RCooper: A Real-world Large-scale Dataset for Roadside Cooperative Perception

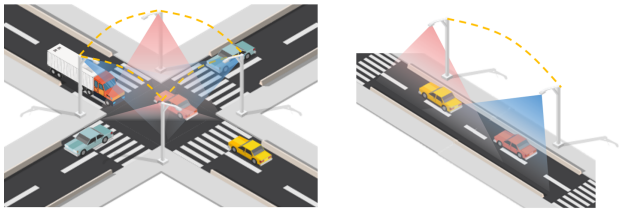
Ruiyang Hao[1,†], Siqi Fan[1,†], Yingru Dai[1,2], Zhenlin Zhang[3], Chenxi Li[3], Yuntian Wang[3], Haibao Yu[1,4], Wenxian Yang[1], Jirui Yuan[1], Zaiqing Nie[1,*]
[1]Institute for AI Industry Research (AIR), Tsinghua University
[2] Department of Electronic Engineering, Tsinghua University
[3] China Automotive Innovation Corporation [4] The University of Hong Kong

Abstract

*The value of roadside perception, which could extend the boundaries of autonomous driving and traffic management, has gradually become more prominent and acknowledged in recent years. However, existing roadside perception approaches only focus on the single-infrastructure sensor system, which cannot realize a comprehensive understanding of a traffic area because of the limited sensing range and blind spots. Orienting high-quality roadside perception, we need **R**oadside **Coo**perative **Per**ception **(RCooper)** to achieve practical area-coverage roadside perception for restricted traffic areas. Rcooper has its own domain-specific challenges, but further exploration is hindered due to the lack of datasets. We hence release the first real-world, large-scale RCooper dataset to bloom the research on practical roadside cooperative perception, including detection and tracking. The manually annotated dataset comprises 50k images and 30k point clouds, including two representative traffic scenes (i.e., intersection and corridor). The constructed benchmarks prove the effectiveness of roadside cooperation perception and demonstrate the direction of further research. Codes and dataset can be accessed at: https://github.com/AIR-THU/DAIR-RCooper.*

1. Introduction

With the development of the Internet of Things (IoT), 5G, and artificial intelligence technologies, the value of roadside perception has gradually become more prominent, which has drawn broad attention in recent years [13, 42–46]. Roadside perception is of great significance to both autonomous driving and traffic management. For autonomous driving, roadside sensor systems provide intelligent vehicles with complementary on-road messages beyond the on-board perspective, assisting vehicles to have a more comprehensive and clear understanding of surrounding environments, thereby trending to better and safer driving of L5 autonomous [18]. As for intelligent transportation systems, the downstream traffic management task, e.g., traffic flow control, traffic participants monitoring, and illegal activities monitoring, can be further improved with more comprehensive understanding by roadside perception [50]. Therefore, orienting high-quality autonomous driving and traffic management, how to achieve practical area-coverage roadside perception for a restricted traffic area is a significant task.



(a) Intersection RCooper scene (b) Corridor RCooper scene

Figure 1. **R**oadside **Coo**perative **Per**ception **(RCooper)** is expected to achieve practical area-coverage roadside perception for restricted traffic areas, which would further promote both the autonomous driving and traffic management. The complex roadside system is boiled down to two typical roadside settings, i.e., (a) intersection RCooper scenes and (b) corridor RCooper scenes.

Previous roadside perception [13, 21, 42–44, 53] mainly concentrates on the perception from an independent roadside view due to the accessible dataset [45, 46]. However, they can not realize a comprehensive understanding of a traffic area. Single-infrastructure roadside perception is constrained by the installation perspective, leading to the limited sensing range and blind spots, which can be handled via cross-infrastructure cooperation. Observation from various views can extend the sensing range, reduce blind spots, and further enhance the understanding of the same instance. Towards practical applications, **R**oadside

*Corresponding author. † indicates equal contribution. Work done at AIR. For any questions, please email dair@air.tsinghua.edu.cn.

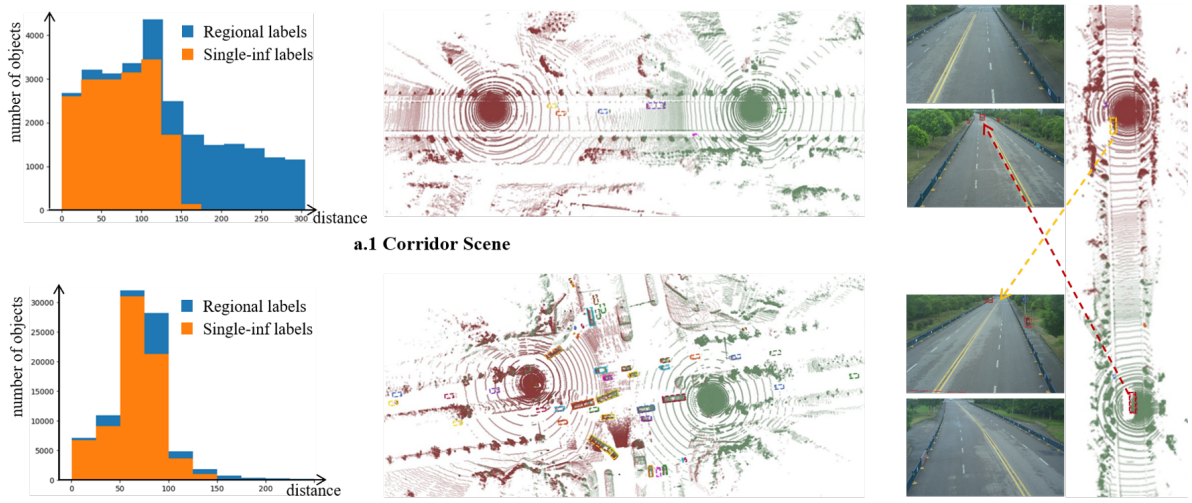


Figure 2. Independent roadside 3D perception (red point clouds) is limited by sensing range and blind spots. (a) The infrastructure-side cooperation can effectively extend the sensing range to cover the whole corridor scene, and the observation from multiple views can weaken the impact of occlusion in the complex intersection scene. (b) The area under the infrastructure is the camera's blind spot, which is perceptible from the adjacent infrastructure's camera.

Cooperative **Per**ception **(RCooper)** is expected to achieve area-coverage roadside perception for restricted traffic areas, which is shown in Fig. 1. The capabilities of RCooper, including extending the sensing range and reducing blind spots, are illustrated in Fig. 2.

Technically, three challenges can be derived from RCooper: *1) Data heterogeneity*. Considering the construction cost, various types of sensors (multiline LiDAR, MEMS LiDAR, and camera) are employed in practice, leading to prominent data heterogeneity for cooperative perception [4, 15]. *2) Cooperative representation merits further enhancement*. Most existing cooperative perception approaches, including well-explored vehicle-vehicle (V2V) and vehicle-infrastructure (V2I) cooperation, are designed for vehicle-centric cooperative tasks [24, 28, 31, 34, 39, 47, 51]. However, to our knowledge, the roadside cooperation approaches have not been investigated before. The inherent characteristics of roadside sensors (like roll, pitch angle, height) make roadside cooperative representation a different playground compared with vehicle-centric cooperation [43–45, 52]. *3) Perception performance needs improvements*. How to achieve high-quality downstream perception tasks based on roadside cooperative representation, e.g., detection, tracking, counting, and monitoring, needs further investigation. For example, tracking by unstable detection results in complex intersection scenes is still challenging. It is necessary to delve into these challenges, but the lack of datasets hinders the exploration of the Garden of Eden.

We hence release the first real-world, large-scale dataset RCooper for this challenging field to open the gate and boost the development of roadside cooperative perception. We follow the installation scheme widely adopted in practical applications. The complex roadside system is boiled down to two typical roadside settings, i.e., intersection and corridor. We select several representative locations with different levels of traffic flow for each traffic scene, resulting in a manually annotated dataset comprising 50k images and 30k point clouds, which covers diverse weather and lighting variations.

Our contributions are summarized as follows:

- The first real-world, large-scale dataset, RCooper, is released to bloom research on roadside cooperative perception for practical applications. All the frames and scenes are captured in real-world scenarios.
- More than 50k images and 30k point clouds manually annotated with 3D bounding boxes and trajectories for ten semantic classes are provided in our RCooper, which enables the training and evaluation of roadside cooperative perception approaches in real-world scenarios.
- Two cooperative perception tasks, including 3D object detection and tracking, are introduced, and comprehensive benchmarks with SOTA methods are reported. The results show the effectiveness of roadside cooperation and demonstrate the direction of further research.

2. Related work

This section briefly reviews three related topics: roadside perception, cooperative perception, and public perception datasets in roadside systems.

Dataset type	Source	Dataset	Year	Coop Mode	RGBs	LiDARs	Catagories	Det task	Trk task
Roadside	real	BoxCars [32]	2018	None	116k	-	12	-	-
		BAAI-VANJEE [11]	2021	None	5k	2.5k	12	3D	-
		Rope3D [45]	2021	None	50k	50k	13	3D	-
		Mona [17]	2022	None	11.7M	-	2	2D	2D
		A9 [9]	2022	None	5.4k	5.3k	10	3D	3D
Cooperative	sim	CODD [1]	2021	V2V	-	13.5k	2	-	-
		OPV2V [38]	2022	V2V	44k	11k	1	3D	-
		V2X-Sim [26]	2022	V2X	60k	10k	1	3D	3D
		V2XSet [37]	2022	V2X	44k	11k	1	3D	-
		DOLPHINS [30]	2022	V2X	127k	84k	3	2D & 3D	-
	real	DAIR-V2X [46]	2022	V2X	39k	39k	10	3D	-
		V2V4Real [41]	2023	V2V	40k	20k	5	3D	3D
Roadside Cooperative	**real**	**Rcooper (Ours)**	**2024**	**Roadside**	**50k**	**30k**	**10**	**3D**	**3D**

Table 1. Comparisons among the representative public perception dataset for road systems.

2.1. Roadside Perception

Benefiting from the release of roadside public datasets, such as Rope3D [45] and DAIR-V2X-I [46], several pioneer roadside perception methods have emerged in recent years. A simple and effective attempt utilizing camera specifications and the ground knowledge is proposed along with the dataset in Rope3D [45]. MonoGAE [43] further proposes a ground-aware embedding to integrate implicit roadside ground information with high-dimensional semantic features. BEVHeight [44] discovers the importance of predicting the height to the ground to ease the optimization process of camera-based roadside perception, and the follow-up work BEVHeight++ [42] further enhances the performance by fusing the height and depth representation. Considering the practical challenges of calibration noises, CBR [13] achieves calibration-free roadside perception via decoupled feature reconstruction. Limited by the available dataset, existing methods do their utmost to pursue better and more robust perception performance with independent roadside sensor systems. However, 'two *eyes* are better than one,' we believe the cross-infrastructure cooperation can further boost the roadside perception performance.

2.2. Cooperative Perception

According to the collaboration stage, cooperative perception can be divided into early, intermediate, and late fusion [3, 6, 18]. Since the 3D point cloud has inherent aggregation convenience, early fusion approaches usually adopt LiDAR as a sensor [2, 8], and collaboration on raw data with comprehensive information makes it always become the upper bound of cooperative perception. However, the massive amount of data also introduces high transmission and computation costs. On the other hand, late fusion is bandwidth-economic for only transferring perception results [2, 46]. The fusion strategy is physically explicable but relies on accurate individual predictions. Recent methods focus more on intermediate fusion to balance the trade-off between performance and cost. They focus on either fusion strategy for better performance [7, 10, 20, 25, 27, 33, 37, 40], or feature selection for transmission efficiency [19, 49]. Unlike scene-level feature cooperation, instance-level query cooperation is proposed for interpretable flexible feature interaction [14]. Compared with well-explored vehicle-centric cooperative perception, the potential of roadside systems has yet to be fully exploited. We introduce a new roadside cooperative perception playground called RCooper.

2.3. Perception Datasets in Road System

The blooming development of data-driven perception for autonomous driving and traffic management has dramatically benefited from abundant public traffic scene datasets. The pioneering work, KITTI [16], released the first well-known dataset for autonomous driving, and nuScenes [5] provided 360° view multimodal data to boost single-vehicle perception research further. To fuel the development in cooperative perception, various multi-agent datasets have emerged in recent years, but most of them are derived from simulators (e.g., CARLA [12] and OpenCDA [36]) due to the difficulty of collecting actual data [1, 2, 26, 30, 37, 38]. DAIR-V2X [46, 48] and V2V4Real [41] are the two large-scale real-world datasets for vehicle-centric cooperative perception, which is significant to practical application in natural scenes. Roadside perception has drawn more attention for its comprehensive perception capability, and several single-infrastructure datasets [9, 11, 17, 32, 45] are publicly available. However, there is a lack of public datasets for roadside cooperative perception. To facilitate the exploration of this exciting and challenging field, we release the first real-world, large-scale dataset, RCooper, in this paper. The comparisons among the representative public dataset for perception in road systems are reported in Tab. 1.

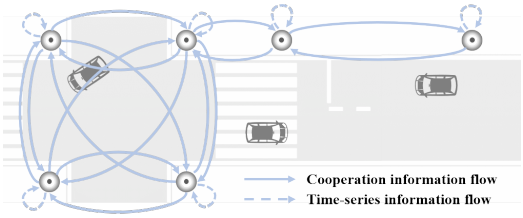


Figure 3. RCooper scenes can be modeled as the structural-stable graph, where the infrastructures and connections are regarded as nodes and edges. Two basic units in graphics, line segment and loop, correspond to the corridor and intersection in actual scenes.

3. RCooper Dataset

To advance roadside cooperative perception, we introduce **RCooper**, a real-world, large-scale, multi-modal dataset annotated with 3D bounding boxes and trajectories. We commence with the data acquisition methodology, delineate the annotation process, and present data analysis.

3.1. Data Acquisition

Scenario Selection Figure 3 depicts the road network as a graph with line segments and loops representing corridors and intersections, respectively. These form the basis for two primary traffic scene types. Besides, the dataset is enriched by capturing scenarios across various times and weather conditions throughout the year, ensuring a wide range of environmental and lighting diversities.

Sensor System Design We follow the typical installation scheme of the infrastructure-side sensor system in practical applications, as shown in Fig.4. Considering the characteristics of scenes and construction cost, there are three schemes for the infrastructure agents. Different from the vehicle-side sensor system, 2 LiDARs with different beams (80 beams + 32 beams) are combined as a group for roadside systems since the mounting height makes a single LiDAR can not sense the area directly below it. Three specific installation schemes are illustrated below, shown in Fig.4.

- **2 Cameras + Multiline LiDARs Group** is adopted in corridor scenes. The corridor area is long and narrow, which is difficult for a single agent to cover. Fig. 2 (a.1) shows that the sensing area of the two neighboring LiDAR systems is intersected, which achieves full area coverage. 2 cameras on the same agent are mounted in reverse, and the cameras of neighboring agents can capture the blind spot directly below itself (Fig. 2 (b)).
- **1 Camera + Multiline LiDARs Group** is adopted in intersection scenes. The camera is mounted towards the junction to capture the RGB video, while the multiline LiDAR can sense half the area of the scene. Such 2 agents are placed in opposite to cover the most of area.

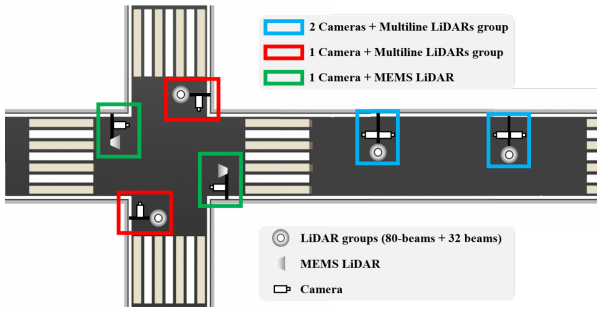


Figure 4. Diagram of the infrastructure-side sensor system. For intersection scenes, we adopt a hybrid scheme for LiDAR-based systems where 2 LiDARs groups (80-beams + 32-beams) and 2 MEMS LiDAR are utilized, because the integration of multiline and MEMS LiDAR is gaining traction for its cost-effectiveness compared to using multiline LiDAR alone. For corridor scenes, each sensor agent include a LiDARs group to cover the region.

- **1 Camera + MEMS LiDAR** is also adopted in intersection scenes to achieve blind area coverage. Compared with corridor scenes, the traffic flow is busier and more complex, and occlusion is more likely to occur. The augment of such setting is of great necessity in intersections.

The detailed parameters are listed in Tab. 2

Data Gathering The most representative 410 scenarios of 15 seconds duration are selected from a vast raw data pool. Sampling frequency is set as 3Hz, resulting in 30K frames of LiDAR point cloud (PC) and 50K frames of RGB images. Each frame of the corridor scene includes 4 RGB images and 2 pre-merged PCs (the PC of the multiline LiDARs group is pre-merged), and that of the intersection scene consists of 4 RGB images and 4 PCs (2 pre-merged PCs and 2 raw PCs of MEMES LiDARs). The synchronization between sensor agents is less than 50ms.

3.2. Coordinates and Data Annotation

Coordinate System There are three different coordinate systems of our RCooper, i.e., the LiDAR coordinate sys-

Sensors		Details
Camera	RGB	42Hz, 1920×1200
LiDAR	80-beams	10Hz, 360° horizontal FOV, -25° to 15° vertical FOV, 1m to 230m capture range, $\pm 3cm$ error
	32-beams	10Hz, 360° horizontal FOV, 90° vertical FOV, 0.1m to 30m capture range, $\pm 3cm$ error
	MEMS	10Hz, $\pm 60^\circ$ horizontal FOV, $\pm 12.5^\circ$ vertical FOV, 0.5m to 200m capture range, $\pm 3cm$ error

Table 2. Sensor specifications in RCooper.

tem, the camera coordinate system, and the world coordinate system. The LiDAR coordinate system is regarded as the bridge, and we provide LiDAR-to-Camera and LiDAR-to-World calibration parameters for each frame. Besides, we annotate the 3D bounding boxes separately based on each infrastructure's LiDAR coordinate system such that each agent's sensor data alone can also be treated as independent roadside view perception tasks. The relative position between the two infrastructures in the same scene is mapped via the world coordinate system, and the system's origin is a virtual point of the local map.

Labeling Approach We adopt a 3-steps labeling approach, including manual labeling for single infrastructure annotations, automatic labeling for cooperative annotations, and a final manual refinement step. We employ groups of professional annotators, and they exhaustively label each object in PC with the 7-degree-of-freedom 3D bounding box containing x, y, z for centric location and l, w, h, yaw for bounding box extent and orientation. There are ten semantic classes in total, belonging to five major classes, i.e., Vehicles (*car, bus, truck, and huge_vehicle*), Cyclists (*bicycle, tricycle, and motorcycle*), Pedestrians, and Constructions (*traffic_sign and construction*). Each annotated object is assigned a unique object ID for the tracking task, and the object ID of the same object in one sequence is unique even when it is wholly occluded in some frames. To automatically generate the cooperative annotations based on the independent labels, we transform the objects from different LiDAR coordinates to the unified world coordinates and match the 3D bounding boxes via the Hungarian algorithm with Euclidean distance. We assign the same object ID for the matched objects and refine the bounding box according to all the corresponding annotations. If there is no matching object, it is introduced as a complement. Finally, we manually supervise and adjust the cooperative annotations and object IDs to obtain more accurate annotations. In addition, the whole dataset is desensitized before public release.

3.3. Statistic and Scene Analysis

Data Statistics It can be observed from Fig. 5 that most of the objects (60%) in RCooper belong to the Car class, and the other three vehicle classes take up 10% in corridor scenes and 20% in intersection scenes. The cyclists class ranks second, and the motorcycle has the most quantities. Since we focus more on vehicles than other road users, the ratio of pedestrians is limited. Besides, there are more construction labels in the intersection scene.

Difference Analysis between Typical Scenes There are two main differences between these two typical scenes:

- **Spatial distribution of data** differs due to the topological characteristics. The corridor is long and narrow, while the intersection is spatial-wise centralized. As shown in Fig. 2(a.1), the cross-infrastructure data in corridor scenes is more clearly distributed in space, extending the sensing range and complementing the blind spot. Differently, the data from multiple infrastructures entwine each other and enhance the observation from various views in complex intersection scenes. Therefore, observation complementary is dominant in corridors, and the observation enhancement is dominant in intersections.
- **Data heterogeneity in intersection scenes** is the specific challenge in practical due to the hybrid-type of LiDAR systems as described in Fig. 4. Both multiline and MEMS LiDARs (with differing operating principles) are employed considering the construction cost, introducing severe data heterogeneity, which may result in failures of the existing cooperative perception methods (Tab. 4).

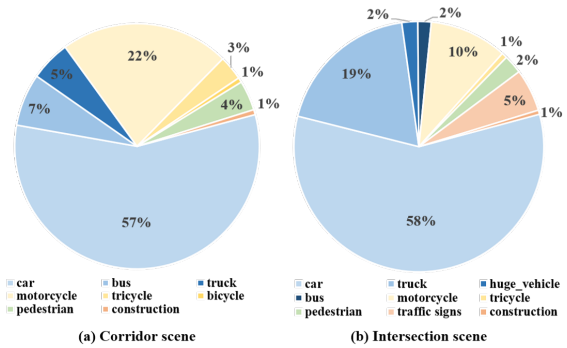


Figure 5. The distribution of semantic classes.

The abovementioned differences make our dataset an interesting but challenging playground. Not only the specialized approach for each scene, but also a unified approach for the entire roadside system, demand prompt solutions.

4. Task

Our dataset or its extension could further support multiple cooperative perception tasks, including detection, tracking, prediction, localization, counting, monitoring, etc. This paper focuses on cooperative detection and tracking tasks, i.e., integrating cross-infrastructure information to localize, recognize, and track 3D objects at a fixed traffic scene.

4.1. Task Overview

Theoretically, the task in this paper can be boiled down to two main sequential sub-tasks: roadside cooperative 3D detection and roadside 3D tracking. A straightforward collaborative approach is to achieve result-level fusion, i.e., late fusion, which merges multi-view detection results to the final detection results. The mainstream cooperative 3D detection framework employs early or intermediate fusion techniques, which first encode the roadside cooperative representation and employ a 3D detector to output the coordi-

nates of 3D bounding boxes and object categories. Based on detected 3D bounding boxes, tracking by detection framework matches the objects and forms the trajectory ID. Two sub-tasks are illustrated in detail below.

4.2. Roadside Cooperative Detection

Task Description The roadside cooperative detection task requires leveraging multiple LiDAR views to perform 3D object detection toward the corresponding district. Compared to single-view roadside detection, roadside cooperative detection has these challenges: data heterogeneity as mentioned before, and cooperative representation merits further enhancement. Our current benchmark does not aim to address these challenging issues absolutely but instead seeks to demonstrate the existence of these gaps and pave the way for subsequent research.

Input and Groundtruth The input of roadside cooperative detection comprises multi-agent sequential frames and their relative posture. Take infrastructure node i of a traffic region graph $\mathcal{G}$ as an example:

- Sequential frames $\{C_i(t_i^{'})|t_i^{'} \leq T_i\}$ of node i and $\{C_{N_i}(t_{N_i}^{'})|t_{N_i}^{'} \leq T_{N_i}\}$ of neighbors, where T_i is the perception moment, $T_{N_i} \leq T_i$ is the capturing moments of neighbors, and $C(\cdot)$ denotes capturing function.
- Relative posture M_i and M_{N_i}.

The perception outputs are the detected objects in the fixed traffic region, usually consisting of coordinates of detected 3D bounding boxes and the confidence score of the object category. Correspondingly, the groundtruth is the set of objects appearing in the region anytime and anywhere, which can be formulated as $GT = (GT_i \cup GT_{N_i}) \cap R$. GT_i and GT_{N_i} are groundtruth from node i and neighbors, and R is the interested region at fixed location.

Benchmark Methodology Most commonly adopted four fusion strategies are employed for roadside cooperative perception with state-of-the-art cooperative methods.

- No Fusion: Only the single LiDAR point clouds are employed for detection, which is the baseline for comparing cooperative and non-cooperative methods.
- Late Fusion: 3D objects are detected for each LiDAR utilizing its sensor observations. Then, non-maximum suppression is adopted to merge and produce final outputs.
- Early Fusion: All the point clouds from multiple LiDARs are aggregated to form a more comprehensive point cloud, which can preserve complete information. Then, obey the no-fusion pipeline to generate detection results.
- Intermediate Fusion: The point clouds from each LiDAR are projected to a selected coordinate system and then are fed into neural feature extractors to encode intermediate features. Afterward, the encoded features are merged for cooperative feature fusion. Our benchmark employs several representative intermediate fusion methods, including AttFuse[38], F-Cooper[7], Where2Comm[19], and CoBEVT[40].

Evaluation Metrics $400m \times 400m$ areas for fixed traffic scenes are chosen for perception evaluation. The common detection metric AP is used for 3D object detection evaluation. Specifically, AP values under different 3D IoU thresholds are reported for a more comprehensive evaluation.

4.3. Roadside Cooperative Tracking

Task Description The roadside cooperative tracking task is expected to show the superiority of the roadside cooperative temporal perception. There are two typical object tracking modes: joint detection and tracking, and tracking by detection. We concentrate on the latter in this paper.

Input and Groundtruth The input of roadside cooperative tracking is the roadside cooperative detection prediction, including coordinates of detected 3D bounding boxes and the confidence score of the object category. Moreover, the groundtruth is the correlation between trajectory IDs and object IDs.

Benchmark Methodology We follow the previous works [41, 48], and implement AB3Dmot tracker [35] in our benchmark. Based on the predictions from the cooperative detection models, the 3D Kalman filter and the Hungarian algorithm are employed by the AB3Dmot tracker to achieve efficient and high-quality tracking.

Evaluation Metrics The same evaluation metrics in [5] and [35] are adopted for roadside cooperative tracking evaluation, including 1) average multi-object tracking accuracy (AMOTA), 2) average multi-object tracking precision (AMOTP), and 3) scaled average multi-object tracking accuracy (sAMOTA), 4) multi-object tracking accuracy (MOTA), 5) mostly tracked trajectories (MT), and 6) mostly lost trajectories (ML).

5. Benchmark Experiments

In this section, we build two benchmarks respectively for roadside cooperative detection and tracking, expecting to provide effective and competitive benchmarks and pave the way for subsequent research.

5.1. Implementation Details

The dataset is split into the train/validation set as 4:1 (ratio of scenario), respectively. Consistent with [41], different categories are merged as the same class. For the RCooper

Method	AP@0.3	AP@0.5	AP@0.7
No Fusion	40.0	29.2	11.1
Late fusion	44.5	29.9	10.8
Early fusion	**69.8**	54.7	30.3
AttFuse [38]	62.7	51.6	32.1
F-Cooper [7]	65.9	55.8	36.1
Where2Comm [19]	67.1	55.6	34.3
CoBEVT [40]	67.6	**57.2**	**36.2**

Table 3. Roadside cooperative detection benchmark of the corridor scene (%).

Method	AP@0.3	AP@0.5	AP@0.7
No Fusion	58.1	44.1	23.8
Late fusion	**65.1**	**47.6**	24.4
Early fusion	50.0	33.9	18.3
AttFuse [38]	45.5	40.9	27.9
F-Cooper [7]	49.5	32.0	12.9
Where2Comm [19]	50.5	42.2	29.9
CoBEVT [40]	53.5	45.6	**32.6**

Table 4. Roadside cooperative detection benchmark of the intersection scene (%).

detection task, PointPillar [23] is adopted as the backbone of all models to extract features from the points cloud. They are trained for 50 epochs with a batch size of 16. The initial learning rate is set as 2×10^{-3} and is scheduled according to cosine annealing [29]. Adam optimizer [22] is adopted with a weight decay of 1×10^{-4}. For the tracking task, $F_{min} = 1$ and $Age_{max} = 2$ are set for the birth and death memory module according to the label criteria for the trajectory with shelters. In the data association module, we use $GIoU3D_{min} = -0.2$ as the threshold to filter the matching, consistent with the original AB3Dmot tracker [35].

5.2. Roadside Cooperative Detection Results

The roadside cooperative detection benchmark results for corridor scenes and intersection scenes are reported in Tab. 3 and Tab. 4, respectively.

It can be seen from Tab. 3 that all the cooperative approaches perform better than the no-fusion one, which aligns with our expectations. The cross-infrastructure cooperation in corridor scenes acts in a typical mode to achieve a comprehensive understanding of the long and narrow traffic area. Benefiting from the extension of the sensing range, the performance improvement is significant. The LiDAR-based CoBEVT [40] achieves the best performance in AP@0.5 and AP@0.7, while early fusion method achieves the best performance in AP@0.3. The corridor scenes provide a typical playground for cooperative detection research but from the roadside view.

Experimental results reported in Tab. 4 show the detection performance in intersection scenes. Unlike the supplementary role in corridor scenes, cooperation at intersections is expected to learn a cooperative representation with observation from various views to understand the complex traffic scenario better. Note that the late-fusion achieves the best performance on $AP@0.3$ and $AP@0.5$ among all the cooperative methods, and the early fusion even performs worse than the no-fusion method. As for the intermediate fusion, CoBEVT [40] performs better on AP@0.7. Considering the data heterogeneity challenge in intersection scenes, the abovementioned phenomenon is unexpected but understandable. ***Firstly***, the impact of data heterogeneity is more evident for early fusion since the simple integration (without fine-grained design) of data cannot cope with the heterogeneity problem and even makes the data distribution more complex to model, resulting in worse performance compared with the no-fusion method. ***Secondly***, late fusion theoretically overcomes the problem via the result-level fusion with bounding boxes, and the performance advantages demonstrate the effectiveness of cooperative perception at intersections. The intermediate fusion approach can deal with the challenge to some extent by feature-level cooperation, and the advantages of cooperative representation lead to a better performance at a higher IoU threshold, i.e., $AP@0.7$. To further explore the impact of data heterogeneity, a data-level ablation study is reported in Fig. 6. The excluding of point clouds from MEMS LiDAR directly makes the cooperative approaches act in a typical manner. Therefore, further research can utilize RCooper to study how to fully leverage the sensing data and overcome the heterogeneity challenge in practical scenes. Referring to multimodal learning techs, we think a possible way lies in encoding heterogeneous data into a unified feature space by feature extraction incorporating distribution consistency constraints, such as aligning distribution via KL divergence.

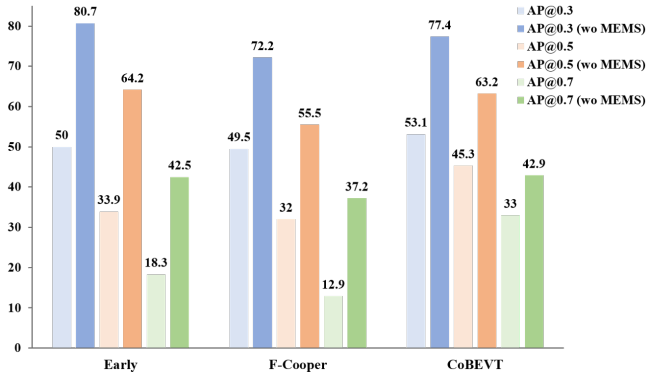


Figure 6. Influences of data heterogeneity. The performance of both early fusion and two representative feature fusion methods is boosted enormously when the two MEMS LiDARs are discarded, even if the number of collaborators decreases.

Apart from the failure of SOTA methods in intersection

Method	AMOTA ↑	AMOTP ↑	sAMOTA ↑	MOTA ↑	MT ↑	ML ↓
No Fusion	8.28	22.74	34.05	23.89	17.34	42.71
Late fusion	9.60	25.77	35.64	24.75	24.37	42.96
Early fusion	**23.78**	**38.18**	59.16	44.30	**53.02**	**12.81**
AttFuse [38]	21.75	35.31	57.43	44.50	45.73	22.86
F-Cooper [7]	22.47	35.54	58.49	45.94	47.74	22.11
Where2Comm [19]	22.55	36.21	**59.60**	46.11	50.00	19.60
CoBEVT [40]	21.54	35.69	53.85	**47.32**	47.24	18.09

Table 5. Roadside cooperative detection benchmark of the corridor scene (%).

Method	AMOTA ↑	AMOTP ↑	sAMOTA ↑	MOTA ↑	MT ↑	ML ↓
No Fusion	18.11	39.71	58.29	49.16	35.32	41.64
Late fusion	**21.57**	43.40	**63.02**	**50.58**	**42.75**	**34.20**
Early fusion	21.38	**47.71**	62.93	50.15	36.80	42.75
AttFuse [38]	11.84	36.63	46.92	39.32	29.00	53.90
F-Cooper [7]	-4.86	14.71	0.00	-45.66	11.52	50.56
Where2Comm [19]	14.21	38.48	50.97	42.27	29.00	45.72
CoBEVT [40]	14.82	38.71	49.04	44.67	33.83	35.69

Table 6. Roadside cooperative tracking benchmark of the intersection scene (%).

scenes, some SOTA methods are not as effective as simple fusion method (early or late) in both scenes. Another reason may lie in the scenario gap. Some SOTA Methods designed for vehicle-centric scenarios, which can leverage vigorous development in vehicle-side perception, struggle with infrastructure-specific challenges (e.g., larger variations of mounting heights and pitch angles compared with vehicle-side) [44, 45]. Specific methods for roadside cooperative perception deserve further investigation.

5.3. Roadside Cooperative Tracking Results

The roadside cooperative tracking benchmark results for corridor scenes and intersection scenes are reported in Tab. 5 and Tab. 6, respectively.

For corridor scenes, the cooperative tracking performance is better than that of the no-fusion method, demonstrating the effectiveness of roadside cooperative temporal perception, as shown in Tab. 5. Compared methods present fierce competition: early fusion method achieves the best performance in AMOTA, AMOTP, MT, and ML metrics, Where2Comm [19] achieves the best performance in sAMOTA metric, while CoBEVT [40] achieves the best performance in MOTA metric.

For intersection scenes, the late-fusion strategy outperforms others, as shown in Tab. 6. Since the detection predictions are affected by the data heterogeneity, the experimental results of AB3Dmot present a similar pattern as the detection. Besides, the tracking results also depend on the temporal-wise continuity of detection predictions, so it cannot generate satisfactory trajectories if the instance is not detected stably in adjacent frames, which results in the worse performance of F-Cooper (whose MT value reduces to 11.52%). The tracking-by-detection strategy is susceptible to the detection performance. How to learn a roadside cooperative representation for the tracking task and how to leverage the spatial-temporal contexts in an end-to-end manner in the roadside scenes need further exploration.

6. Conclusion

To extend the boundaries of both autonomous driving and traffic management, **R**oadside **Coo**perative **Per**ception **(RCooper)**, a real-world, large-scale dataset, is released in this paper, which is expected to boost roadside cooperative perception towards round-the-clock area-coverage perception for a restricted traffic area. Our dataset consists of 30K manually annotated sensor data groups (images and point-cloud) covering two typical traffic scenes (i.e., intersection and corridor). A competitive benchmark is constructed to pave the way for subsequent research, and the experimental results demonstrate the effectiveness of roadside cooperation perception and reveal the direction of future work.

Limitation. In this paper, we release the RCooper and build the corresponding benchmarks with several representative approaches. Although it opens the gate, the Garden of Eden still deserves further exploration. As the aforementioned discussion, foreseeable future works include: 1) learning a unified roadside cooperative representation to not only overcome the data heterogeneity challenges but also leverage spatial-temporal multimodal sensor data; 2) exploring a unified approach for end-to-end perception and other tasks; 3) finding a solution for practical challenges, like calibration noises, leveraging the advantages of cross-infrastructure cooperation.

References

[1] Eduardo Arnold, Sajjad Mozaffari, and Mehrdad Dianati. Fast and robust registration of partially overlapping point clouds. *IEEE Robotics and Automation Letters (RAL)*, 7(2): 1502–1509, 2021.

[2] Eduardo Arnold, Mehrdad Dianati, Robert de Temple, and Saber Fallah. Cooperative perception for 3d object detection in driving scenarios using infrastructure sensors. *IEEE Transactions on Intelligent Transportation Systems (TITS)*, 23(3):1852–1864, 2022.

[3] Zhengwei Bai, Guoyuan Wu, Matthew J Barth, Yongkang Liu, Emrah Akin Sisbot, Kentaro Oguchi, and Zhitong Huang. A survey and framework of cooperative perception: From heterogeneous singleton to hierarchical cooperation. *arXiv preprint arXiv:2208.10590*, 2022.

[4] Zhengwei Bai, Guoyuan Wu, Matthew J Barth, Yongkang Liu, Emrah Akin Sisbot, and Kentaro Oguchi. Vinet: Lightweight, scalable, and heterogeneous cooperative perception for 3d object detection. *Mechanical Systems and Signal Processing*, 204:110723, 2023.

[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 multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11621–11631, 2020.

[6] Antoine Caillot, Safa Ouerghi, Pascal Vasseur, Rémi Boutteau, and Yohan Dupuis. Survey on cooperative perception in an automotive context. *IEEE Transactions on Intelligent Transportation Systems (TITS)*, 23(9):14204–14223, 2022.

[7] Qi Chen, Xu Ma, Sihai Tang, Jingda Guo, Qing Yang, and Song Fu. F-cooper: Feature based cooperative perception for autonomous vehicle edge computing system using 3d point clouds. In *Proceedings of the ACM/IEEE Symposium on Edge Computing*, pages 88–100, 2019.

[8] Qi Chen, Sihai Tang, Qing Yang, and Song Fu. Cooper: Cooperative perception for connected autonomous vehicles based on 3d point clouds. In *IEEE International Conference on Distributed Computing Systems*, pages 514–524, 2019.

[9] Christian Creß, Walter Zimmer, Leah Strand, Maximilian Fortkord, Siyi Dai, Venkatnarayanan Lakshminarasimhan, and Alois Knoll. A9-dataset: Multi-sensor infrastructure-based dataset for mobility research. In *IEEE Intelligent Vehicles Symposium (IV)*, pages 965–970, 2022.

[10] Jiaxun Cui, Hang Qiu, Dian Chen, Peter Stone, and Yuke Zhu. Coopernaut: end-to-end driving with cooperative perception for networked vehicles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17252–17262, 2022.

[11] Yongqiang Deng, Dengjiang Wang, Gang Cao, Bing Ma, Xijia Guan, Yajun Wang, Jianchao Liu, Yanming Fang, and Juanjuan Li. Baai-vanjee roadside dataset: Towards the connected automated vehicle highway technologies in challenging environments of china. *arXiv preprint arXiv:2105.14370*, 2021.

[12] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. CARLA: An open urban driving simulator. In *Conference on robot learning (CoRL)*, pages 1–16, 2017.

[13] Siqi Fan, Zhe Wang, Xiaoliang Huo, Yan Wang, and Jingjing Liu. Calibration-free bev representation for infrastructure perception. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 9008–9013, 2023.

[14] Siqi Fan, Haibao Yu, Wenxian Yang, Jirui Yuan, and Zaiqing Nie. QUEST: Query stream for vehicle-infrastructure cooperative perception. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2024.

[15] Jin Fang, Dingfu Zhou, Jingjing Zhao, Chulin Tang, Cheng-Zhong Xu, and Liangjun Zhang. Lidar-cs dataset: Lidar point cloud dataset with cross-sensors for 3d object detection. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2024.

[16] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3354–3361, 2012.

[17] Luis Gressenbuch, Klemens Esterle, Tobias Kessler, and Matthias Althoff. Mona: The munich motion dataset of natural driving. In *IEEE International Intelligent Transportation Systems Conference (ITSC)*, pages 2093–2100, 2022.

[18] Yushan Han, Hui Zhang, Huifang Li, Yi Jin, Congyan Lang, and Yidong Li. Collaborative perception in autonomous driving: Methods, datasets and challenges. *IEEE Intelligent Transportation Systems Magazine*, 15(6):131–151, 2023.

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

[20] Yue Hu, Yifan Lu, Runsheng Xu, Weidi Xie, Siheng Chen, and Yanfeng Wang. Collaboration helps camera overtake lidar in 3d detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9243–9252, 2023.

[21] Lei Huang and Wenzhun Huang. RD-YOLO: an effective and efficient object detector for roadside perception system. *Sensors*, 22(21):8097, 2022.

[22] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2014.

[23] 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 (CVPR)*, pages 12697–12705, 2019.

[24] Zixing Lei, Shunli Ren, Yue Hu, Wenjun Zhang, and Siheng Chen. Latency-aware collaborative perception. In *European Conference on Computer Vision (ECCV)*, pages 316–332, 2022.

[25] Yiming Li, Shunli Ren, Pengxiang Wu, Siheng Chen, Chen Feng, and Wenjun Zhang. Learning distilled collaboration graph for multi-agent perception. In *Conference on Neural Information Processing Systems (NeurIPS)*, pages 29541–29552, 2021.

[26] Yiming Li, Dekun Ma, Ziyan 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 (RAL)*, 7(4): 10914–10921, 2022.

[27] Yencheng Liu, Junjiao Tian, Nathaniel Glaser, and Zsolt Kira. When2com: Multi-agent perception via communication graph grouping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4106–4115, 2020.

[28] Yencheng Liu, Junjiao Tian, Chih-Yao Ma, Nathan Glaser, Chia-Wen Kuo, and Zsolt Kira. Who2com: Collaborative perception via learnable handshake communication. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 6876–6883, 2020.

[29] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2018.

[30] Ruiqing Mao, Jingyu Guo, Yukuan Jia, Yuxuan Sun, Sheng Zhou, and Zhisheng Niu. DOLPHINS: Dataset for collaborative perception enabled harmonious and interconnected self-driving. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, pages 4361–4377, 2022.

[31] Donghao Qiao and Farhana Zulkernine. Adaptive feature fusion for cooperative perception using lidar point clouds. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1186–1195, 2023.

[32] Jakub Sochor, Jakub Špaňhel, and Adam Herout. Boxcars: Improving fine-grained recognition of vehicles using 3-d bounding boxes in traffic surveillance. *IEEE Transactions on Intelligent Transportation Systems (TITS)*, 20(1):97–108, 2018.

[33] Tsun-Hsuan Wang, Sivabalan Manivasagam, Ming Liang, Bin Yang, Wenyuan Zeng, and Raquel Urtasun. V2vnet: Vehicle-to-vehicle communication for joint perception and prediction. In *European Conference on Computer Vision (ECCV)*, pages 605–621, 2020.

[34] Zhe Wang, Siqi Fan, Xiaoliang Huo, Tongda Xu, Yan Wang, Jingjing Liu, Yilun Chen, and Yaqin Zhang. Emiff: Enhanced multi-scale image feature fusion for vehicle-infrastructure cooperative 3d object detection. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2024.

[35] Xinshuo Weng, Jianren Wang, David Held, and Kris Kitani. 3d multi-object tracking: A baseline and new evaluation metrics. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 10359–10366, 2020.

[36] 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 *IEEE International Intelligent Transportation Systems Conference (ITSC)*, pages 1155–1162, 2021.

[37] 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 (ECCV)*, pages 107–124, 2022.

[38] 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 *IEEE International Conference on Robotics and Automation (ICRA)*, pages 2583–2589, 2022.

[39] Runsheng Xu, Weizhe Chen, Hao Xiang, Xin Xia, Lantao Liu, and Jiaqi Ma. Model-agnostic multi-agent perception framework. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 1471–1478, 2023.

[40] Runsheng Xu, Zhengzhong Tu, Hao Xiang, Wei Shao, Bolei Zhou, and Jiaqi Ma. CoBEVT: Cooperative bird's eye view semantic segmentation with sparse transformers. In *Conference on robot learning (CoRL)*, pages 989–1000, 2023.

[41] 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 (CVPR)*, pages 13712–13722, 2023.

[42] Lei Yang, Tao Tang, Jun Li, Peng Chen, Kun Yuan, Li Wang, Yi Huang, Xinyu Zhang, and Kaicheng Yu. BEVHeight++: Toward robust visual centric 3d object detection. *arXiv preprint arXiv:2309.16179*, 2023.

[43] Lei Yang, Jiaxin Yu, Xinyu Zhang, Jun Li, Li Wang, Yi Huang, Chuang Zhang, Hong Wang, and Yiming Li. MonoGAE: Roadside monocular 3d object detection with ground-aware embeddings. *arXiv preprint arXiv:2310.00400*, 2023.

[44] Lei Yang, Kaicheng Yu, Tao Tang, Jun Li, Kun Yuan, Li Wang, Xinyu Zhang, and Peng Chen. BEVHeight: A robust framework for vision-based roadside 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21611–21620, 2023.

[45] 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 (CVPR)*, pages 21341–21350, 2022.

[46] Haibao Yu, Yizhen Luo, Mao Shu, Yiyi Huo, Zebang Yang, Yifeng Shi, Zhenglong Guo, Hanyu Li, Xing Hu, Jirui Yuan, and Zaiqing Nie. 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 (CVPR)*, pages 21361–21370, 2022.

[47] Haibao Yu, Yingjuan Tang, Enze Xie, Jilei Mao, Jirui Yuan, Ping Luo, and Zaiqing Nie. Vehicle-infrastructure cooperative 3d object detection via feature flow prediction. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

[48] 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 (CVPR)*, pages 5486–5495, 2023.

[49] Yunshuang Yuan, Hao Cheng, and Monika Sester. Keypoints-based deep feature fusion for cooperative vehicle detection of autonomous driving. *IEEE Robotics and Automation Letters (RAL)*, 7(2):3054–3061, 2022.

[50] Junping Zhang, Fei-Yue Wang, Kunfeng Wang, Wei-Hua Lin, Xin Xu, and Cheng Chen. Data-driven intelligent transportation systems: A survey. *IEEE Transactions on Intelligent Transportation Systems (TITS)*, 12(4):1624–1639, 2011.

[51] Yang Zhou, Jiuhong Xiao, Yue Zhou, and Giuseppe Loianno. Multi-robot collaborative perception with graph neural networks. *IEEE Robotics and Automation Letters (RAL)*, 7(2): 2289–2296, 2022.

[52] Walter Zimmer, Joseph Birkner, Marcel Brucker, Huu Tung Nguyen, Stefan Petrovski, Bohan Wang, and Alois C Knoll. Infradet3d: Multi-modal 3d object detection based on roadside infrastructure camera and lidar sensors. In *IEEE Intelligent Vehicles Symposium (IV)*, pages 1–8, 2023.

[53] Walter Zimmer, Christian Creß, Huu Tung Nguyen, and Alois C Knoll. Tumtraf intersection dataset: All you need for urban 3d camera-lidar roadside perception. In *IEEE International Intelligent Transportation Systems Conference (ITSC)*, pages 1030–1037, 2023.