

HoloVIC: Large-scale Dataset and Benchmark for Multi-Sensor Holographic Intersection and Vehicle-Infrastructure Cooperative

Cong Ma¹, Lei Qiao¹, Chengkai Zhu¹, Kai Liu¹, Zelong Kong¹,
Qing Li¹, Xueqi Zhou¹, Yuheng Kan¹, Wei Wu^{1,2*}
¹SenseAuto Research ²Tsinghua University

<https://holovic.net>

Abstract

Vehicle-to-everything (V2X) is a popular topic in the field of Autonomous Driving in recent years. Vehicle-infrastructure cooperation (VIC) becomes one of the important research area. Due to the complexity of traffic conditions such as blind spots and occlusion, it greatly limits the perception capabilities of single-view roadside sensing systems. To further enhance the accuracy of roadside perception and provide better information to the vehicle side, in this paper, we constructed holographic intersections with various layouts to build a large-scale multi-sensor holographic vehicle-infrastructure cooperation dataset, called HoloVIC. Our dataset includes 3 different types of sensors (Camera, Lidar, Fisheye) and employs 4 sensor-layouts based on the different intersections. Each intersection is equipped with 6-18 sensors to capture synchronous data. While autonomous vehicles pass through these intersections for collecting VIC data. HoloVIC contains in total on 100k+ synchronous frames from different sensors. Additionally, we annotated 3D bounding boxes based on Camera, Fisheye, and Lidar. We also associate the IDs of the same objects across different devices and consecutive frames in sequence. Based on HoloVIC, we formulated four tasks to facilitate the development of related research. We also provide benchmarks for these tasks.

1. Introduction

Autonomous Driving has experienced notable progression in recent years. In order to further enhance the safety and overall perception, V2X (Vehicle-to-everything) has emerged as a new generation research focus in autonomous driving, which hopefully maximizes the potential of autonomous driving by interaction between Vehicle-to-Vehicle (V2V) and Vehicle-to-Infrastructure (V2I). Currently, Vehicle-Infrastructure Cooperation (VIC) has become a significant research area within V2X. Due to sen-

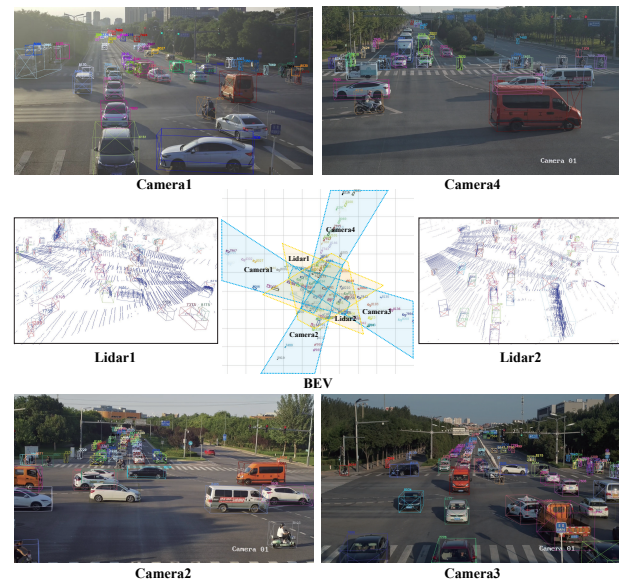


Figure 1. An example from HoloVIC dataset: The data and annotated 3D boxes on Camera, Lidar, and BEV, the same targets from different devices are labeled with the same Global ID.

sors from roadside at a higher viewpoint, the sensors cover wider field compared to the perspective of the vehicle, thus the captured data can provide information for the blind spots and farther areas that are beyond the sight of single-vehicle.

V2X has been gradually attracting more attention recently, and some pioneering datasets related to V2X have been released [15, 24, 33, 34]. V2X-sim [15] and Deep-Accident [24] are generated through simulations using CARLA [7] and SUMO [12]. On the other hand, DAIR-V2X [33] and V2X-seq [34] are collected from real-world scenarios, but the dataset rely on a single viewpoint by a pair of Camera and Lidar to capture data from different intersections. However, due to the complexity of traffic conditions, the targets captured from Camera are frequently occluded. Therefore, collecting data from a single-viewpoint sensor greatly limits the roadside perception capability.

To further accelerate research in the field of vehicle-infrastructure cooperation, in this paper, we constructed

*Corresponding Author

Table 1. Comparison of popular Datasets in Autonomous Driving and V2X. C: Camera, L: Lidar, F: Fisheye

Dataset	Source	View	With Trajectory	Multi-view Overlapping	Sensors Layouts of Vehicle	Sensors Layouts of Infrastructure	Synchronized Frames
KITTI [9]	real	vehicle	✓	-	4C+1L	-	14999
nuScenes [5]	real	vehicle	✓	-	6C+1L	-	200k
Waymo Open [21]	real	vehicle	✓	-	5C+5L	-	600k
ApolloScape [10]	real	vehicle	✓	-	1L	-	12360
DAIR-V2X-V [33]	real	vehicle	✗	-	2C+2L	-	22325
OPV2V [29]	simulated	V2V	✓	-	4C1L	-	11464
V2VReal [30]	real	V2V	✓	-	2C+1L	-	20000
DAIR-V2X-I [33]	real	infrastructure	✗	✗	-	1C+1L	10084
AICITY22 [17]	real	infrastructure	✓	✓	-	4C	2132
V2X-Sim [15]	simulated	V2V,VIC	✓	-	6C+1L	4C+1L	10000
V2XSet [28]	simulated	V2V,VIC	✓	✓	-	4C+1L	11447
DeepAccident [24]	simulated	V2V,VIC	✓	✓	6C+1L	6C+1L	57000
CARTI [2]	simulated	VIC	✓	✓	1L	1L	11000
DAIR-V2X-C [33]	real	VIC	✗	✗	2C+2L	1C+1L	38845
V2X-seq [34]	real	VIC	✓	✗	2C+2L	1C+1L	15000
HoloVIC (Ours)	real	VIC	✓	✓	2C+2L	4C+2L 12C+4F+2L	100k

several holographic intersections from diverse perspectives, where the areas captured by multiple sensors overlapping with each other. The intersections consists of 3 different types of sensors (C: Camera, L: Lidar, F: Fisheye). Each intersection is equipped with 6-18 sensors to capture synchronous data. We designed 4 sensor-layouts for different intersections, which includes 4C+2L; 8C+2L; 12C+4F+2L; 4C+2F+2L. Based on these intersections, we build a large-scale multi-viewpoint, multi-sensor dataset and benchmark, named HoloVIC. Meanwhile, autonomous vehicles pass through these intersections and capture data simultaneously with roadside for constructing VIC dataset.

HoloVIC consists of a total of 100k+ frames of synchronized data. Furthermore, the data are obtained from different sensors both on vehicle and road sides. We annotated more than 11.47M 2D&3D bounding boxes based on 3 types of sensors, and also associate the IDs of the same objects across different devices and consecutive frames in sequence. Then, we formed global trajectories for each individual object from a Bird’s-Eye View (BEV) perspective. Based on the annotation of HoloVIC, we generally formulate multiple tasks and benchmark: 1. Monocular 3D Detection (Mono3D); 2. Lidar 3D Detection 3. Multiple Object Tracking (MOT); 4. Multi-sensor Multi-object Tracking (MSMOT); 5. Vehicle-Infrastructure Cooperation Perception (VIC Perception)

The main contributions of our work are as follows:

- We construct several holographic intersections, which adopt 4 different sensor-layouts with Cameras, Fisheyes and Lidars for collecting synchronized data from all sensors in intersections.
- We release the first large-scale multi-viewpoint multi-sensor holographic intersection and vehicle-infrastructure cooperation dataset, named HoloVIC.
- We annotate 3D bounding boxes on 100k+ synchro-

nized frames based on all the sensors from road-side and vehicle-side, and associate the same targets with unique IDs to form global trajectories.

- We formulate five tasks and benchmark to promote the development of research on road-side perception and vehicle-infrastructure cooperative.

2. Related Work

We summarized open datasets in the field of autonomous driving and V2X, as shown in Tab.1. Based on the view of the data, datasets are categorized into single vehicle, V2V (Vehicle-to-Vehicle), Infrastructure (Roadside), and Vehicle-Infrastructure Cooperation (VIC). "Multi-view Overlapping" indicates whether the captured areas of road-side sensors overlapping with each other. We have provided information on the number and types of sensors on both the vehicle-side and road-side for each dataset, as well as the number of captured synchronized frames in the dataset.

Vehicle-Side Datasets [5, 9, 10, 21] primarily serve perception algorithms for autonomous driving. Common single-vehicle datasets include KITTI [9], a pioneering dataset that includes synchronized video frames obtained from stereo Cameras and Lidar, totaling 14,999 frames. It is used for pointcloud detection and tracking. The ApolloScape [10] includes additional pixel-wise masks, which is utilized for vehicle-side segmentation. The Waymo [5] contains a large amount of Camera and Lidar data, including pointclouds, images, and vehicle localization, it is primarily used for perception, prediction, and planning. And nuScenes [5] provides sensor data from a variety of sources, including Lidar, Radar, and Camera, with a significant number of calibrated frames.

Simulated V2X Datasets [15, 24, 28, 29] are generated by simulators. At the beginning of V2X, researchers utilized simulators such as CARLA [7], SUMO [12], and

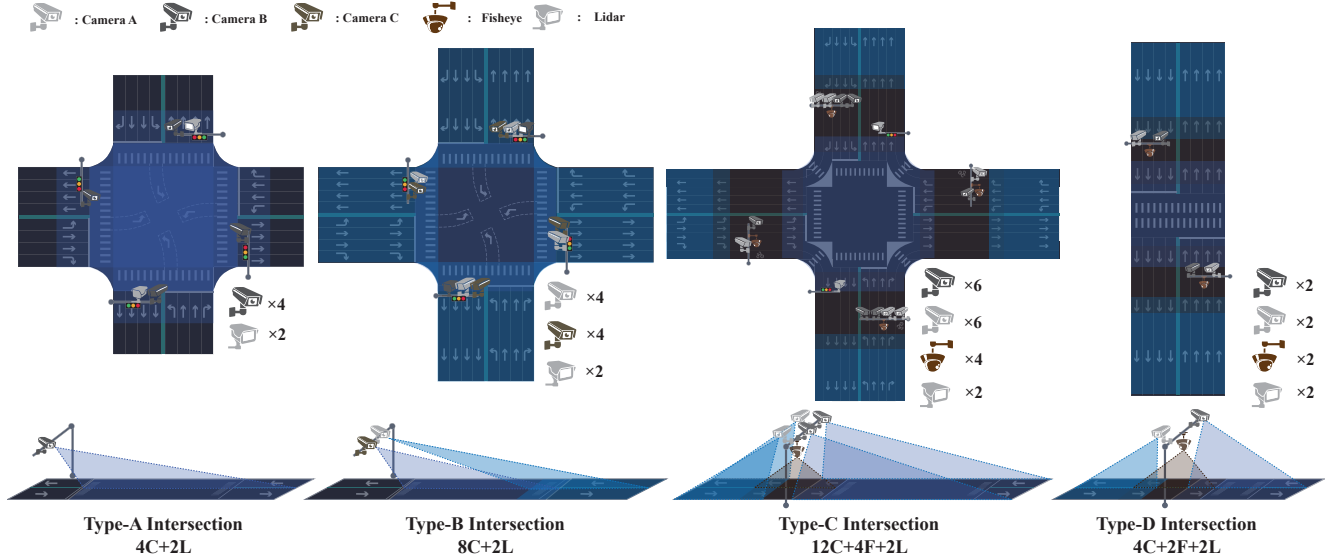


Figure 2. The configuration of holographic intersections: The figure illustrates three different sensors (C: Camera, L: Lidar, F: Fisheye) and four various sensor-layouts (4C+2L, 8C+2L, 12C+4F+2L, 4C+2F+2L) in holographic intersections

OpenCDA [27] to generate V2V and VIC data. The simulators are able to generate large-scale V2X data easily. OPV2V [29] is the first large-scale open simulated dataset for Vehicle-to-Vehicle perception. V2X-sim [15] and V2X-set [28] employs simulators to obtain traffic flows and collect sensor stream. DeepAccident [24] utilizes CARLA to generate approximately 7 times more than nuScenes, and formulated end-to-end motion and accident prediction task.

Real-world Roadside and VIC Datasets [6, 17, 33, 34] are difficult to be collected and annotated due to the dependency on infrastructure development. However, researchers are not satisfied with simulation datasets. In recent years, real-world datasets have been released. DAIR-V2X [33] is the first large-scale real-world VIC dataset, and V2X-Seq [34] is the first VIC sequential perception dataset. Additionally, multi-sensor with overlapping Datasets in roadside have been released such as WildTrack [6] and AIC-ITY2022 [17]. These datasets involve synchronized data streams from 4-7 Cameras and provide annotations for the same objects across multiple Cameras. Due to the difficulty of annotation, these datasets only provide annotations for a few thousand synchronized frames.

3. HoloVIC Open Dataset

3.1. Multi-Sensor Layouts

Due to the varying width of road, number of lanes, and shapes of intersection, we adopt four distinct sensor layouts to ensure optimal coverage of the intersection areas, as shown in Fig.2.

Type-A Intersection (4C+2L) utilizes four Cameras mounted on signal poles in all four directions to perceive

the central area of the intersection. Additionally, two Lidars are deployed in opposing directions in signal poles to form a layout of 4C+2L.

Type-B Intersections (8C+2L) encompass the central area of the intersection and the area beyond the stop line to increase the coverage of perception. We have installed a set of short-focus and telephoto Cameras on signal poles in all four directions to capture images of both the inside and outside of intersection respectively. Furthermore, two Lidar sensors are mounted in opposing directions, resulting in a sensor layout of 8C+2L.

Type-C Intersections (12C+4F+2L) have more lanes. To ensure better coverage of both the inside and outside areas of the intersection, we deploy 2 Cameras on both sides of the monitoring poles for four directions. In addition, to account for blind spots directly beneath the monitoring poles that are not captured by the Cameras, we install a Fisheye on each pole to effectively cover these blind spots. Similarly, a set of two Lidars are deployed in opposing directions, and the layout of type-C is 12C+4F+2L.

Type-D intersections (4C+2F+2L) only allows pedestrians cross laterally, while vehicles are restricted to going straight or making U-turns. Therefore, similar to the installation plan for Type C intersections, we install two Cameras and Fisheye in each of the two directions. And we deploy two Lidars to compose 4C+2F+2L.

3.2. Coordinate Systems

We introduce all of the coordinate systems involved in dataset as shown in Fig.3. All of the coordinate systems follow the right-hand rule. The definition, calibration and transformation of coordinate systems are as follows:

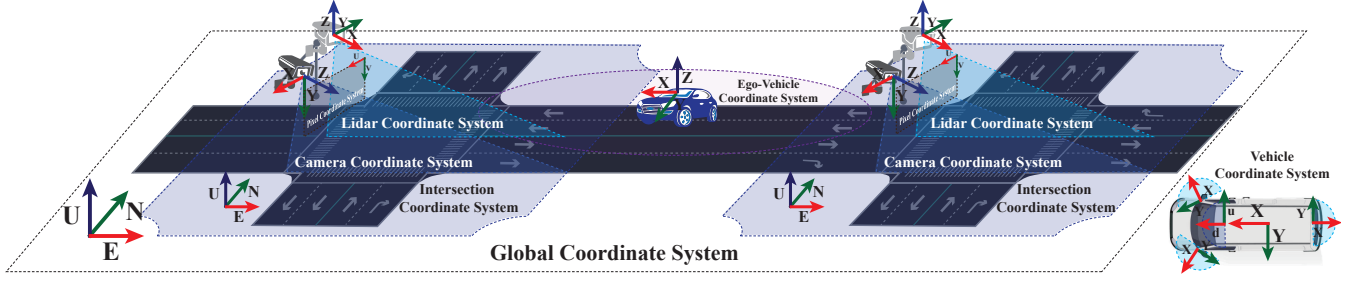


Figure 3. The coordinate systems in the HoloVIC dataset, involving all sensors on both vehicle and road sides.

Global Coordinate: To align the coordinates of intersections and vehicles, we select a point within the entire range of intersections as the origin of the global coordinate $(\omega_{x_0}, \omega_{y_0}, \omega_{z_0})$. The global coordinate is constructed based on the East-North-Up (ENU) coordinate system. Any position $(\omega_x, \omega_y, \omega_z)$ within the scene is calculated the relative actual distance to the origin in east-west and north-south as the x-axis and y-axis, respectively. The distance between two points is based on the WGS84, which is collected by Real-Time Kinematic (RTK). Both road and vehicle are aligned by converting their respective coordinate systems to the global coordinate system.

Intersection Coordinate also utilizes ENU coordinate system $(\sigma_x, \sigma_y, \sigma_z)$. We select a point as the origin of the intersection coordinate system for each intersection. The Intersection Coordinate System is the most crucial coordinate system for roadside perception, which is used for aligning all the sensor coordinate systems from the intersection.

Ego-Vehicle Coordinate $(\delta_x, \delta_y, \delta_z)$ is defined with the vehicle center as the origin. The forward, left and upward direction represent the positive X, Y and Z, respectively. All sensors on the vehicle will eventually be converged into the vehicle coordinate system, which further utilized for vehicle-infrastructure cooperative perception.

Lidar Coordinate: is a 3D coordinate system that includes the x , y , and z dimensions to represent the spatial position of pointclouds (x, y, z) . The transformation between the Lidar and Intersection/Ego-Vehicle coordinate can be achieved by calibrating multiple points between different coordinate systems and solving for the rotation $R_L \in \mathbb{R}^{3 \times 3}$ and translation $T_L \in \mathbb{R}^{3 \times 1}$ by Kabsch [11], projection equations in homogeneous coordinates is defined as:

$$\begin{pmatrix} x \\ y \\ z \\ 1 \end{pmatrix} = \begin{bmatrix} R_L^{3 \times 3} & T_L^{3 \times 1} \\ 0 & 1 \end{bmatrix} \begin{pmatrix} \sigma_x \\ \sigma_y \\ \sigma_z \\ 1 \end{pmatrix} \quad (1)$$

Camera Coordinate: is a 3D coordinate system, where the z -axis represents the depth in the Camera Coordinate System (c_x, c_y, c_z) . The transformation from the intersection coordinate system to the Camera Coordinate System

can be defined as:

$$\begin{pmatrix} c_x \\ c_y \\ c_z \\ 1 \end{pmatrix} = S^{4 \times 4} \begin{bmatrix} R_C^{3 \times 3} & T_C^{3 \times 1} \\ 0 & 1 \end{bmatrix} \begin{pmatrix} \sigma_x \\ \sigma_y \\ \sigma_z \\ 1 \end{pmatrix} \quad (2)$$

where $S \in \mathbb{R}^{4 \times 4}$ is used for mapping between Intersection-Camera Coordinates axes. Rotation $R_C^{3 \times 3}$ and Translation $T_C^{3 \times 1}$ are calculated by solving PnP (Perspective-n-Point).

Pixel Coordinate is obtained by projecting the Camera coordinate onto the imaging plane, which is transformed into 2D coordinate system (u, v) . The transformation from Camera coordinate to pixel coordinate is formulated as:

$$Z_c \begin{pmatrix} u \\ v \\ 1 \end{pmatrix} = K^{-3 \times 3} \begin{pmatrix} c_x \\ c_y \\ c_z \end{pmatrix}, \quad K^{-3 \times 3} = \begin{bmatrix} f_x & -1 & u_0 \\ 0 & f_y & v_0 \\ 0 & 0 & 1 \end{bmatrix} \quad (3)$$

$K \in \mathbb{R}^{3 \times 3}$ indicates the intrinsic matrix of Camera, f_x, f_y denote the focal of the Camera in x -axis, y -axis. u_0, v_0 represent the center of image. We calibrated each Camera before capturing the data by Chessboard Calibration.

3.3. Ground Truth Labels

In the HoloVIC dataset, we provide high-quality annotation, the ground truth include:

1. We annotated 3D bounding boxes $[x, y, z, l, w, h, \theta]$ and category in both the Camera coordinate for images and in the Lidar coordinate for pointclouds, where x, y, z represent the center of 3D box in 3D coordinate, while l, w, h correspond to length, width and height, and θ is the orientation (yaw) of 3D box. The category η include "Vehicle", "Cyclist" and "Pedestrian".

2. We assigned Track ID to the same target across the temporal sequence of each sensor. For the same target within a sequence, it has unique tracking ID, τ . The bounding boxes with the same τ are linked together to form a complete trajectory.

3. We associated global ID to the same object across different sensors at the same timestamp. For the same target within a intersection, it only has unique global ID ρ .

Additionally, we also generate the 3D box of Target ρ in intersection coordinate $\rho : [\sigma_x, \sigma_y, \sigma_z, l, w, h, \theta, \eta]$ based on all of the boxes with global ID of ρ .

4. We matched the the same objects between from the vehicle-side ν and road-side ρ as the unique global ID.

4. Tasks & Metrics

Based on HoloVIC data and annotation, we formulated the tasks based on Single-Sensor, Multi-Sensor and VIC Perception, which are introduced as following.

4.1. Single-Sensor Perception

4.1.1 3D Detection

The 3D Detection task of the HoloVIC includes Monocular 3D (Mono3D) and Lidar 3D Detection. Given an image frame or pointcloud, the detection model is used to obtain the position, shape, orientation, and category of targets in the form of 3D bounding boxes $[x, y, z, l, w, h, \theta, \eta]$. We refer to the metrics from Rope3D [32], KITTI [9], and nuScenes [9] and adopt mAP (mean Average Precision) and mAOS (mean Average Orientation Similarity) to evaluate detectors for both the tasks.

Mean Average Precision (mAP) is used to evaluate the accuracy of object detection. The mAP score is influenced by the position, size, orientation, classification, and confidence of the predicted bounding boxes, which is defined as:

$$mAP = \frac{1}{\mathbb{C}} \sum_{\eta \in \mathbb{C}} AP_{\eta} \quad (4)$$

where \mathbb{C} is the set of category, AP indicates Average Precision [8]. For 2D Detection task, we utilize 3D intersection over union (3D IOU) to match the prediction and ground truth, which is formulated as:

$$AP|_n = \frac{1}{|Rc|} \sum_{r \in Rc} \max_{\tilde{r}: \tilde{r} > r} Pr(\tilde{r}) \quad (5)$$

where $Pr(\tilde{r})$ is the precision at a certain recall threshold $r \in \{\frac{1}{n}, \frac{2}{n}, \dots, 1\}$, $|Rc| = n + 1$ indicates the number of elements in Rc , we set n to 40.

Mean Average Orientation Similarity is utilized to measure the precision of yaw angle, which is defined as:

$$mAOS = \frac{1}{\mathbb{C}} \sum_{\eta \in \mathbb{C}} AOS_{\eta} \quad (6)$$

where AOS (Average Orientation Similarity) [21], for 3D Detection task, we only consider the angle around the z-axis rotation (yaw), which is formulated as:

$$AOS|_n = \frac{1}{|Rc|} \sum_{r \in Rc} \max_{\tilde{r}: \tilde{r} > r} Ori(\tilde{r}) \quad (7)$$

$$Ori(r) = \frac{1}{|D(r)|} \sum_{i \in D(r)} \frac{1 + \cos \Delta_{\theta}^{(i)}}{2} \delta_i \quad (8)$$

where D is set of true positive samples, $\Delta_{\theta}^{(i)}$ is the angle difference of sample i . To penalize multiple boxes matching to the same ground truth, we set $\delta_i = 1$ for box i if it has already been matched to a ground truth, otherwise $\delta_i = 0$.

4.1.2 Tracking

The Tracking task of the HoloVIC consists of 2D Tracking on videos and 3D Tracking both on video and point-cloud sequences. Given the sequence and the detection result as inputs, tracking model aims to associate the bounding boxes for same target across the sequence and assign Track ID τ for each target. We refer to the MOT metrics [4, 19] and adopt MOTA (MOT Accuracy), IDF1 (ID F1 Score) to measure Tracking task.

The definition of MOTA metric is as follows:

$$MOTA = 1 - \frac{|FP| + |FN| + |IDSw|}{|gtDet|} \quad (9)$$

where FP (False Positive) and FN (False Negative) indicate the wrongly detected and missed from detection. IDSw (ID switch) represents the tracker incorrectly assign different IDs for same target or swap the IDs of two objects. We also evaluate MT (Mostly Tracked), and ML (Mostly Lost) as additional reference metrics from CLEAR-MOT [4].

IDF1 calculates one-to-one mapping the Trajectories between prediction and ground truth, which is defined as:

$$IDF1 = \frac{2|IDTP|}{2|IDTP| + |IDFP| + |IDFN|} \quad (10)$$

where IDTP (identity true positives) are matches on overlapping part of trajectories that are matched. IDFN (identity false negatives) and IDFP (identity false positives) are calculated from both non-overlapping of matched trajectories, and the remaining trajectories that are not matched. We also evaluate IDP (ID-Precision) and IDR (ID-Recall) as reference metrics from Identity Metrics [19].

4.2. Multi-Sensor Perception

The Multi-Sensor Perception task is primarily used to analyze the overall situation at intersections. It involves multiple sensors capturing data simultaneously in intersections. The perception task is divided into detection and tracking, which utilize data and the spatial-temporal relationships between devices to output perception results in the intersection coordinate system, which are presented in the form of 3D boxes $[\sigma_x, \sigma_y, \sigma_z, l, w, h, \theta, \eta]$. The ground truth 3D

Table 2. The details of HoloVIC dataset, including the proportion allocation between Infrastructure and VIC; The proportion of different scenes; The distribution of training, testing and validation sets; The count of synchronized frames and annotated 3D boxes for each scene.

View	Scene	Layout	Distribution		Ratio of Dataset			Num of Frames	3D Boxes (Global)	3D Boxes (Local)
			View	Scene	Train	Test	Valid			
Infrastructure	Int-1	4C+2L	70%	30%	50%	40%	10%	21k	600k	2M
	Int-2	8C+2L		30%				21k	480k	3.6M
	Int-3	12C+4F+2L		10%				7k	85k	1.2M
	Int-4	4C+2F+2L		20%				14k	120k	420k
	Int-5	12C+4F+2L		10%				7k	87k	1.6M
VIC	VIC-1	4C+2L	30%	30%	50%	40%	10%	18k	210k	650k
	VIC-2	8C+2L		30%				18k	100k	580k
	VIC-3	12C+4F+2L		10%				3k	40k	580k
	VIC-4	4C+2F+2L		20%				6k	65k	205k
	VIC-5	12C+4F+2L		10%				3k	45k	640k
Total	-	-	100%		45%	46%	9%	100k	1.8M	11.47M

boxes from each sensor are merged in the intersection coordinate system to create ground truth for the multi-sensor perception task, which is defined as:

$$GT_i = \{GT_s, s \in S\} \quad (11)$$

where GT_s is the ground truth in s -th Sensor, S is the set of Sensors in Intersection. We refer to 3D Detection metrics [9] and adopt mAP and mAOS as the evaluation metrics for multi-sensor 3D detection. In addition, we utilized metrics of Multi-target Multi-camera Tracking (MTMCT) [19] and selected MOTA and IDF1 as the metrics for multi-sensor 3D Tracking. All the definitions of metrics are consistent with the relevant descriptions in Eq.4-10.

4.3. VIC Perception

The VIC perception task focuses on the cooperative perception between vehicles and infrastructure. Given synchronized data from both on vehicle and road sides, VIC is used to evaluate the capability fusing information and assess the benefits brought to vehicles by roadside perception.

We firstly align the position of the 3D boxes from the intersection coordinate to the ego-vehicle coordinate. The ground truth 3D boxes are merged from both vehicle-side and road-side at the same timestamp t :

$$GT_{vic}^t = GT_v^t \cup GT_{i|v}^t \quad (12)$$

$$GT_{i|v}^t = \{\rho_{i \rightarrow v} \in GT_i^t, \|\rho_{i \rightarrow v}[\sigma] - \nu[\sigma]\| < \varepsilon\} \quad (13)$$

where GT_i^t is the ground truth from roadside at t , $\rho_{i \rightarrow v}$ indicates the position of target from roadside after transformed to the ego vehicle coordinate. We defined a range around the position of ego vehicle $\nu[\sigma]$, any roadside 3D box that exceeds the distance ε is discarded. We refer to the metrics from DAIR-V2X [33, 34] and adopt mAP, mAOS for Detection, MOTA and IDF1 for Tracking. We evaluate results separately with same ground truth GT_{vic}^t : 1. ONLY

vehicle-side data as input 2. Both vehicle-side and roadside as input. This approach allows for a better comparison to quantify the benefits brought to ego vehicle perception by incorporating roadside data.

5. Experiments

5.1. Benchmarks Setup

Our HoloVIC dataset contains 100k frames of synchronized data, with 70% dedicated to holographic intersection data and 30% allocated for vehicle-infrastructure cooperation (VIC) data. To ensure privacy, we have applied blurring to all faces, vehicle plates and road signs in all data. The details of the dataset are illustrated in Tab.2 The holographic intersection data is collected from five different intersections with four distinct sensor layouts. The VIC data collection begins when a vehicle enters an intersection and ends when it exits the corresponding intersection area.

HoloVIC is divided into training, testing, and validation sets, with a ratio of 50%, 40%, and 10% respectively. The training and validation sets include ground truth labels, while the testing set only provides data. One of the five holographic intersections is exclusively assigned to the testing set, while the remaining intersections are included in all three sets. Algorithms can be submitted to our benchmark for online evaluation of the corresponding task.

5.2. Sensor Specifications

The detailed specifications of all devices as shown in Tab.5. Due to the varying distances between the areas and poles, we select Cameras with different focal lengths and Field of View (FOV) to cover the areas. All devices are synchronized in time via Network Time Protocol (NTP) before data collection, we utilize a time interval of 100ms as the global timestamp for intersections, and match the frame from each device with the nearest timestamp adjacent to the global timestamp. This process ultimately yields synchronized multi-sensor data at a frame rate of 10 fps.

Table 3. Mono3D and Lidar 3D Detection results on validation sets

Task	Method	Vehicle@0.2/0.7			Cyclist@0.2/0.5			Pedestrian@0.2/0.5		
		3d AP	bev AP	AOS	3d AP	bev AP	AOS	3d AP	bev AP	AOS
Mono3D Detection	FCOS3D [22]	47.21	37.17	68.05	11.68	11.60	47.96	11.31	6.99	53.79
	PGD [23]	52.54	43.21	74.23	17.31	19.15	53.18	8.67	4.78	58.35
Lidar 3D Detection	SECOND [31]	70.54	81.35	60.84	77.35	80.86	57.80	26.25	29.81	29.90
	Pointpillars [14]	70.17	81.69	60.61	73.04	78.84	57.8	27.43	31.31	24.76
	PVRCNN [20]	72.78	82.16	60.50	77.06	82.22	58.01	21.91	24.47	22.54
	Transfusion-L [1]	69.79	82.86	60.49	77.51	81.64	58.32	45.51	52.57	44.54

Table 4. The performance of 2D/3D MOT on validation sets

Task	Detector	Tracker	IDF1 \uparrow	IDP \uparrow	IDR \uparrow	MOTA \uparrow	MT \uparrow	ML \downarrow
2D MOT	YOLOv8 [18]	DeepSort [26]	53.72	67.41	45.53	41.72	26.38	34.74
		Tractor [3]	59.34	71.58	53.69	57.70	39.85	24.21
		FairMOT [35]	68.17	72.35	62.57	63.35	47.98	17.45
Mono3D MOT	FCOS3D [22]	AB3DMOT [25]	42.85	57.36	32.17	39.34	21.81	27.37
Lidar 3D MOT	Pointpillars [14]	AB3DMOT [25]	57.62	69.59	43.86	51.45	24.70	30.71

Table 5. Sensor Specifications in HoloVIC

Scene	Sensor	Details
Infrastructure	Camera A	RGB, 25hz, 1920 \times 1080 FOV:[31.6 $^\circ$, 17.0 $^\circ$]
	Camera B	RGB, 25hz, 1920 \times 1080 FOV:[47.7 $^\circ$, 25.2 $^\circ$]
	Camera C	RGB, 25hz, 1920 \times 1080 FOV:[111.78 $^\circ$, 63.16 $^\circ$]
	Fisheye	RGB, 25hz, 2048 \times 2048 FOV:[180.0 $^\circ$, 180.0 $^\circ$]
	Lidar	150 beams, 10hz, 1.6M pps FOV:[100.0 $^\circ$, 40.0 $^\circ$]
Vehicle	Camera A	RGB, 25hz, 1920 \times 1080 FOV:[30.0 $^\circ$, 17.0 $^\circ$]
	Camera B	RGB, 25hz, 1920 \times 1080 FOV:[120.0 $^\circ$, 17.0 $^\circ$]
	Lidar A	40 beams, 10hz, 720k pps FOV:[286.48 $^\circ$, -25 $^\circ$ to 15 $^\circ$]
	Lidar B	64 beams, 10hz, 384k pps FOV:[360.0 $^\circ$, -25 $^\circ$ to 15 $^\circ$]

5.3. Baselines

5.3.1 3D Detection

Monocular 3D Detection To evaluate the performance of image-only monocular 3D detection, we selected widely-used methods as our Mono3D baseline models FCOS3D [22] for validation, which is a fully convolutional single-stage detector and transforms 7-DoF 3D targets to the image domain and decouple them as 2D and 3D attributes. In addition, we choose PGD [23] as a comparison to our baseline method.

Lidar 3D Detection To demonstrate the capabilities of well-known Lidar Detectors in our Lidar 3D task, we implemented four methods with different architectures: SECOND [31], Pointpillars [14], PVRCNN [20], and Transfusion-L [1], the methods utilized different techniques such as Voxelization, Pillar-based, Two-stage Region Proposals, and Transformers to achieve accurate and efficient detection in pointcloud, where we select Pointpillars as our baseline for Lidar 3D Detection.

5.3.2 Multiple Object Tracking

2D/3D MOT We follow the Tracking-by-Detection paradigm using 2D or 3D bounding boxes as inputs. The 2D boxes are generated by the Yolov8 [18] detector pretrained on the COCO dataset [16]. We selected DeepSort [26], Tractor [3], and FairMOT [35] as our 2D MOT baseline. And the 3D boxes from image and pointclouds are provided by FCOS3D and Pointpillars respectively. As for the 3D tracker, we select AB3DMOT [25] as 3D MOT baseline.

5.3.3 Multi-Sensor Tracking

Multi-Sensor Multi-Object Tracking We have developed a multi-sensor late-fusion framework to fuse the local tracklets as global trajectories based on the intersection coordinate system. In this framework, we utilize the 3D MOT baselines to generate tracking results, which serve as inputs for the fusion process. To associate the same objects from different Cameras, we employ the Hungarian Algorithm [13] to solve the bipartite graph problem. Finally, we integrate the local information from each device to generate global trajectory information.

5.3.4 VIC Perception

Vehicle-Intersection Cooperative Perception focuses on fusing both data and perception results from road and vehicle sides. We design a VIC late-fusion framework as the baseline. Both sides generate perception results as inputs for late-fusion. The 3D boxes from the road-side perception results are transformed to the ego-vehicle coordinate system based on the relative positions between coordinates. We utilize the Hungarian Algorithm [13] to match the 3D boxes belonging to the same targets on both sides. The 3D bounding boxes are associated in consecutive frames using the AB3DMOT algorithm to generate the fused trajectory between the vehicle and the road.

Table 6. Detection and tracking accuracy of only-vehicle, VIC with "1C+1L" and "4C+2L" on Int-1 and VIC-1 validation sets.

View	Detection				Tracking			
	Metric	Range(m)			IDF1↑	MOTA↑	MT↑	ML↓
		[0,30]	[30,50]	[50,70]				
Vehicle (V-only)	bev AP	97.85	90.97	46.95	86.57	75.64	50.01	15.38
	AOS	98.16	94.01	49.67				
VIC (V+I: 1C+1L)	bev AP	98.09	91.51	59.82	90.02	81.19	61.54	7.69
	AOS	98.42	94.21	63.00				
VIC (V+I 4C+2L)	bev AP	98.09	92.06	85.75	95.47	90.19	92.31	3.85
	AOS	98.42	94.75	90.43				

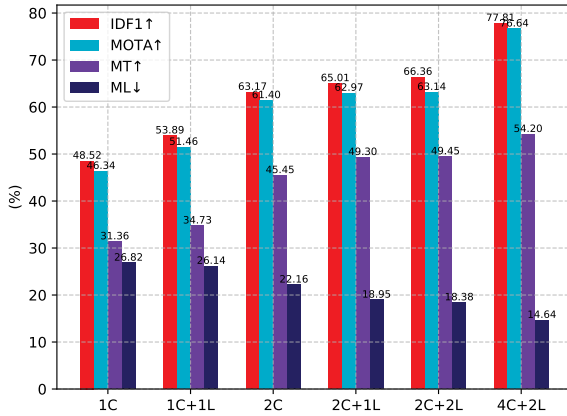


Figure 4. The impact of different sensor-layouts on multi-sensor tracking accuracy in Int-1 validation set.

5.4. Analysis

Performance of Single-Sensor Perception, we evaluated the perception results on single sensor in the HoloVIC validation set. The ground truth was defined by covering objects visible in both the Camera and Lidar. The detection results based on Mono3D and Lidar 3D are shown in Tab. 3, where we evaluated 3D AP, BEV AP, and AOS separately for three different categories. Lidar 3D generally higher than Mono3D for AP, especially for cyclists and pedestrians. However, in terms of orientation error, the AOS metric is generally higher for image-based perception compared to Lidar 3D. Regarding tracking, as shown in Tab. 4, Mono3D-based tracking exhibits relatively poorer performance due to weaker detection results compared to other methods. Lidar 3D tracking is performed in 3D space, which mitigates the impact of occlusion between objects that often leads to tracklet mixing in 2D tracking. However, AB3DMOT [25] only considers the position and motion of the target for tracking and cannot reconnect lost targets through strategies like ReID used in 2D MOT. Therefore, the MOTA and IDF1 metrics fall between the performance of 2D MOT methods.

Performance with different Sensor-layouts, we evaluated the performance using the Int-1 (4C+2L) validation set. The ground truth data was obtained by targets within the areas covered by the field of view of all devices. As depicted in Fig.4, we evaluate IDF1, MOTA, MT, and ML in the intersection coordinate system based on six differ-

ent sensor layouts. As the number of sensors increased, all metrics showed gradual improvement. Multiple viewpoints provided better coverage of blind spots and enhanced trajectory continuity. Therefore, the 2C method achieved a higher improvement in the 1C results compared to 1C+1L, with an increase of 9.28%, 9.94%, 10.72%, and 3.98% respectively. Similarly, 4C+2L exhibited improvements of 11.36%, 13.5%, 4.75%, 3.74% compared to 2C+2L.

Performance from VIC, we evaluated the detection and tracking performance within ranges of 30m, 50m, and 70m around the center of the ego-vehicle. The vehicle travels within a range of 70m from the intersection coordinate system. The ground truth includes all bounding boxes on both the road-side and vehicle-side within the corresponding evaluation range. We separately evaluated the detection and tracking performance of only the vehicle-side perception, VIC perception with 1C+1L, and 4C+1L based on Scene of Int-1. As shown in Tab.6, the performance of VIC perception is better than vehicle-side in all metrics. Moreover, increasing the number of sensors from the roadside can greatly improve the performance for the vehicle. The perception capability beyond 50m deteriorates rapidly due to target occlusion and the limited sensing range of the ego-vehicle sensors. Through VIC, the performance is improved up to 38.8% and 40.76% in terms of AP and AOS, respectively. The accuracy in the 50-70m range on the vehicle-side remains consistent with the range of 0-50m. Additionally, roadside perception provides an 8.9% and 14.55% improvement in IDF1 and MOTA for the vehicle-side.

6. Conclusion

In this paper, we constructed several holographic intersections with three different types of sensors and four different sensor-layouts. We annotated 11.47M 3D Boxes in total on 100k synchronous frames from different sensors. We propose a large-scale holographic intersection and vehicle-infrastructure cooperative dataset, HoloVIC. Furthermore, the dataset is divided into multiple tasks in various dimensions for further research on perception models. In the future, we plan to expand more tasks and benchmarks based on HoloVIC, such as trajectory prediction, and explore the additional benefits that roadside perception can bring to vehicle-side perception.

References

- [1] Xuyang Bai, Zeyu Hu, Xinge Zhu, Qingqiu Huang, Yilun Chen, Hongbo Fu, and Chiew-Lan Tai. Transfusion: Robust lidar-camera fusion for 3d object detection with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1090–1099, 2022. 7
- [2] Zhengwei Bai, Guoyuan Wu, Matthew J Barth, Yongkang Liu, Emrah Akin Sisbot, and Kentaro Oguchi. Pillargrid: Deep learning-based cooperative perception for 3d object detection from onboard-roadside lidar. In *2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC)*, pages 1743–1749. IEEE, 2022. 2
- [3] Philipp Bergmann, Tim Meinhardt, and Laura Leal-Taixe. Tracking without bells and whistles. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 941–951, 2019. 7
- [4] Keni Bernardin and Rainer Stiefelwagen. Evaluating multiple object tracking performance: the clear mot metrics. *EURASIP Journal on Image and Video Processing*, 2008:1–10, 2008. 5
- [5] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multi-modal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020. 2
- [6] Tatjana Chavdarova, Pierre Baqué, Stéphane Bouquet, Andrii Maksai, Cijo Jose, Timur Bagautdinov, Louis Lettry, Pascal Fua, Luc Van Gool, and François Fleuret. Wildtrack: A multi-camera hd dataset for dense unscripted pedestrian detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5030–5039, 2018. 3
- [7] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. Carla: An open urban driving simulator. In *Conference on robot learning*, pages 1–16. PMLR, 2017. 1, 2
- [8] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88:303–338, 2010. 5
- [9] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3354–3361. IEEE, 2012. 2, 5, 6
- [10] Xinyu Huang, Xinjing Cheng, Qichuan Geng, Binbin Cao, Dingfu Zhou, Peng Wang, Yuanqing Lin, and Ruigang Yang. The apolloscape dataset for autonomous driving. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 954–960, 2018. 2
- [11] Wolfgang Kabsch. A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography*, 32(5):922–923, 1976. 4
- [12] Daniel Krajzewicz, Jakob Erdmann, Michael Behrisch, and Laura Bieker. Recent development and applications of sumo simulation of urban mobility. *International journal on advances in systems and measurements*, 5(3&4), 2012. 1, 2
- [13] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955. 7
- [14] Alex H Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. Pointpillars: Fast encoders for object detection from point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12697–12705, 2019. 7
- [15] Yiming Li, Dekun Ma, Ziyang An, Zixun Wang, Yiqi Zhong, Siheng Chen, and Chen Feng. V2x-sim: Multi-agent collaborative perception dataset and benchmark for autonomous driving. *IEEE Robotics and Automation Letters*, 7(4):10914–10921, 2022. 1, 2, 3
- [16] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 7
- [17] M. Naphade, S. Wang, D. C. Anastasiu, Z. Tang, M. Chang, Y. Yao, L. Zheng, M. Shaiqur Rahman, A. Venkatachalapathy, A. Sharma, Q. Feng, V. Ablavsky, S. Sclaroff, P. Chakraborty, A. Li, S. Li, and R. Chellappa. The 6th ai city challenge. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 3346–3355. IEEE Computer Society, 2022. 2, 3
- [18] Dillon Reis, Jordan Kupec, Jacqueline Hong, and Ahmad Daoudi. Real-time flying object detection with yolov8. *arXiv preprint arXiv:2305.09972*, 2023. 7
- [19] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *European conference on computer vision*, pages 17–35. Springer, 2016. 5, 6
- [20] Shaoshuai Shi, Chaoxu Guo, Li Jiang, Zhe Wang, Jianping Shi, Xiaogang Wang, and Hongsheng Li. Pv-rcnn: Point-voxel feature set abstraction for 3d object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10529–10538, 2020. 7
- [21] Pei Sun, Henrik Kretschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2446–2454, 2020. 2, 5
- [22] Tai Wang, Xinge Zhu, Jiangmiao Pang, and Dahua Lin. Fcos3d: Fully convolutional one-stage monocular 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 913–922, 2021. 7
- [23] Tai Wang, ZHU Xinge, Jiangmiao Pang, and Dahua Lin. Probabilistic and geometric depth: Detecting objects in perspective. In *Conference on Robot Learning*, pages 1475–1485. PMLR, 2022. 7
- [24] Tianqi Wang, Sukmin Kim, Wenxuan Ji, Enze Xie, Chongjian Ge, Junsong Chen, Zhenguo Li, and Ping Luo. Deepaccident: A motion and accident prediction

- benchmark for v2x autonomous driving. *arXiv preprint arXiv:2304.01168*, 2023. [1](#), [2](#), [3](#)
- [25] Xinshuo Weng, Jianren Wang, David Held, and Kris Kitani. Ab3dmot: A baseline for 3d multi-object tracking and new evaluation metrics. *arXiv e-prints*, 2020. [7](#), [8](#)
- [26] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *2017 IEEE international conference on image processing (ICIP)*, pages 3645–3649. IEEE, 2017. [7](#)
- [27] Runsheng Xu, Yi Guo, Xu Han, Xin Xia, Hao Xiang, and Jiaqi Ma. Opencda: an open cooperative driving automation framework integrated with co-simulation. In *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*, pages 1155–1162. IEEE, 2021. [3](#)
- [28] Runsheng Xu, Hao Xiang, Zhengzhong Tu, Xin Xia, Ming-Hsuan Yang, and Jiaqi Ma. V2x-vit: Vehicle-to-everything cooperative perception with vision transformer. In *European conference on computer vision*, pages 107–124. Springer, 2022. [2](#), [3](#)
- [29] Runsheng Xu, Hao Xiang, Xin Xia, Xu Han, Jinlong Li, and Jiaqi Ma. Opv2v: An open benchmark dataset and fusion pipeline for perception with vehicle-to-vehicle communication. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 2583–2589. IEEE, 2022. [2](#), [3](#)
- [30] Runsheng Xu, Xin Xia, Jinlong Li, Hanzhao Li, Shuo Zhang, Zhengzhong Tu, Zonglin Meng, Hao Xiang, Xiaoyu Dong, Rui Song, 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*, pages 13712–13722, 2023. [2](#)
- [31] Yan Yan, Yuxing Mao, and Bo Li. Second: Sparsely embedded convolutional detection. *Sensors*, 18(10):3337, 2018. [7](#)
- [32] Xiaoqing Ye, Mao Shu, Hanyu Li, Yifeng Shi, Yingying Li, Guangjie Wang, Xiao Tan, and Errui Ding. 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*, pages 21341–21350, 2022. [5](#)
- [33] Haibao Yu, Yizhen Luo, Mao Shu, Yiyi Huo, Zebang Yang, Yifeng Shi, Zhenglong Guo, Hanyu Li, Xing Hu, Jirui Yuan, 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*, pages 21361–21370, 2022. [1](#), [2](#), [3](#), [6](#)
- [34] Haibao Yu, Wenxian Yang, Hongzhi Ruan, Zhenwei Yang, Yingjuan Tang, Xu Gao, Xin Hao, Yifeng Shi, Yifeng Pan, Ning Sun, 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*, pages 5486–5495, 2023. [1](#), [2](#), [3](#), [6](#)
- [35] Yifu Zhang, Chunyu Wang, Xinggong Wang, Wenjun Zeng, and Wenyu Liu. Fairmot: On the fairness of detection and re-identification in multiple object tracking. *International Journal of Computer Vision*, 129:3069–3087, 2021. [7](#)