual lanes. A visual illustration of the calibration tool is provided in Fig .3. This process ensures that all cameras are calibrated uniformly to the same vector map, thereby guaranteeing cohesive 2D to 3D projection.
- **Calibration Algorithm**: To address the concern of sensor drifting, we present an automatic calibration algorithm. This algorithm leverages a segmentation model to extract visual lanes and optimizes the 3D matching score between vector map lanes and visual lanes using a bundle adjustment algorithm. The calibration algorithm runs on a daily basis to maintain the timely accuracy of the calibration. Furthermore, the algorithm's effectiveness is rigorously authenticated by human operators.

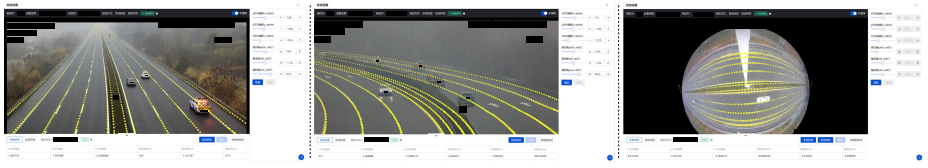

**Fig. 3:** Calibration Tool Examples. Images from left to right are the calibration interfaces for short-range, long-range, and fisheye cameras, respectively. Yellow dots are the projection of the vector map to the image plane. Non-English characters are not relevant to the core contents of the paper and can be ignored. Sensitive information is masked out during calibration and GT generation for safety concerns.

### 3.2 Data Acquisition & GT Generation

**Collection** Once the sensors and vector map are accurately calibrated, we proceed to decode the camera streams and capture synchronized frames from the five cameras (2 long-range cameras, 2 short-range cameras, and 1 fisheye camera) at a rate of 10 Hz. These frames, along with the current calibration parameters, are stored using ros bag. To ensure diversity, meticulous human efforts were invested in collecting over 1.94 million frames in total, encompassing the following aspects:

- **Scenes**. The dataset comprehensively covers nearly all highway scenes, including long straight roads, curved roads, interchanges, ramps, and more.
- **Weather**. The collected data encapsulates a wide spectrum of weather and time conditions, ranging from daytime and nighttime to varied atmospheric phenomena such as rainy, foggy, dusty, dawn, cloudy, etc.
- **Traffic**. Traffic conditions spans normal traffic flow, congested traffic flow, and instances of sparse traffic.

Any invalid frames arising from network fluctuations are meticulously removed (all synchronized five frames are deleted if any one of them is deemed invalid) to uphold the high quality of the data.

**GT Generation** In the context of roadside infrastructure 3D perception, the ground truth is typically annotated on the LiDAR point cloud, as in [35, 36]. However, in highway scenarios, the mounting of LiDARs on masts is impractical due to high cost, maintenance, durability, sensing range issues. As a viable alternative, we propose a novel BEV ground truth generation pipeline tailored for the camera-only highway scenario, as illustrated in Fig. 4.

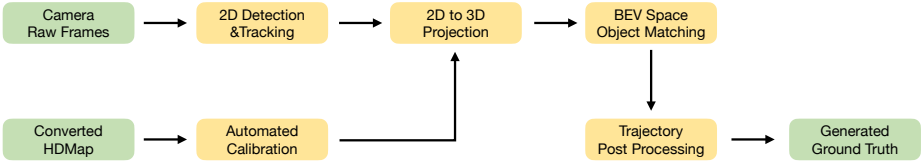


**Fig. 4:** Ground Truth Generation Pipeline.

- **2D Detection and Tracking**. We chose the bounding box as the object representation and train a cloud foundation model using 122,445 annotated data. We incorporate semi-supervised learning to utilize more data, resulting in 370,000 frames in total, yielding a mean average precision of 0.95@IoU=0.5, which provides solid results for subsequent modules. After obtaining the single frame object detection bounding box, a 2D multiple object tracking algorithm is added to smooth the bounding box size and provide a tracking ID for each object in the image plane.
- **2D to 3D Projection**. Bboxes in the image plane are subsequently projected to the vector map coordinate system. Outliers resulting from false detections are filtered during projection by leveraging road boundary information from the vector map.
- **Object Matching**. For objects that travel across sensors, a minimum Euclidean distance Hungarian matching strategy is introduced to merge the same object observed by different cameras.
- **Post Processing**. An Extended Kalman Filter motion model for each object in BEV space to further smooth object trajectories. The tracking IDs for each object in BEV space are also generated in the post-processing module.

Once the proposed GT generation pipeline is processed, human operators must further validate the data quality. They need to remove the samples that contain ghost trajectories (trajectories in the BEV space that do not match any object in the image frame) and trajectory breaks (trajectories that break abnormally or the same trajectory but tracking ID changes). Then, the remaining samples are labeled with four classes: sedan, van, bus, truck. These class labels are human-labeled during the 2D detection dataset construction phase, detected via 2D detectors, and passed through the pipeline. Velocity and heading angle are provided by radar observations. Finally, 17.56 million samples are generated.

### 3.3 Statistics Analysis

In total, we provide over 1.9 million samples and 17.6 million objects, with detailed statistical plots depicted in Fig 5. Figure (a) illustrates the object distribution based on the distance from the mast. Figure (b) showcases the category

distribution, highlighting sedans and trucks as dominant categories. Figure (c) displays the velocity distribution, indicating that most vehicles travel at reasonably high speeds. Figure (d) presents the average velocity per category. Figure (e) exhibits the track length of vehicles, where tracks longer than 400 meters constitute 78.4% of the dataset. Lastly, Figure (f) illustrates the average track length for each category.

In essence, these statistics reveal the distinctive data distribution characteristics of highway scenarios in H-V2X. The dataset differs significantly from urban scene data, underscoring its value in highway perception research.

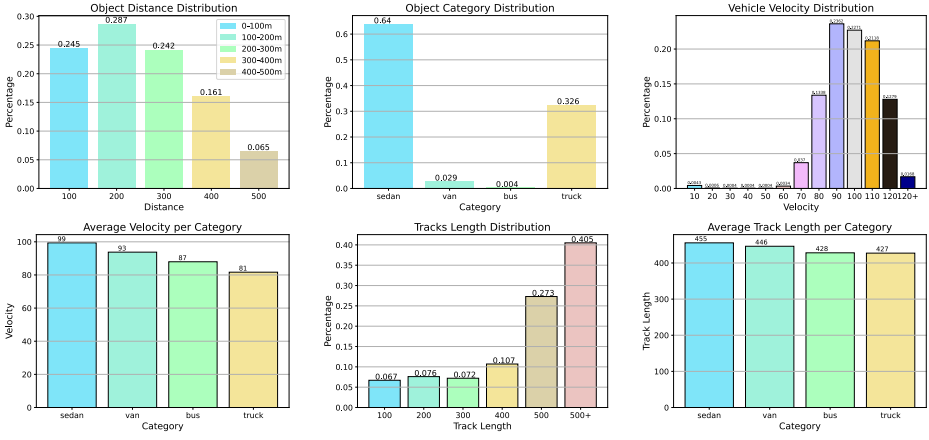


**Fig. 5:** H-V2X Dataset Statistics Analysis.

# 4 Tasks

## 4.1 Task 1: BEV Detection (H-V2X-Det)

**Task Description** In a manner akin to the task definition in autonomous driving BEV perception [13], BEV detection in the H-V2X dataset aims to pinpoint objects in BEV space by extracting positions, angles, and velocities from images. However, this task exhibits distinctive characteristics as outlined below:

• **Extended detection range**: The coverage typically spans approximately ±500 meters around the local principal point, significantly surpassing the detection range in traditional autonomous driving tasks (e.g., ±60m in [13], ±75m in [19]). Consequently, this task introduces challenges related to long-range detection in BEV settings, thereby influencing the design and integration of grid maps within the BEV algorithm.

• **Diverse camera types**: In highway scenarios, where comprehensive sensor coverage and an expanded field of view are imperative, various camera types are deployed, including long-focus, short-focus, and especially fisheye cameras.

• **Integration of HDMap**: The High-Definition Map (HDMap) serves as a critical input for the task, with lanes and boundaries depicted as line dots accompanied by appropriate labels, further enhancing the contextual information available for analysis.

**Formulation** Given multiple sensor inputs at each timestamp $t$, BEV detection tries to find a fitting neural net $F$ which calculates the 3D information of objects in all image views, forming a ten-tuple:

$$(cls, conf, x, y, z, w, l, h, yaw, color) = F((I_0, I_1, I_2, I_3, I_4), M), \quad (1)$$

where:

- $I_i$ represents each image of the five cameras in JPEG BGR format. Resolutions of long-range and short-range camera images are 1920x1080x3 and 1080x1080x3 for fisheye cameras.
- $M$ represents the HDMap, which provides static road information by lines of 3D points.
- $cls$, $conf$ represent class label, confidence respectively.
- $x, y, z, w, l, h$ represent 3D location in BEV space.
- $yaw$ represents the heading angle of the object, defined by moving direction.
- $color$ represents the object's color if the object belongs to the vehicle class.

**Evaluation & Metrics** We choose AP and FPS as BEV detection metrics.

**Average Precision (AP)**. We choose 40 recall positions to eval detection performance, i.e., $\text{AP}|_{R40}$ :

$$AP|_{R40} = \frac{1}{|R|} \sum_{r \in R} \max_{\tilde{r}:\tilde{r} \geq r} \rho(\tilde{r}) \quad (2)$$

where $\rho(\tilde{r})$ is the precision at a certain recall theshold r $\in \{1/40, 2/40, 3/40, ..., 1\}$.

**Frame Per Second(FPS)**. We use the FPS metric to measure the real-time performance of the neural net, i.e.,$FPS = \frac{1}{t(s)}$, where $t(s)$ is the cost time required to detect a single frame input given by specific GFLOPS.

### 4.2 Task 2: OneID MOT (H-V2X-Trk)

**Task Description** The MOT task is centered around establishing and maintaining a unique ID for each relevant traffic participant. There are two primary paradigms in MOT based on whether detection is performed separately:

- **Tracking-by-detection**: This method follows a two-stage process. Initially, a detection algorithm locates objects in each frame. Subsequently, a tracking algorithm utilizes the detection results from consecutive frames to associate new objects with historical trajectories, determining optimal matches and assigning correct IDs to newly detected objects.
- **Joint detection and tracking**: In this integrated approach, detection and tracking occur simultaneously through an end-to-end algorithm, often a specialized deep neural network. This network not only learns detection features but also captures embedding or re-identification (reid) features to differentiate between individual objects. Consequently, the network can simultaneously determine object positions and IDs during each inference step on successive image pairs.

**Formulation** The input consists of sequential images, while the output encompasses comprehensive 3D object details, including position and attributes, with a unique ID assigned to each object. The trajectory of each object can be extracted and represented as $S = \{o_1, o_2, o_3, ..., o_T\}$, where $T$ denotes the timeframe. Subsequent applications can benefit from velocity estimation, orientation optimization, and post-processing techniques based on historical trajectories.

**Evaluation & Metrics** In line with the recent trend [4,25,38], we adopt widely accepted evaluation metrics MOTA (Multiple Object Tracking Accuracy) and MOTP (Multiple Object Tracking Precision) to evaluate the performance of tracking methods. The definitions are as follows:

$$MOTA = 1 - \frac{\sum_t (FN_t + FP_t + IDS_t)}{\sum_t GT_t}, \tag{3}$$

where:

- $FN_t$ represents the number of false negatives at the time (t),
- $FP_t$ represents the number of false positives at the time (t),
- $IDS_t$ represents the number of identity switches (mismatches in object associations) at the time (t),
- $GT_t$ represents the number of ground truth objects at a time (t).

$$MOTP = \frac{\sum_{i,j} d_{ij}}{\sum_i C_i}, \tag{4}$$

where:

- $d_{ij}$ represents the Euclidean distance between the center of the predicted bounding box and the center of the corresponding ground truth bounding box for object $i$ at frame $j$,
- $C_i$ represents the number of correct associations for object $i$.

In addition, we employ IDS (ID Switch) as a standalone metric to assess the tracking algorithm's proficiency in preserving target identity consistency. The IDS metric is calculated by dividing the total number of ID switches by the overall count of ground-truth objects.

### 4.3 Task 3: Trajectory Prediction (H-V2X-Prediction)

**Task Description** The trajectory prediction task aims to predict the continuous trajectories of objects over a future period based on the observed history trajectories.

**Formulation** Given the history trajectory of one object, the trajectory prediction task tries to find the optimal future trajectory of the object by modeling a sequential forecasting net $F$:

$$p_{t+1}, p_{t+2}, p_{t+3}, ..., p_{t+T} = F(p_1, p_2, p_3, ..., p_t, M) \tag{5}$$

where:

- $p_t$ denotes the state of the object at timestamp $t$, $T$ represents the number of timestamps to predict, $M$ denotes HDMap, same as task 1.
- The state of the object is defined by the output of task 2, i.e., 3D information with tracking ID.

**Evaluation & Metrics** We choose Average Displacement Error (ADE) and Final Displacement Error (FDE) as trajectory prediction metrics.

**ADE**. Measuring the average error between the prediction position and the actual position. Defined as:

$$ADE = \sum 1/n \sqrt{(x_{p_i} - x_{g_i})^2 + (y_{p_i} - y_{g_i})^2} \tag{6}$$

**FDE**. Measuring the final error between the prediction final position and the actual final position. Defined as:

$$FDE = \sqrt{(x_{p_f} - x_{g_f})^2 + (y_{p_f} - y_{g_f})^2} \tag{7}$$

where n the prediction sequence length, $(x_{p_i}, y_{p_i}), (x_{g_i}, y_{g_i})$ is the object prediction position and actual position at time i. $(x_{p_f}, y_{p_f}), (x_{g_f}, y_{g_f})$ is the object prediction final position and actual final position.

## 5 Benchmarks

### 5.1 Track1: BEV Detection

**Baselines** We propose two baseline methods for BEV detection: late and early fusion. Both use 8:1:1 train/validation/test split.

- **Baseline1 Late Fusion.** Late fusion follows the same pipeline as GT generation in section 3.2. However, due to limited computing resources on the 32TOPS NVIDIA Xavier, it uses a lightweight YOLOs [9] detection model instead of a cloud-side foundation model.
- **Baseline2 Early Fusion.** We propose H-BEV (Highway BEV) model, a BEV detection neural net incorporating fisheye camera and vector map. The model is built on BEVDet [13], learns depth distribution and constructs frustum point cloud, illustrated in Fig. 6. To create the projection coordinates of each camera's pixel points in the BEV perspective, we used 3D lane information for interpolation. We generated the frustum point cloud for each camera through a table lookup and added a attribute branch to predict vehicle attribute information.

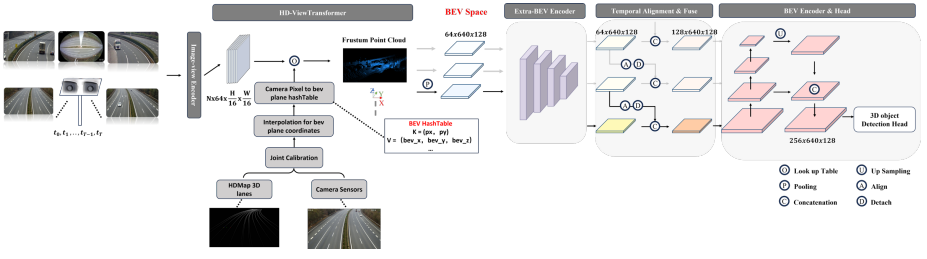


**Fig. 6:** H-BEV Model Architecture.

**Analysis**

- **Late Fusion vs Early Fusion.** Early fusion outperforms Late Fusion in terms of detection performance, but Late Fusion requires less time. Late Fusion is more sensitive to calibration parameters and performs poorly in detecting truncated and occluded objects, whereas early fusion is more robust.
- **HDMap information is important.** H-BEV streamlines the projection process for fisheye and pinhole cameras by incorporating HDMap data. Unlike BevDet [13], H-BEV eliminates the need to learn a multitude of discrete depth values for each image pixel. Instead, it derives a single depth value

from HDMap information, resulting in a drastic reduction in the image feature map size by 500 times. This optimization significantly boosts the speed and efficiency of the early fusion model.

**Table 3:** Evaluation results for the baseline methods. The H-BEV model is converted to TensorRT engine. The resolution of all image inputs is resized to 720 * 1280.

| Method | Backbone | $\text{mAP}_{3D\|R40}\uparrow$ | $\text{mAP}_{bev\|R40}\uparrow$ | FPS↑ |
|---|---|---|---|---|
| Late Fusion | - | 31.75 | 33.24 | 10 |
| Early Fusion (H-BEV) | ResNet-18 | **35.49** | **38.40** | **6** |

### 5.2 Track2: One-ID MOT

**Baseline** In this section, we present a tracking-by-detection MOT algorithm as the baseline approach. We utilize BEV detection results from multiple cameras to trace the trajectory of each object within the perception region, illustrated in Fig. 7. We adapt a variant of the SORT algorithm [1] to establish associations between objects across successive frames. This leads to the assignment of consistent IDs to individual objects, resulting in continuous object trajectories.

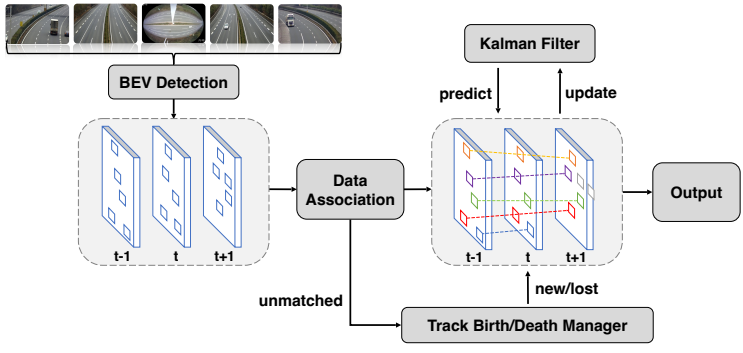


**Fig. 7:** One-ID MOT Framework: a baseline tracking-by-detection procedure

Each detected traffic object at time $t-1$ is denoted as $o^i_{t-1}$, where $i$ represents the object index. The set of objects in frame $t-1$ is represented as $I_{t-1}$, with $o^i_{t-1} \in I_{t-1}$. At each timestamp $t$, the primary objective of the tracking process is to link the objects from the previous frame ($I_{t-1}$) and the current frame ($I_t$), assigning accurate IDs to new detections. To achieve this, we compute similarities between objects in $I_{t-1}$ and $I_t$ based on their feature descriptors, facilitating the establishment of continuous object trajectories through optimal matches.

For the baseline implementation, we adopt the 3D position as the object descriptor, measuring object similarities using Euclidean distances between 3D points. Object matching between frames is accomplished using the Hungarian algorithm. Given the fast movement of most highway objects, an effective motion estimation technique is essential to compensate for positional changes between consecutive frames. Here, we employ the Kalman filter algorithm, utilizing the local coordinates $(x, y, z)$ of 3D points as measurements and incorporating

first-order differences (velocity components $(vx, vy, vz)$) into the state transition matrix.

**Analysis** The evaluation results of the baseline MOT method can be found in Table 4. In highway environments, conducting multi-object tracking on 3D points presents various challenges such as high-speed motion, occlusions, missing or false detections, positional ambiguity among nearby objects, and more. Common features in highway scenes like overpasses and oversized vehicles often lead to scenarios where smaller cars remain completely occluded for extended periods. Addressing these complexities requires the integration of effective prediction strategies and motion constraints. Additionally, when multiple vehicles are in close proximity and moving concurrently, relying solely on simple Euclidean distance matching between points can result in confusion and inaccuracies.

**Table 4:** Evaluation results for One-ID MOT algorithm.

| **Metrics** | MOTA↑ | MOTP↑ | IDS↓ |
|---|---|---|---|
| **Baseline** | 0.85 | 0.95 | 0.05 |

### 5.3 Track3: Trajectory Prediction

**Baselines** We first build trajectory prediction models based on SocialGAN [10] and further propose an HD-GAN network by incorporating HDMap.

- **Basline1 Vanilla SocialGAN**: During training, we augment trajectory data by adding random noise to observed positions (0.5 prob. for [-2m, 2m]) and applying Gaussian noise to observed directions ([-0.785 rad, 0.785 rad]). We train the model for 500000 iterations, leaving other settings unchanged.
- **Baseline2 HD-GAN**: We propose HD-GAN, a network that uses global map information to extract the global positional feature of an object. The HDMap Layer takes in lane information and calculates the object's global position via interpolation, normalizing it to [0,1] in the HDMap Normalization Layer. The HDMap Encoder Layer encodes the object's global map positional features, which are then combined with the local positional features of the observed trajectory. The generator and discriminator network remain unchanged, and the training strategy is the same as baseline 1.

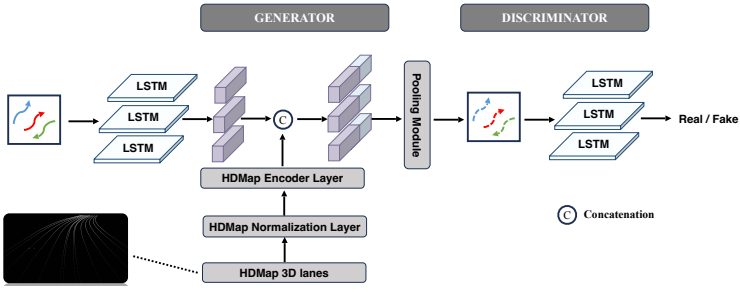


**Fig. 8:** HD-GAN Model Architecture.

**Analysis**

- **Global Map Information is useful.** In comparison to the re-implemented vanilla SocialGAN, HD-GAN demonstrates superior performance with lower ADE and FDE metrics. These results indicate that as the prediction horizon extends, the utilization of global map information becomes increasingly valuable. Additionally, as illustrated in Fig. 9 (a) vs (b), the vanilla SocialGAN struggles to generate predictions aligned with the road map under identical observations. These findings underscore the efficacy of incorporating global map data in HD-GAN for enhancing trajectory prediction accuracy.
- **Benefits of Data Augmentation.** As depicted in Fig. 9 (c) vs (d), the absence of data augmentation leads to increased prediction errors when the observed trajectory exhibits jitter. Conversely, models employing our data augmentation techniques showcase enhanced resilience to observed trajectory variations, as evidenced by their improved anti-interference capabilities compared to models without such augmentation methods.

**Table 5:** Evaluation results for different baselines on val dataset. The metrics respectively means the predicted trajectory lengths is 1s-5s.

| Method | ADE↓ | FDE↓ |
|---|---|---|
| Vanilla SocialGAN [10] | 0.432 / 0.894 / 1.430 / 2.018 / 2.683 | 0.830 / 1.849 / 3.031 / 4.382 / 5.833 |
| HD-GAN | **0.427** / **0.843** / **1.291** / **1.847** / **2.457** | **0.827** / **1.696** / **2.657** / **3.879** / **5.188** |

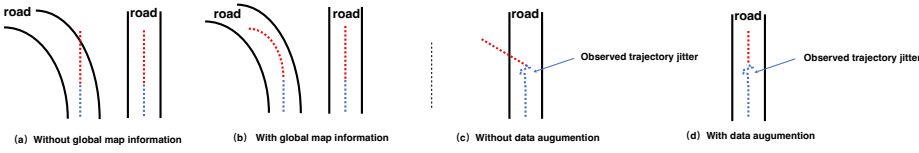


**Fig. 9:** Trajectory prediction results with and without global map information (a, b), with and without data augmentation (c, d). The blue points are the observed trajectory inputs and the red points are the prediction trajectory outputs.

## 6 Conclusion

This paper introduces the H-V2X dataset, the first large-scale highway dataset captured by real-world sensors. H-V2X offers synchronized data from multiple sensors with ground truth annotations, accompanied by a vector map, enabling end-to-end highway BEV perception. The tasks of BEV detection, One-ID MOT, and trajectory prediction are outlined, alongside benchmark methodologies for each task. Novel approaches to highway BEV detection and trajectory prediction, integrating the vector map, are proposed and evaluated through both quantitative and qualitative experiments to validate their efficacy. In our upcoming work, we aim to release a highway traffic event task focusing on rare occurrences such as illegal parking, emergency stops, accidents, and more, which will be systematically collected and annotated.

## References


1. Bewley, A., Ge, Z., Ott, L., Ramos, F., Upcroft, B.: Simple online and realtime tracking. In: 2016 IEEE international conference on image processing (ICIP). pp. 3464–3468. IEEE (2016)
2. Brazil, G., Liu, X.: M3d-rpn: Monocular 3d region proposal network for object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9287–9296 (2019)
3. Brazil, G., Pons-Moll, G., Liu, X., Schiele, B.: Kinematic 3d object detection in monocular video. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIII 16. pp. 135–152. Springer (2020)
4. 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)
5. Chen, Z., Shi, Y., Jia, J.: Transiff: An instance-level feature fusion framework for vehicle-infrastructure cooperative 3d detection with transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 18205–18214 (2023)
6. Creß, C., Zimmer, W., Strand, L., Fortkord, M., Dai, S., Lakshminarasimhan, V., Knoll, A.: A9-dataset: Multi-sensor infrastructure-based dataset for mobility research. In: 2022 IEEE Intelligent Vehicles Symposium (IV). pp. 965–970. IEEE (2022)
7. 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)
8. Fan, S., Yu, H., Yang, W., Yuan, J., Nie, Z.: Quest: Query stream for vehicle-infrastructure cooperative perception. arXiv preprint arXiv:2308.01804 (2023)
9. Ge, Z., Liu, S., Wang, F., Li, Z., Sun, J.: Yolox: Exceeding yolo series in 2021. arXiv preprint arXiv:2107.08430 (2021)
10. Gupta, A., Johnson, J., Fei-Fei, L., Savarese, S., Alahi, A.: Social gan: Socially acceptable trajectories with generative adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2255–2264 (2018)
11. Howe, M., Reid, I., Mackenzie, J.: Weakly supervised training of monocular 3d object detectors using wide baseline multi-view traffic camera data. arXiv preprint arXiv:2110.10966 (2021)
12. 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)
13. Huang, J., Huang, G., Zhu, Z., Ye, Y., Du, D.: Bevdet: High-performance multi-camera 3d object detection in bird-eye-view. arXiv preprint arXiv:2112.11790 (2021)
14. Jiang, W., Xiang, H., Cai, X., Xu, R., Ma, J., Li, Y., Lee, G.H., Liu, S.: Optimizing the placement of roadside lidars for autonomous driving. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 18381–18390 (2023)
15. Krajewski, R., Bock, J., Kloeker, L., Eckstein, L.: The highd dataset: A drone dataset of naturalistic vehicle trajectories on german highways for validation of highly automated driving systems. In: 2018 21st international conference on intelligent transportation systems (ITSC). pp. 2118–2125. IEEE (2018)

16. Krajzewicz, D., Erdmann, J., Behrisch, M., Bieker, L.: Recent development and applications of sumo-simulation of urban mobility. International journal on advances in systems and measurements **5**(3&4) (2012)
17. 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)
18. 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)
19. 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)
20. Liu, S., Gao, C., Chen, Y., Peng, X., Kong, X., Wang, K., Xu, R., Jiang, W., Xiang, H., Ma, J., et al.: Towards vehicle-to-everything autonomous driving: A survey on collaborative perception. arXiv preprint arXiv:2308.16714 (2023)
21. Ma, X., Zhang, Y., Xu, D., Zhou, D., Yi, S., Li, H., Ouyang, W.: Delving into localization errors for monocular 3d object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4721–4730 (2021)
22. Qian, R., Lai, X., Li, X.: 3d object detection for autonomous driving: A survey. Pattern Recognition **130**, 108796 (2022)
23. Rukhovich, D., Vorontsova, A., Konushin, A.: 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. pp. 2397–2406 (2022)
24. Sindagi, V.A., Zhou, Y., Tuzel, O.: Mvx-net: Multimodal voxelnet for 3d object detection. In: 2019 International Conference on Robotics and Automation (ICRA). pp. 7276–7282. IEEE (2019)
25. 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)
26. Wang, B., Zhang, L., Wang, Z., Zhao, Y., Zhou, T.: Core: Cooperative reconstruction for multi-agent perception. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8710–8720 (2023)
27. 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)
28. Yan, Y., Mao, Y., Li, B.: Second: Sparsely embedded convolutional detection. Sensors **18**(10), 3337 (2018)
29. Yan, Z., Li, P., Fu, Z., Xu, S., Shi, Y., Chen, X., Zheng, Y., Li, Y., Liu, T., Li, C., et al.: Int2: Interactive trajectory prediction at intersections. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8536–8547 (2023)
30. Yang, B., Luo, W., Urtasun, R.: Pixor: Real-time 3d object detection from point clouds. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 7652–7660 (2018)
31. Yang, K., Yang, D., Zhang, J., Li, M., Liu, Y., Liu, J., Wang, H., Sun, P., Song, L.: Spatio-temporal domain awareness for multi-agent collaborative perception. In:

Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 23383–23392 (2023)
32. Yang, L., Tang, T., Li, J., Chen, P., Yuan, K., Wang, L., Huang, Y., Zhang, X., Yu, K.: Bevheight++: Toward robust visual centric 3d object detection. arXiv preprint arXiv:2309.16179 (2023)
33. Yang, L., Yu, J., Zhang, X., Li, J., Wang, L., Huang, Y., Zhang, C., Wang, H., Li, Y.: Monogae: Roadside monocular 3d object detection with ground-aware embeddings. arXiv preprint arXiv:2310.00400 (2023)
34. Yang, L., Yu, K., Tang, T., Li, J., Yuan, K., Wang, L., Zhang, X., Chen, P.: Bevheight: A robust framework for vision-based roadside 3d object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21611–21620 (2023)
35. Ye, X., Shu, M., Li, H., Shi, Y., Li, Y., Wang, G., Tan, X., Ding, E.: Rope3d: The roadside perception dataset for autonomous driving and monocular 3d object detection task. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21341–21350 (2022)
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)
37. Yu, H., Tang, Y., Xie, E., Mao, J., Luo, P., Nie, Z.: Flow-based feature fusion for vehicle-infrastructure cooperative 3d object detection. arXiv preprint arXiv:2311.01682 (2023)
38. Yu, H., Yang, W., Ruan, H., Yang, Z., Tang, Y., Gao, X., Hao, X., Shi, Y., Pan, Y., Sun, N., et al.: V2x-seq: A large-scale sequential dataset for vehicle-infrastructure cooperative perception and forecasting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5486–5495 (2023)
39. Yurtsever, E., Lambert, J., Carballo, A., Takeda, K.: A survey of autonomous driving: Common practices and emerging technologies. IEEE access **8**, 58443–58469 (2020)
40. Zhang, Y., Lu, J., Zhou, J.: Objects are different: Flexible monocular 3d object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3289–3298 (2021)
41. Zhang, Z.: A flexible new technique for camera calibration. IEEE Transactions on pattern analysis and machine intelligence **22**(11), 1330–1334 (2000)
42. Zhao, H., Gao, J., Lan, T., Sun, C., Sapp, B., Varadarajan, B., Shen, Y., Shen, Y., Chai, Y., Schmid, C., et al.: Tnt: Target-driven trajectory prediction. In: Conference on Robot Learning. pp. 895–904. PMLR (2021)
43. Zhou, Y., Tuzel, O.: Voxelnet: End-to-end learning for point cloud based 3d object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4490–4499 (2018)
44. Zhou, Z., Ye, L., Wang, J., Wu, K., Lu, K.: Hivt: Hierarchical vector transformer for multi-agent motion prediction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8823–8833 (2022)
45. Zimmer, W., Birkner, J., Brucker, M., Nguyen, H.T., Petrovski, S., Wang, B., Knoll, A.C.: Infradet3d: Multi-modal 3d object detection based on roadside infrastructure camera and lidar sensors. arXiv preprint arXiv:2305.00314 (2023)
46. Zimmer, W., Creß, C., Nguyen, H.T., Knoll, A.C.: A9 intersection dataset: All you need for urban 3d camera-lidar roadside perception. arXiv preprint arXiv:2306.09266 (2023)