UrbanIng-V2X: A Large-Scale Multi-Vehicle, Multi-Infrastructure Dataset Across Multiple Intersections for Cooperative Perception

Karthikeyan Chandra Sekaran[1*] **Markus Geisler**[1*] **Dominik Rößle**[1*]

Adithya Mohan[1] **Daniel Cremers**[2] **Wolfgang Utschick**[2]

Michael Botsch[1] **Werner Huber**[1] **Torsten Schön**[1]

[1]**Technische Hochschule Ingolstadt** [2]**Technical University of Munich**

Abstract

Recent cooperative perception datasets have played a crucial role in advancing smart mobility applications by enabling information exchange between intelligent agents, helping to overcome challenges such as occlusions and improving overall scene understanding. While some existing real-world datasets incorporate both vehicle-to-vehicle and vehicle-to-infrastructure interactions, they are typically limited to a single intersection or a single vehicle. A comprehensive perception dataset featuring multiple connected vehicles and infrastructure sensors across several intersections remains unavailable, limiting the benchmarking of algorithms in diverse traffic environments. Consequently, overfitting can occur, and models may demonstrate misleadingly high performance due to similar intersection layouts and traffic participant behavior. To address this gap, we introduce UrbanIng-V2X, the first large-scale, multi-modal dataset supporting cooperative perception involving vehicles and infrastructure sensors deployed across three urban intersections in Ingolstadt, Germany. UrbanIng-V2X consists of 34 temporally aligned and spatially calibrated sensor sequences, each lasting 20 seconds. All sequences contain recordings from one of three intersections, involving two vehicles and up to three infrastructure-mounted sensor poles operating in coordinated scenarios. In total, UrbanIng-V2X provides data from 12 vehicle-mounted RGB cameras, 2 vehicle LiDARs, 17 infrastructure thermal cameras, and 12 infrastructure LiDARs. All sequences are annotated at a frequency of 10 Hz with 3D bounding boxes spanning 13 object classes, resulting in approximately 712k annotated instances across the dataset. We provide comprehensive evaluations using state-of-the-art cooperative perception methods and publicly release the codebase, dataset, HD map, and a digital twin of the complete data collection environment via `https://github.com/thi-ad/UrbanIng-V2X`.

1 Introduction

Reliable perception is fundamental to autonomous driving, particularly in complex urban intersections where a comprehensive understanding of the scene is essential for safe decision-making [1, 35, 2]. In such environments, single-agent systems are inherently constrained by their Field-of-View (FOV) and often fail to detect critical objects that are occluded by other vehicles and infrastructure [11]. To

[*]Equal contribution. Authors listed in alphabetical order.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Track on Datasets and Benchmarks.

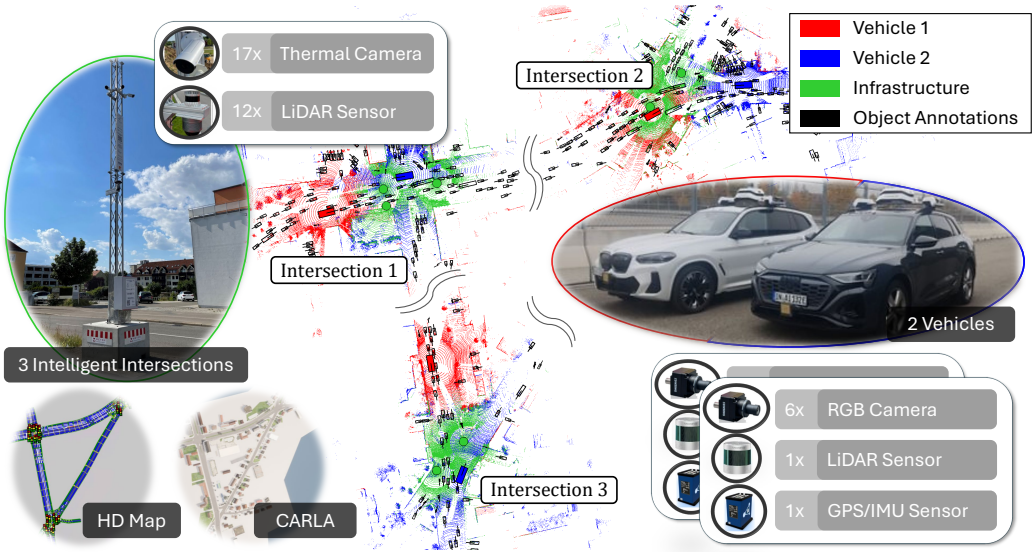


Figure 1: This illustration provides a comprehensive overview of the **UrbanIng-V2X** cooperative perception dataset environment. For each intersection, a globally fused point cloud of a representative scenario is visualized. Point clouds from individual agents are color-coded, highlighting two vehicles and sensor poles at three intersections as cooperation partners. Further, the complete sensor setup, along with a bird's-eye view of both the HD map and a high-fidelity CARLA map, is shown.

address these limitations, recent research has increasingly focused on cooperative perception [21, 32], leveraging information sharing among multiple agents. To advance this paradigm, substantial effort has been invested in the development of real-world Vehicle-to-Everything (V2X) datasets [33, 27, 37], despite significant challenges such as precise temporal and spatial synchronization, high financial and logistical costs, and the complexity of multi-agent hardware setups [12].

However, none of the existing V2X datasets capture urban scenarios involving multi-vehicle, multi-infrastructure setups across multiple intersections. This combination is crucial to assess the scalability and real-world applicability of cooperative perception systems, especially in urban environments where heterogeneous sensor views and varying infrastructure layouts demand robust generalization and reliable performance. To address this gap, we introduce the **UrbanIng-V2X** dataset—a large-scale cooperative perception dataset collected at three distinct urban intersections within the High-Definition Testfield [3, 22]. In this environment, we recorded 34 sequences, each lasting 20 seconds, involving two vehicles and up to three infrastructure-mounted sensor poles cooperating in coordinated scenarios.

The main contributions of UrbanIng-V2X are as follows:

- UrbanIng-V2X is the first real-world cooperative perception dataset featuring multiple vehicles and extensive infrastructure sensing at three distinct urban intersections, see Figure 1.
- UrbanIng-V2X introduces the largest number of cooperating sensors in real-world datasets to date and improves multi-modality by including thermal cameras. Each scenario involves 2 connected vehicles, each with 6 RGB cameras and a rooftop LiDAR, cooperating with up to 6 thermal cameras and 4 LiDARs mounted on 3 infrastructure poles per intersection. All sensors are spatially and temporally aligned, with additional sweep data provided for intermediate frames.
- The dataset supports a broad range of cooperative perception benchmark tasks, including 3D object detection, tracking, trajectory prediction, and localization. All sequences are annotated at 10 Hz with 3D bounding boxes for 13 object classes, totaling approximately 712k annotated instances.
- A comprehensive benchmark evaluation is conducted using state-of-the-art (SOTA) algorithms for cooperative perception. The results highlight key challenges and opportunities across diverse sensor setups and intersection configurations.

- A developer toolkit is provided, including format converters for OpenCOOD [30] and nuScenes [5], for enabling integration with single- and multi-modal perception pipelines.
- The codebase and dataset, including High-Definition maps (HD maps) in Lanelet2 format [20] and a geo-referenced CARLA [7] digital twin to support situation interpretation, synthetic data generation, and domain adaptation research, are publicly available.

2 Related Work

Single Agent Perception Datasets: Large-scale single-vehicle datasets such as KITTI [10], nuScenes [5], and the Waymo Open Dataset [24] have been fundamental in advancing perception tasks. They provide multi-modal sensor data captured from individual vehicles in diverse urban and suburban settings. In contrast, datasets such as LUMPI [4] and the TUMTraf Intersection Dataset [36] offer multi-modal sensor data, collected from infrastructure-mounted sensors. While all these datasets support tasks such as 3D object detection and tracking, they inherently lack multi-agent interactions, limiting their utility in cooperative perception environments.

Cooperative Perception Datasets: Several synthetic datasets such as OPV2V [30], V2XSet [29], and V2X-Sim [17] have been developed in the past to explore cooperative perception with simulated multi-vehicle and infrastructure scenarios. Although these datasets offer flexible and scalable environments, they fail to capture the full complexity and noise characteristics of real-world settings. Consequently, several real-world datasets have emerged to support cooperative perception research. **V2V4Real** [31] enables Vehicle-to-Vehicle (V2V) cooperative perception across diverse driving scenes with rich annotations such as track IDs, but lacks infrastructure sensor data, limiting it to non-V2X scenarios. **DAIR-V2X-C** [33] includes both Vehicle-to-Infrastructure (V2I) data in 28 intersections—the most among existing datasets—but involves only one vehicle and lacks track IDs, HD maps, diverse sensors, and dense urban scenes. While its extension **V2X-Seq** [34] adds track IDs and HD maps for some sequences, the regional availability restrictions of DAIR-V2X-C [33] and V2X-Seq [34] limit international usability. **TUMTraf-V2X** [37] offers labeled single vehicle and infrastructure data at a single intersection with HD maps and day/night coverage. However, its small scale and limited geographic diversity restrict its applicability. **V2X-Real** [27] combines data from two vehicles and infrastructure, but lacks coverage of multiple intersections and HD maps.

We introduce UrbanIng-V2X to fill the gap of missing real-world datasets that cover a combination of multi-agent coordination, multi-modal sensing (including thermal cameras), and multiple intersection layouts. A detailed comparison to the existing datasets is shown in Table 1.

3 Dataset

Two connected vehicles and three smart infrastructures are used for data collection. Each intersection has 2 to 3 sensor poles. Each of these two vehicles is equipped with a high-precision Inertial Measurement Unit (IMU), a 128-ray high-end LiDAR sensor, and six Full High Definition (FHD) RGB cameras oriented in six directions, providing a full 360° FOV. The vehicles also receive Real Time Kinematic (RTK) correction data, achieving localization accuracy up to $1\,\mathrm{cm}$. At each intersection, 2 to 3 poles are installed, each equipped with 1 to 3 VGA thermal cameras. Additionally, six of the seven poles are equipped with a LiDAR setup, comprising a 64-ray midrange LiDAR and a 32-ray short-range blind-spot LiDAR. Detailed sensor descriptions and FOV coverage for infrastructure sensors are provided in the supplementary material.

3.1 Sensor Synchronization

The UTC clock is employed as a unified time reference to synchronize both vehicle- and infrastructure-mounted sensors. The IMU synchronizes to UTC via GPS signals and acts as the Precision Time Protocol (PTP) [14] master within the vehicle. The camera capture card operates as a PTP slave, inheriting the UTC reference from the IMU, while the LiDAR system obtains UTC timestamps independently through a dedicated GPS mouse. Beyond time synchronization, sensor data acquisition is precisely coordinated. The LiDAR is phase-locked so that its zero-degree orientation consistently aligns with integer multiples of its rotation cycle, establishing a deterministic angular reference. Camera triggering is hardware-based and event-driven: instead of simultaneous image capture, each

Table 1: Comparison of real-world cooperative V2X datasets with the proposed UrbanIng-V2X dataset (I=Infrastructure, V=Vehicle). †Images are not published yet.

Property	**V2V4Real [31]**	**DAIR-V2X-C[33]**	**V2X-Seq[34]**	**TUMTraf-V2X[37]**	**V2XReal [27]**	**UrbanIng-V2X (ours)**
Year	2022	2022	2023	2024	2024	2025
V2X	V2V	V2I	V2I	V2I	V2V&I	V2V&I
Intersections	0	28	28	1	1	3
Vehicles	2	1	1	1	2	2
RGB Images	40k†	39k	15k	5k	171k	81.6k
IR Images	0	0	0	0	0	38.8k
LiDAR frames	20k	39k	15k	2k	33k	27.2k
3D Boxes	240k	464k	10.45k	29k	1.2M	712k
Classes	5	10	9	8	10	13
Digital Twin	No	No	No	No	No	Yes
Av. worldwide	Yes	No	No	Yes	Yes	Yes
HD Maps	Yes	No	Yes	Yes	No	Yes
Attributes	No	No	No	Yes	No	Yes
Track IDs	Yes	No	Yes	Yes	Yes	Yes
Traffic light	No	No	No	No	No	Yes
Sensors (I | V)	0 | 8	2 | 3	2 | 3	5 | 4	8 | 12	10 | 16
City	Ohio	Beijing	Beijing	Munich	N.A.	Ingolstadt
Country	USA	China	China	Germany	N.A.	Germany

camera is triggered exactly when the LiDAR beam passes through its FOV. This setup minimizes intermodal latency and ensures high-precision spatio-temporal alignment between LiDAR and camera data.

Each intersection is equipped with UTC-synchronized PTP and Network Time Protocol (NTP) time servers. Thermal cameras are synchronized via the NTP service, while all LiDAR units receive UTC timestamps through dedicated GPS mouses. Similar to the vehicles, infrastructure LiDARs are phase-locked to ensure that their rotational cycles start and end simultaneously across devices. Unlike vehicle-mounted cameras, the infrastructure thermal cameras operate asynchronously in free-run mode and are not externally triggered. Due to the heterogeneous placement and FOV of these sensors, synchronized hardware triggering would not yield optimal alignment across all LiDAR-camera pairs. Instead, during post-processing, the thermal image closest in time to each annotated LiDAR scan is selected. With thermal cameras operating at 30 Frames Per Second (FPS) and LiDARs at 20 FPS, every second LiDAR frame aligns with every third thermal camera frame with a maximum possible temporal misalignment of $16.6\,\mathrm{ms}$, which corresponds to half the cycle time of the thermal cameras.

3.2 Calibration

Intrinsic calibration: All thermal and RGB cameras are calibrated using a checkerboard pattern. Side-mounted vehicle cameras use a fisheye projection model due to their wide FOV, while all other cameras use a standard pinhole model.

Extrinsic calibration: Each sensor on the infrastructure (thermal camera, LiDAR) and vehicle (camera, LiDAR, IMU) has a local coordinate frame. The vehicle coordinate frame is defined at the geometric center of the car. Each intersection uses a fixed GPS location as its local origin, with the coordinate frame aligned to the East-North-Up (ENU) convention. All coordinate systems use a right-handed convention. Figure 2 shows the coordinate systems for the vehicles and intersections.

The extrinsic transformation between the IMU and the vehicle coordinate frame ($\mathbf{T}_{\text{IMU}\rightarrow\text{Vehicle}}$) is precomputed using a total station. The IMU reports its pose in global coordinates via latitude, longitude, altitude, roll, pitch, and yaw, allowing precise computation of the vehicle's global pose. To estimate extrinsic transformation matrices between sensor pairs, we use a cone-shaped calibration target with a highly reflective marker in the center. The cone is placed in multiple positions within the sensors' FOV. Its global position is measured with a $2\,\mathrm{cm}$ precise RTK GPS device (Trimble SP80). The reflective marker is manually annotated in both LiDAR point clouds and in camera images.

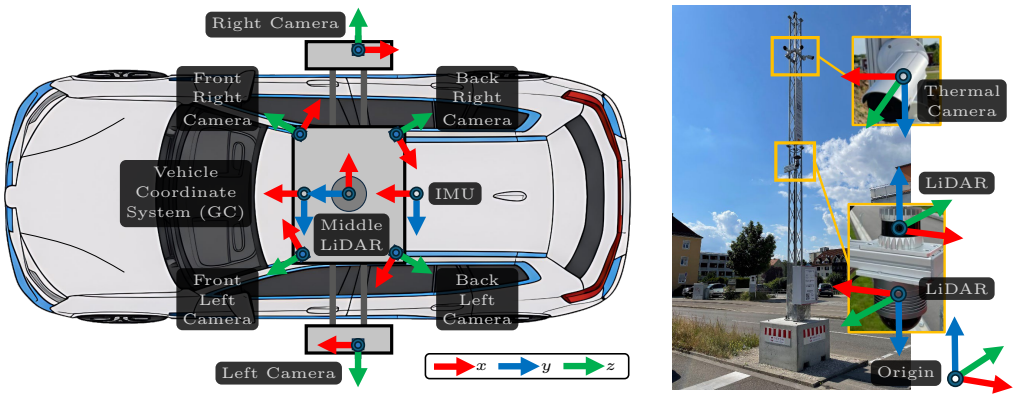


Figure 2: Sensor setup and coordinate frame. The left figure shows details of one vehicle, and the right figure shows details of one pole of a crossing. GC describes the geometric center.

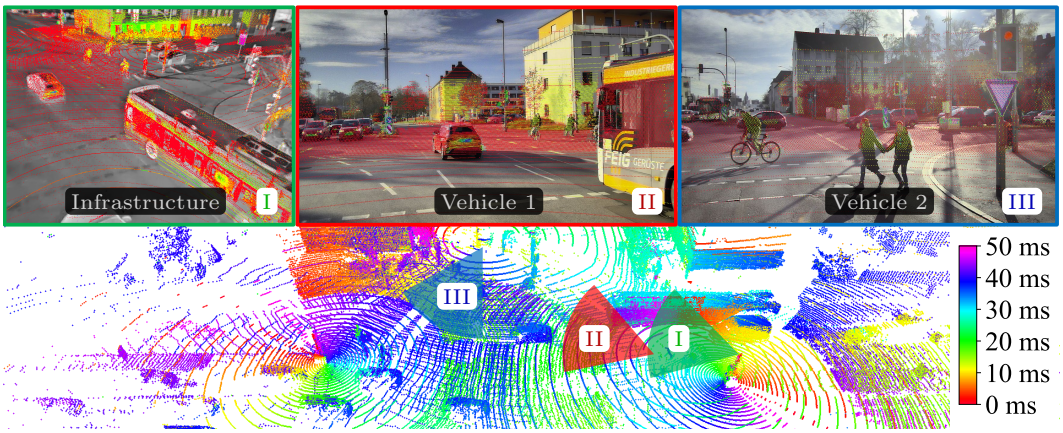


Figure 3: Result of the spatially calibrated and temporally aligned multi-modal sensor sources. The point cloud image highlights the time deviation in a globally fused cooperative LiDAR frame, particularly critical when LiDARs of multiple agents are capturing the same object. The top row shows the overlaid projections of the point cloud into three exemplary sensor perspectives.

Using these annotations, we numerically optimize for transformation matrices by minimizing the reprojection error [19]. This process yields the following extrinsic transformations:

- $\mathbf{T}_{\text{Cam}\rightarrow\text{Global}}$ and $\mathbf{T}_{\text{LiDAR}\rightarrow\text{Global}}$ for infrastructure cameras and LiDARs respectively
- $\mathbf{T}_{\text{Cam}\rightarrow\text{Vehicle}}$ and $\mathbf{T}_{\text{LiDAR}\rightarrow\text{Vehicle}}$ for vehicle-mounted cameras and LiDARs respectively.

3.3 LiDAR Motion Compensation and Data Fusion

Due to the rotational nature of the LiDAR sensor, each point within the point cloud is captured at a slightly different timestamp and vehicle pose. To accurately fuse point clouds from multiple sources, it is necessary to apply transformations on a per-point basis. Each point, captured at time t with LiDAR pose $\mathbf{P}_t$, is transformed into the reference frame $\mathbf{P}_0$ at scan start time $t = 0$ using the relative vehicle motion. Such intra-scan motion compensation is crucial for achieving accurate fusion, especially in dynamic scenes where overlapping observations are acquired by multiple sensors undergoing relative motion.

Using the estimated intrinsic and extrinsic parameters, all vehicle and infrastructure LiDAR point clouds are transformed into a shared global coordinate frame. This unified representation enables the projection of any LiDAR point into any camera image. Figure 3 presents a fused multi-modal visualization of sensor data of a specific frame. The bottom image displays the aggregated point cloud data from all vehicle-mounted and infrastructure-based LiDAR sensors. Each point is color-coded

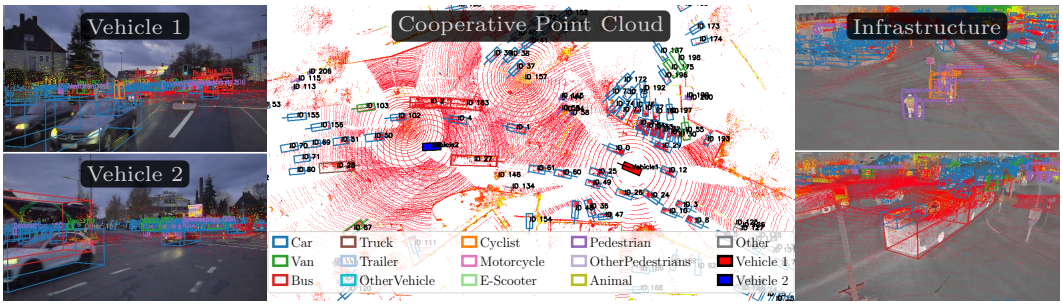


Figure 4: Projection of 3D annotations at one timestamp into three exemplary views: front left camera (left), bird's-eye view fused point cloud (center), and two infrastructure cameras (right) are shown.

based on its temporal offset from the start of the frame, highlighting that objects are observed by different LiDARs at varying timestamps. Assuming a maximum object speed of $50\,\mathrm{km\,h^{-1}}$ and accounting for the worst-case sensor misalignment, the maximum spatial error within a frame is estimated to be $0.7\,\mathrm{m}$. This error is unavoidable, as any object within the scene may move in arbitrary directions. Additionally, Figure 3 shows images from infrastructure and vehicle cameras with overlaid projections of the point cloud.

3.4 Scenario Selection and Annotation

The UrbanIng-V2X dataset is carefully curated from approximately eight hours of recorded data collected across three intersections. Based on the raw recordings, we selected a set of 34 representative 20-second scenarios that capture diverse traffic situations and flow patterns, with a focus on varied vehicle behaviors and object categories. The dataset comprises a wide range of illumination conditions, including 10 daytime, 5 cloudy, 6 moderate-light, 5 late-evening, and 8 nighttime scenarios. All faces and license plates are anonymized using a Gaussian blur to comply with data protection regulations [9]. Annotations are applied to the fused point cloud data from all infrastructure sensors and vehicles, ensuring both spatial and temporal consistency across all modalities. Data quality was rigorously validated through multiple rounds of quality control. Each object is annotated with detailed 3D bounding boxes at a frequency of 10 Hz, specifying their spatial position (x, y, z) and orientation. Additionally, each object is assigned a unique tracking ID per sequence and categorized into one of 13 object classes. These annotations are further enriched with six attribute types, described in the supplementary materials. Figure 4 illustrates one scenario sample from different sensor perspectives. Annotation characteristics are analyzed in detail across trajectory, frame, and object levels. This includes visualizations such as trajectory overlays on HD maps, polar density maps, object category distributions, and statistics on object and track counts (see Figures 5, 6, 7, and 8).

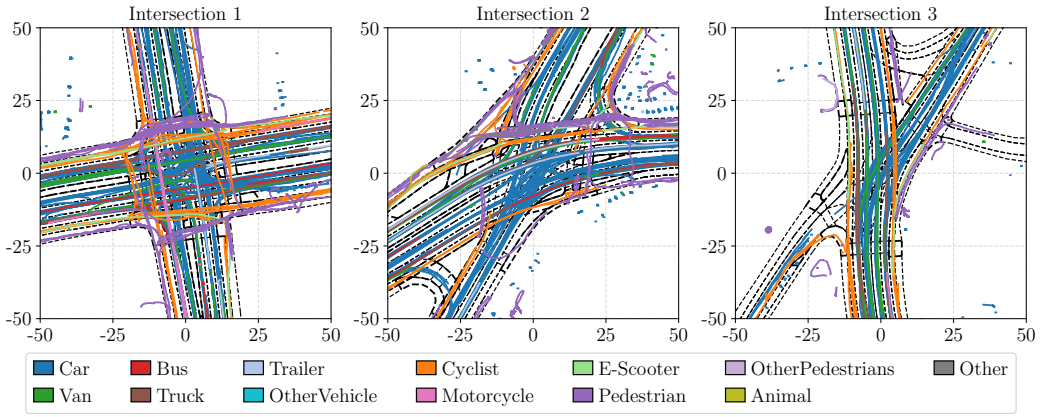


Figure 5: Trajectories projected onto the HD map of each intersection, color-coded by object category, illustrating the quality, density, and variation across the intersection layouts. In total, 2156 trajectories of Intersection 1, 1895 trajectories of Intersection 2, and 835 tracks of Intersection 3 are shown.

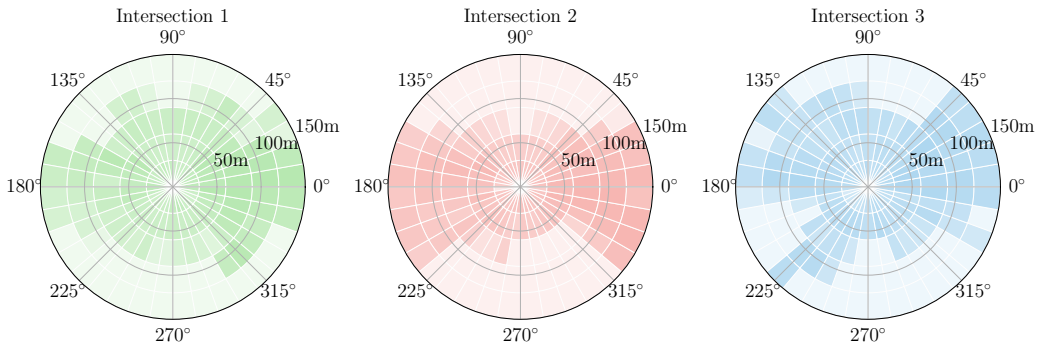


Figure 6: Polar density map showing object distributions by range and angle relative to vehicle agents, separated by intersection. Bin shading indicates object density, with 0° aligned to the vehicle's forward direction. Objects are densely distributed up to 150 meters. While high density along the vehicle axis is expected, the maps also reveal increased angular spread influenced by intersection layouts, which supports benchmark evaluations on generalization.

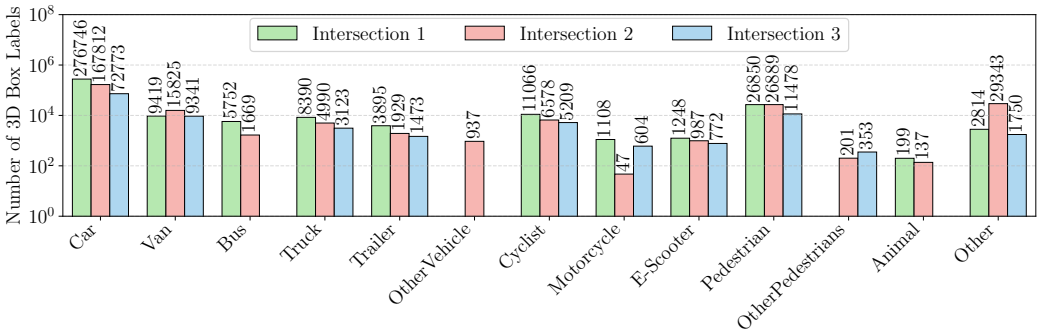
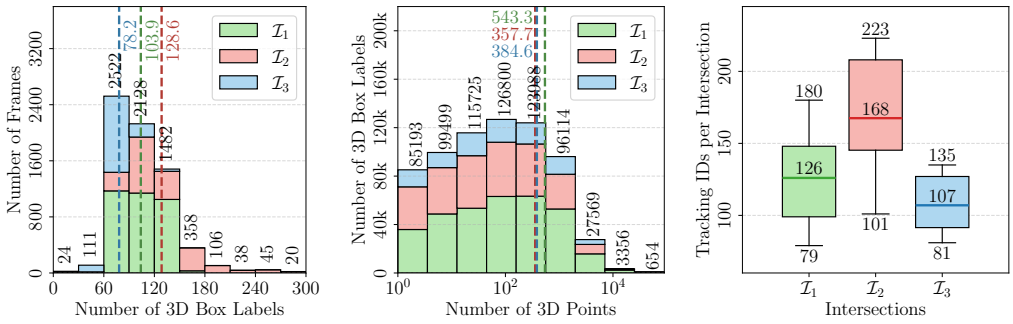


Figure 7: The number of annotated 3D bounding boxes for all 13 object classes, grouped by intersection. While cars constitute the highest amount of 3D annotations, also other categories such as pedestrians, and cyclists are well represented. The distribution is relatively evenly across the intersections. The only exception is OtherVehicle, which is predominantly represented by an excavator located exclusively in Intersection 2 and does not appear in other intersections.



(a) 3D box labels per frame. (b) 3D points per 3D box label. (c) Tracking IDs per intersection.

Figure 8: Intersection 1, 2, and 3 are abbreviated as $\mathcal{I}_1, \mathcal{I}_2, \mathcal{I}_3$, respectively. (a) Frames contain an average of 103.9 objects in $\mathcal{I}_1$, 128.6 in $\mathcal{I}_2$, and 78.2 in $\mathcal{I}_3$. (b) 3D box labels contain an average of 543.3 points in $\mathcal{I}_1$, 357.7 in $\mathcal{I}_2$, and 384.6 in $\mathcal{I}_3$, based on the fused point cloud. (c) Sequences contain a median of 126 tracks in $\mathcal{I}_1$, 168 in $\mathcal{I}_2$, and 107 in $\mathcal{I}_3$. Scene complexity, object density, and observation quality differ significantly across intersection types: $\mathcal{I}_1$ yields the highest point-level visibility per object, $\mathcal{I}_2$ features the most dynamic and densely populated scenarios, and $\mathcal{I}_3$ corresponds to sparser environments with reduced perceptual coverage.

4 Tasks

UrbanIng-V2X provides comprehensive 3D annotations supporting multiple tasks, including object detection—the primary focus of this work—as well as object tracking, trajectory prediction, and localization. The dataset further enables the evaluation of vehicle and infrastructure agents operating in various cooperative modes, allowing for evaluations of the performance of V2V, V2I, and Infrastructure-to-Infrastructure (I2I) (at a sensor pole level) at all three intersections.

3D Object Detection. For a structured analysis, we group the 13 annotated object categories into four superclasses: *Vehicle* (Car, Van), *Two-Wheelers* (Cyclist, Motorcycle, E-Scooter), *Heavy Vehicle* (Truck, Bus, Trailer, Other Vehicle), and *Pedestrian* (Pedestrian, OtherPedestrian). The classes *Animal* and *Other* are excluded due to their low sample counts and high intra-class variability. Bounding boxes beyond ± 100 meters in the x-direction and ± 40 meters in the y-direction are excluded [27]. Furthermore, only objects with at least five LiDAR points in the fused point cloud of the selected agents are considered during both training and evaluation. During training, ego agents are selected randomly to enable viewpoint diversity and improved model generalization. For evaluation, we select one ego agent as an autonomous vehicle and the rest as collaborators, similar to V2X-Real [27]. We report detection performance using the mean Average Precision (mAP) metric, evaluated at low IoU thresholds of 0.3 and 0.5 similar to V2XReal [27] and KITTI-360 [18].

5 Experiments

We present 3D object detection LiDAR-only benchmark results in four strategies: no fusion, early fusion, late fusion, and intermediate fusion. We use F-Cooper [6], AttFuse [30], V2X-ViT [29], Where2Comm [13], and CoBEVT [28] for intermediate fusion. All models are implemented using the PointPillars backbone [15].

5.1 Dataset splits

To reliably assess the generalization capabilities of benchmark algorithms, it is crucial to employ dataset splitting strategies that prevent data leakage and enable fair evaluation. Common approaches include frame-wise [37, 27, 8] and sequence-wise splits [5]. **Frame-wise splitting** distributes individual frames across the training, validation, and test sets by optimizing for equal characteristics of the data across the subsets. However, this approach is prone to data leakage due to strong temporal correlation among frames and could lead to undetected overfitting and misleadingly high performance scores. **Sequence-wise splitting** groups temporally consecutive frames (i. e., driving sequences) into the same set, potentially avoiding data leakage. This method ensures a more realistic evaluation of generalization but may result in less balanced distributions of data characteristics across the splits.

To account for limitations on existing split strategies, we propose two approaches, namely **Equal Intersection Split (EIS)** and **Separate Intersection Split (SIS)**. EIS utilizes a sequence-dependent approach to fairly assess performance within known intersections. To account for the possibility of sequence selection bias while maintaining representativeness, we define three randomized splits with non-overlapping validation and test sets across the splits. Each split consists of 21 training, 6 validation, and 7 test sequences, proportionally distributed across all three intersections. **SIS** leverages the presence of three distinct intersections in UrbanIng-V2X to enable **intersection-wise splitting**. This approach strengthens independence by ensuring that all data from a given intersection appears exclusively in either the training, validation, or test split, thereby promoting generalization to entirely unseen locations. SIS follows a leave-one-out scheme across the three intersections, with four configurations: $\text{SIS}_{1/2\text{vs.}3}$, $\text{SIS}_{2/3\text{vs.}1}$, and $\text{SIS}_{1/3\text{vs.}2}$, where indices denote the intersections used for training and validation versus testing. 4 sequences of intersection 1, 3 of intersection 2, and 3 of intersection 3 are consistently sampled for the validation split, if the intersection is part of the training split.

5.2 Benchmark results

We use the $\text{SIS}_{1/2\text{vs.}3}$ split to benchmark on all above-mentioned SOTA algorithms to evaluate their performance on an unseen intersection. The results are presented in Table 2. Intermediate fusion methods generally outperform no fusion, late fusion, and early fusion, although the latter yields

Table 2: Evaluation of SOTA algorithms using AP metrics on the $\text{SIS}_{1/2\text{vs.}3}$ split.

Method	**AP_{Veh}**		**AP_{HVeh}**		**AP_{Ped}**		**AP_{TWheel}**		**mAP**	
IoU	0.3	0.5	0.3	0.5	0.3	0.5	0.3	0.5	0.3	0.5
No Fusion	49.1	40.9	19.2	17.6	2.0	0.7	18.0	13.8	22.1	18.3
Early Fusion	46.1	41.1	26.8	24.8	6.0	3.5	24.1	21.6	25.8	22.8
Late Fusion	28.7	24.6	9.8	6.9	1.9	0.8	16.7	12.1	14.3	11.1
F-Cooper [6]	52.6	46.7	33.1	24.0	4.6	3.1	25.1	23.2	28.9	24.2
AttFuse [30]	52.7	47.6	34.1	27.8	7.1	4.6	23.7	22.1	29.4	25.5
V2X-ViT [29]	52.0	46.2	32.5	22.2	5.8	3.5	19.7	18.0	27.5	22.5
Where2Comm [13]	50.4	45.8	28.4	25.3	5.1	3.1	23.2	20.9	26.7	23.8
CoBEVT [28]	53.2	46.0	33.8	29.6	5.7	3.3	22.5	20.5	28.8	24.9

Table 3: Evaluation of all combinations of EIS and SIS splits based on CoBEVT [28]. EIS_{avg} represents the averaged score across the three different EIS splits.

Data split	**AP_{Veh}**		**AP_{HVeh}**		**AP_{Ped}**		**AP_{TWheel}**		**mAP**	
IoU	0.3	0.5	0.3	0.5	0.3	0.5	0.3	0.5	0.3	0.5
EIS_{avg}	74.6	68.7	44.7	37.3	21.8	13.1	38.7	33.0	45.0	38.2
$\text{SIS}_{1/2\text{vs.}3}$	53.2	46.0	33.8	29.6	5.7	3.3	22.6	20.5	28.8	24.6
$\text{SIS}_{1/3\text{vs.}2}$	45.1	40.2	14.9	11.3	10.2	6.0	22.0	18.7	23.0	19.1
$\text{SIS}_{2/3\text{vs.}1}$	64.8	59.1	41.5	31.1	10.8	7.4	22.6	18.2	34.9	28.9

competitive performance. Late fusion exhibits the weakest cooperative performance, indicating significant challenges in the association of agent-specific object lists. In contrast, AttFuse [30] achieves the best overall performance, surpassing other methods by at least 0.6 mAP@0.5. A category-wise comparison reveals that Vehicles is the best-performing class. In contrast, Heavy Vehicles, Pedestrians, and Two-Wheelers present greater challenges. We attribute this to the fact that Pedestrians are the smallest objects, while the Two-Wheelers and Heavy Vehicle superclasses exhibit the highest intra-class dimension variance, with a minimum of three original annotation categories.

Further, detailed evaluation on the remaining SIS and EIS splits for the most recently published method CoBEVT [28] are shown in Table 3. The performance on EIS_{avg} is 38.2 mAP@0.5, while the average SIS performance drops to 24.2 mAP@0.5. This indicates generalization issues that need to be solved for future cooperative perception applications, an open challenge that UrbanIng-V2X aims to address.

6 Conclusion

We present UrbanIng-V2X, the first large-scale cooperative perception dataset that integrates multi-vehicle, multi-infrastructure, and multi-sensor modalities across multiple urban intersections. By expanding the diversity of sensor types—including up to 12 RGB cameras, 6 thermal cameras, and 6 LiDAR sensors per scene—UrbanIng-V2X enables research into robust multi-modal, multi-view fusion. The dataset is uniquely designed to evaluate generalization by including both familiar and previously unseen intersection layouts, addressing a critical limitation in existing benchmarks. Our initial baseline experiments with SOTA LiDAR-only cooperative detection models reveal a clear gap in generalization performance: while better results are achieved on known intersections, there is a significant drop of 14.0 mAP@0.5 when models are applied to unseen environments. These results highlight the pressing need for research into models that generalize reliably across varied urban scenes.

To support the community, we release the complete dataset alongside a development kit, HD maps, and a geo-referenced digital twin in CARLA to facilitate research in perception, tracking, prediction, and simulation. Despite its contributions, UrbanIng-V2X has certain limitations. The dataset is restricted to three intersections within Ingolstadt, Germany, and broader generalization will require extending the benchmark to more diverse urban settings and adverse weather conditions. We invite the research

community to use UrbanIng-V2X as a robust foundation for advancing cooperative perception and want to encourage research into generalization, data-efficient learning, and synthetic-to-real transfer techniques.

Acknowledgments

The dataset was collected within the High-Definition Testfield, which was constructed within the KIVI project (45KI05C031) funded by the *Federal Ministry for Digital and Transport of Germany (BMDV)*. This work was supported by the Hightech Agenda Bavaria, the SiRaMiS project (H2-F1116.IN/48/2), and the Bavarian Academic Forum - BayWISS, all funded by the *Bavarian State Ministry of Science and the Arts (StMWK)*.

References

[1] Claudine Badue, Rânik Guidolini, Raphael Vivacqua Carneiro, Pedro Azevedo, Vinicius B. Cardoso, Avelino Forechi, Luan Jesus, Rodrigo Berriel, Thiago M. Paixão, Filipe Mutz, Lucas de Paula Veronese, Thiago Oliveira-Santos, and Alberto F. De Souza. Self-driving cars: A survey. *Expert Systems with Applications*, 165:113816, 2021.

[2] National Transportation Safety Board. Collision between a car operating with automated vehicle control systems and a tractor-semitrailer truck. Technical report, National Transportation Safety Board, 2016. URL `https://www.ntsb.gov/Investigations/Accidentreports/Reports/Har1702.pdf`. Accessed: 2025-05-16.

[3] Michael Botsch, Werner Huber, Lakshman Balasubramanian, Alberto Flores Fernández, Markus Geisler, Christian Gudera, Mauricio Rene Morales Gomez, Peter Riegl, Eduardo Sánchez Morales, Michael Weinzierl, and Karthikeyan Chandra Sekaran. Data collection and safety use cases in smart infrastructures. In *Adjunct Proceedings of the 15th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (Automotive UI)*, pages 333–336, 2023.

[4] Steffen Busch, Christian Koetsier, Jeldrik Axmann, and Claus Brenner. Lumpi: The leibniz university multi-perspective intersection dataset. In *Proceedings of the IEEE Intelligent Vehicles Symposium (IV)*, pages 1127–1134, 2022.

[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 11618–11628, 2020.

[6] 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 4th ACM/IEEE Symposium on Edge Computing (SEC)*, pages 88–100, 2019.

[7] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. CARLA: An open urban driving simulator. In *Proceedings of the Conference on Robot Learning (CoRL)*, volume 78, pages 1–16, 2017.

[8] Felix Fent, Fabian Kuttenreich, Florian Ruch, Farija Rizwin, Stefan Juergens, Lorenz Lechermann, Christian Nissler, Andrea Perl, Ulrich Voll, Min Yan, and Markus Lienkamp. Man truckscenes: A multimodal dataset for autonomous trucking in diverse conditions. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pages 62062–62082, 2024.

[9] GDPR. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). `https://eur-lex.europa.eu/eli/reg/2016/679/oj`, April 2016. Accessed: 2025-05-16.

[10] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *International Journal of Robotics Research (IJRR)*, 32(11):1231–1237, 2013.

[11] Jeremias Gerner, Dominik Rößle, Daniel Cremers, Klaus Bogenberger, Torsten Schön, and Stefanie Schmidtner. Enhancing realistic floating car observers in microscopic traffic simulation. In *Proceedings of the IEEE International Conference on Intelligent Transportation Systems (ITSC)*, pages 2396–2403, 2023.

[12] 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.

[13] Yue Hu, Shaoheng Fang, Zixing Lei, Yiqi Zhong, and Siheng Chen. Where2comm: Communication-efficient collaborative perception via spatial confidence maps. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pages 4874–4886, 2022.

[14] IEEE 1588. IEEE Standard for a Precision Clock Synchronization Protocol for Networked Measurement and Control Systems. Standard, IEEE The Institute of Electrical and Electronics Engineers, New York City, US, November 2019.

[15] 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 12689–12697, 2019.

[16] Robert J. Lempert, Benjamin Preston, Sophia M. Charan, Laura Fraade-Blanar, and Marjory S. Blumenthal. The societal benefits of vehicle connectivity. *Transportation Research Part D: Transport and Environment*, 93:102750, 2021.

[17] 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 (RA-L)*, 7(4):10914–10921, 2022.

[18] Yiyi Liao, Jun Xie, and Andreas Geiger. KITTI-360: A Novel Dataset and Benchmarks for Urban Scene Understanding in 2D and 3D . *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 45(3):3292–3310, 2023.

[19] Jorge Nocedal and Stephen J. Wright. *Numerical Optimization*. Springer, 2nd edition, 2006. ISBN 9780387303031.

[20] Fabian Poggenhans, Jan-Hendrik Pauls, Johannes Janosovits, Stefan Orf, Maximilian Naumann, Florian Kuhnt, and Matthias Mayr. Lanelet2: A high-definition map framework for the future of automated driving. In *Proceedings of the IEEE International Conference on Intelligent Transportation Systems (ITSC)*, pages 1672–1679, 2018.

[21] Dominik Rößle, Jeremias Gerner, Klaus Bogenberger, Daniel Cremers, Stefanie Schmidtner, and Torsten Schön. Unlocking past information: Temporal embeddings in cooperative bird's eye view prediction. In *Proceedings of the IEEE Intelligent Vehicles Symposium (IV)*, pages 2220–2225, 2024.

[22] Karthikeyan Chandra Sekaran, Lakshman Balasubramanian, Michael Botsch, and Wolfgang Utschick. Open-set object detection for the identification and localization of dissimilar novel classes by means of infrastructure sensors. In *Proceedings of the IEEE Intelligent Vehicles Symposium (IV)*, pages 1643–1650, 2024.

[23] Anshuman Sharma and Zuduo Zheng. *Connected and Automated Vehicles: Opportunities and Challenges for Transportation Systems, Smart Cities, and Societies*, pages 273–296. Springer Singapore, 2021.

[24] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, Vijay Vasudevan, Wei Han, Jiquan Ngiam, Hang Zhao, Aleksei Timofeev, Scott Ettinger, Maxim Krivokon, Amy Gao, Aditya Joshi, Yu Zhang, Jonathon Shlens, Zhifeng Chen, and Dragomir Anguelov. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2443–2451, 2020.

[25] OpenPCDet Development Team. Openpcdet: An open-source toolbox for 3d object detection from point clouds. `https://github.com/open-mmlab/OpenPCDet`, 2020.

[26] Uwe Winkelhake. *Vision of a Digitized Automotive Industry 2030*, pages 95–167. Springer Berlin Heidelberg, 2025.

[27] Hao Xiang, Zhaoliang Zheng, Xin Xia, Runsheng Xu, Letian Gao, Zewei Zhou, Xu Han, Xinkai Ji, Mingxi Li, Zonglin Meng, Li Jin, Mingyue Lei, Zhaoyang Ma, Zihang He, Haoxuan Ma, Yunshuang Yuan, Yingqian Zhao, and Jiaqi Ma. V2x-real: A largs-scale dataset for vehicle-to-everything cooperative perception. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 455–470, 2024.

[28] 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 *Proceedings of the Conference on Robot Learning (CoRL)*, pages 989–1000, 2022.

[29] 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 *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 107–124, 2022.

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

[31] Runsheng Xu, Xin Xia, Jinlong Li, Hanzhao Li, Shuo Zhang, Zhengzhong Tu, Zonglin Meng, Hao Xiang, Xiaoyu Dong, Rui Song, Hongkai Yu, Bolei Zhou, and Jiaqi Ma. 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.

[32] Kun Yang, Dingkang Yang, Jingyu Zhang, Mingcheng Li, Yang Liu, Jing Liu, Hanqi Wang, Peng Sun, and Liang Song. Spatio-temporal domain awareness for multi-agent collaborative perception. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 23326–23335, 2023.

[33] 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 21329–21338, 2022.

[34] Haibao Yu, Wenxian Yang, Hongzhi Ruan, Zhenwei Yang, Yingjuan Tang, Xu Gao, Xin Hao, Yifeng Shi, Yifeng Pan, Ning Sun, Juan Song, Jirui Yuan, Ping Luo, and Zaiqing Nie. 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.

[35] Zewei Zhou, Ziru Yang, Yuanjian Zhang, Yanjun Huang, Hong Chen, and Zhuoping Yu. A comprehensive study of speed prediction in transportation system: from vehicle to traffic. *iScience*, 25(3):103909, 2022.

[36] 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 *Proceedings of the IEEE International Conference on Intelligent Transportation Systems (ITSC)*, pages 1030–1037, 2023.

[37] Walter Zimmer, Gerhard Arya Wardana, Suren Sritharan, Xingcheng Zhou, Rui Song, and Alois C. Knoll. Tumtraf v2x cooperative perception dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22668–22677, 2024.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We discuss the contributions and novelty of the paper throughout the whole paper. A first overview is provided in Section 1, put into the context of related work in Section 2, analyzed in Section 3, and empirically evaluated in Section 5.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitation to the existing work is reported in the conclusion and elaborated in the supplementary material.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: The paper itself is not proposing a novel theory and does not require a theoretical proof.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: All the necessary code and data are provided and will be made publicly available after acceptance. Further, the paper lists the data split details in Section5 and all hyperparameters in the supplementary material. The necessary code to generate the exact splits used is provided in the submitted GitHub link.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We share our data on dataverse and we will make it publicly available. In addition, a GitHub link to the development kit is shared for the review process and provides sufficient instructions to reproduce the main experimental results.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The main experiments (Section 5) are build on the public framework OpenCOOD and utilize their standardized training procedures. A detailed description of hyperparameters is additionally provided in the supplementary material.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Test and validation splits were implemented according to Machine Learning standards and are shown in section 5. The randomized selection of test sequences is performed three times to provide statistical significance. All results are published.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The utilized computing resources is attached in the supplementary materials. The paper does not introduce a new algorithm, but instead refers to state-of-the-art models with known computing requirements.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Faces and license plates of all RGB images were blurred to protect the privacy of people captured during the dataset recordings 3.4. A license file defines the usability of code and dataset.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The main paper is currently focusing on the positive aspects of cooperative perception 1 and its safety chances. Supplementary materials describe a more in-depth discussion.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [Yes]

Justification: Data is anonymized to prevent the misuse of personal data (Section 3.4). We utilize the Dataverse platform with user questions and license files that limit the permitted usage of the dataset and code.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All utilized code frameworks were referenced according to citation standards. More specifically, the code base is built upon OpenCOOD and the original creators are referenced. (see section 5).

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Our publication has two main assets: a dataset and the code. The dataset is described in detail in the paper in section 3 and in the supplemental materials, and its usability is guaranteed through a development kit. The shared GitHub code provides an environment setup and a Readme to support a good user experience.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: No crowdsourcing has been performed.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: No study participants were involved in the generation of this paper.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: LLMs do not impact any part of the methodology, scientific rigorousness, or originality of the research.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

A Sensor setup

The UrbanIng-V2X dataset was collected using two vehicles and seven infrastructure-mounted sensor poles. The setup spans three intersections: Intersection 1 is equipped with three sensor poles, while intersections 2 and 3 each have two sensor poles. Intersection 1 includes 6 thermal cameras and 4 LiDAR sensors, Intersection 2 includes 5 thermal cameras and 4 LiDAR sensors, and Intersection 3 includes 6 thermal cameras and 4 LiDAR sensors. Table 4 and 5 provide an overview of the infrastructure and vehicle sensor specifications. Figure 9 illustrates the FOV coverage of the thermal cameras and LiDAR sensors installed at each intersection. Additionally, the FOV coverage of our vehicles for the RGB cameras and the LiDAR sensor is depicted in Figure 10.

Table 4: Vehicle sensor specifications (per vehicle)

Sensor	**Details**
RGB Cameras (6×)	Sensing GSML2 SG2-AR0233C-5200-G2A, 20 FPS, 1920 x 1080 resolution, $60°$ horizontal FOV (4x); $100°$ horizontal FOV (2x)
LiDAR (1×)	Robosense Ruby Plus, 20 FPS, 128 rays, $360\,\mathrm{degree}$ horizontal FOV, $-25°$ to $15°$ vertical FOV, $\leq 240\,\mathrm{m}$ range at $\geq 10\,\%$ reflectivity
GPS/IMU (1×)	Genesys ADMA Pro+, 100 FPS, RTK correction, $1\,\mathrm{cm}$ precise position data

Table 5: Infrastructure sensor specifications (per intersection)

Sensor	**Details**
Thermal Cameras (5–6×)	Axis Q1942-E, 30 FPS, 640 x 480 resolution, $63°$ horizontal FOV
LiDARs (2×)	Ouster OS1-64 (Below Horizon) Rev 06, 20 FPS, 64 rays, $\leq 45\,\mathrm{m}$ range at $\geq 10\,\%$ reflectivity, $360\,\mathrm{degree}$ horizontal FOV, $-22.5°$ to $0°$ vertical FOV
LiDARs (2×)	Robosense Bpearl, 20 FPS, 32 rays, $\leq 30\,\mathrm{m}$ range at $\geq 10\,\%$ reflectivity blind spot sensor, $360\,\mathrm{degree}$ horizontal FOV, $-90°$ to $0°$ vertical FOV

B Data annotation

B.1 Annotation process

The annotations underwent a multi-stage quality assurance process. After the initial annotation phase, in total, three review cycles with a manual refinement of bounding boxes by a professional annotation company were performed. At each stage, independent reviewers reported errors to enhance the precision of bounding boxes, object trajectories, and orientation estimates across the dataset.

B.2 Object classes and object attributes

In addition to class labels, we assigned semantic attributes to all annotated objects to capture more detailed characteristics and behavioral states. For the purpose of benchmark evaluation, we grouped specific and closely related object classes into superclasses to perform a more structured detection task. Figure 11 illustrates the structure of these superclasses, their associated object types, and the attribute types applicable to each object type. The object types Animal and Other are not included in the figure, as they were underrepresented in the dataset and thus not grouped into any superclasses. However, both object types are also annotated with the occlusion attribute. Figure 12 provides an overview of all attribute types and their respective subcategories, along with the frequency of their occurrences in the dataset.

B.3 Data anonymization

To ensure privacy compliance, we anonymized the RGB camera data of our dataset. The dataset comprises a total of 163.200 full HD RGB images, recorded at 20 FPS across all sequences. All visible faces and license plates were annotated with 2D bounding boxes and anonymized using a

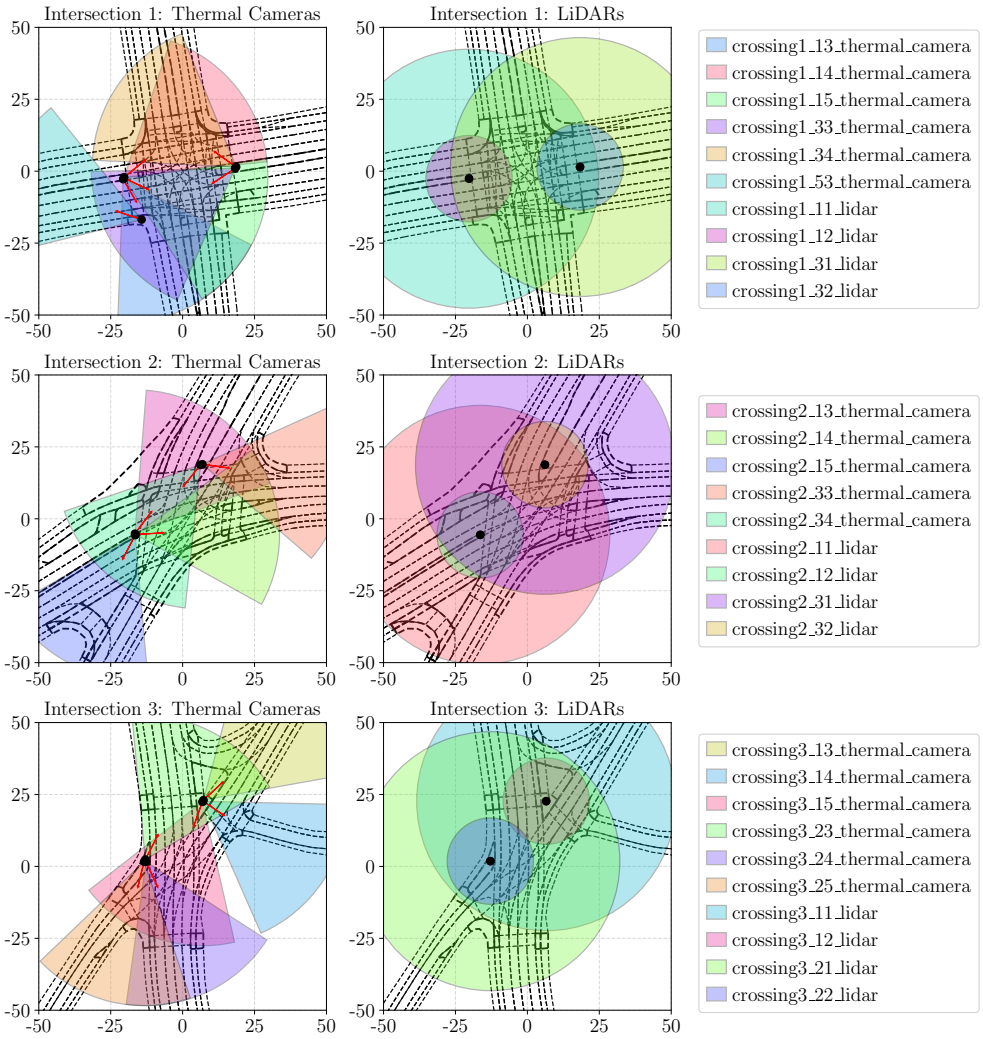


Figure 9: Sensor coverage for each intersection, with legend entries corresponding to folder names in the dataset.

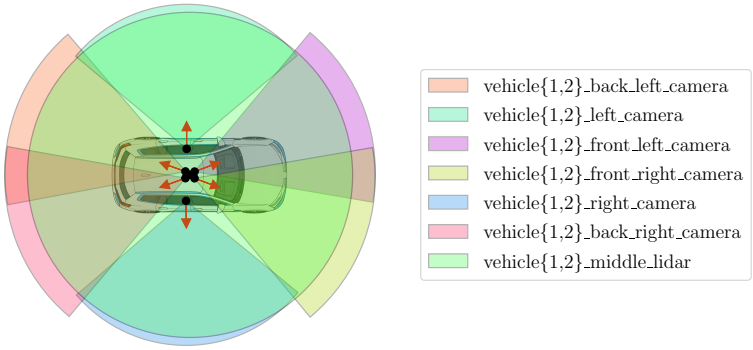


Figure 10: Sensor coverage for each vehicle, with legend entries corresponding to folder names in the dataset.

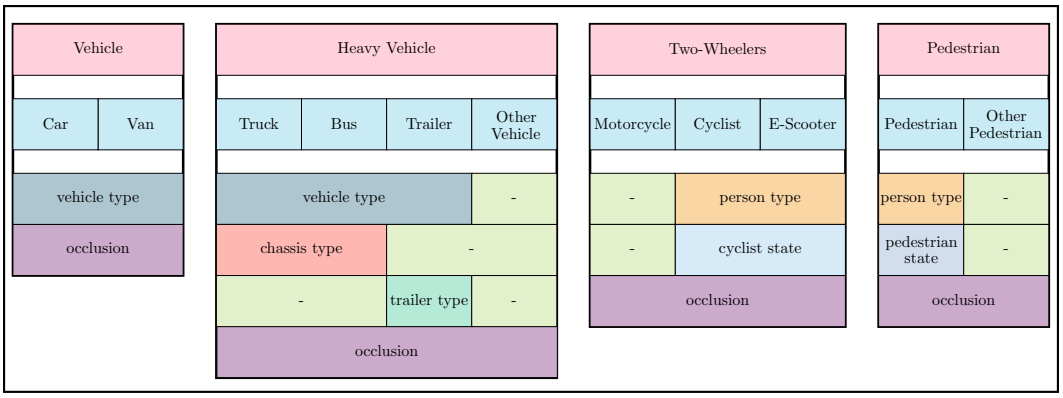


Figure 11: The first row displays the superclasses used for training the multi-object detectors. The second row shows the individual object classes grouped under their respective superclasses. The third row illustrates the attributes associated with each object class, represented by different color codes. The object classes Animal and Other are excluded from this overview, as they are not assigned to any superclass and are only annotated with the occlusion attribute.

Gaussian blur with a 75×75 kernel. The annotations are publicly available in our Git repository to support advanced techniques such as inpainting or synthetic replacement.

B.4 Extended annotation statistics

To provide further insights into the dataset, we present additional statistics for all annotated object classes. Figure 13 shows the distributions of the object dimensions for all provided classes of our dataset. These distributions also reveal intra-superclass variations, for instance, highlighting the significant variance observed among classes within the Heavy Vehicle category. Figure 14 presents the average number of LiDAR points captured per 3D bounding box across varying distances. Each subplot represents a specific object class and compares the point density of vehicle-mounted and infrastructure-mounted LiDAR sensors per agent. While the superclasses Vehicle and Heavy Vehicle have densities up to 100 points per object at the benchmark range of $100\,\mathrm{m}$, the superclasses Pedestrian and Two-Wheelers show significant drops at ranges of approximately $70\,\mathrm{m}$.

C Experiments

C.1 Computer resources

The experiments and computations described in this work were performed on a workstation equipped with an NVIDIA RTX A6000 GPU and an AMD Ryzen Threadripper PRO 5955WX processor with 16 cores, running Ubuntu 22.04.5 LTS.

C.2 Implementation details

Multi-class detection. Our dataset includes 13 distinct object classes: Car, Van, Bus, Truck, Trailer, Vehicle, Cyclist, Motorcycle, E-Scooter, Pedestrian, OtherPedestrians, Animal, Other. These classes were selected to reflect the diversity of road users in urban environments and to enable comprehensive multi-class detection. We follow the standardized training procedure of the OpenCOOD [30] framework and follow the approach of V2XReal [27], leveraging the OpenPCDet framework [25] to enable multi-class evaluation. All models are trained for 60 epochs with a batch size of 4. We use the Adam optimizer with a learning rate of 1×10^{-3} and a weight decay of $\sigma = 10^{-4}$. A cosine annealing learning rate schedule is applied, starting with a warm-up phase over the first 10 epochs. During this phase, the learning rate increases from 2×10^{-4} to its peak and gradually decays to 2×10^{-5} by the final epoch. Model configurations, hyperparameters, and training setups for all approaches are provided. For evaluation, we selected the model checkpoint corresponding to the best validation performance, evaluated every 10 epochs.

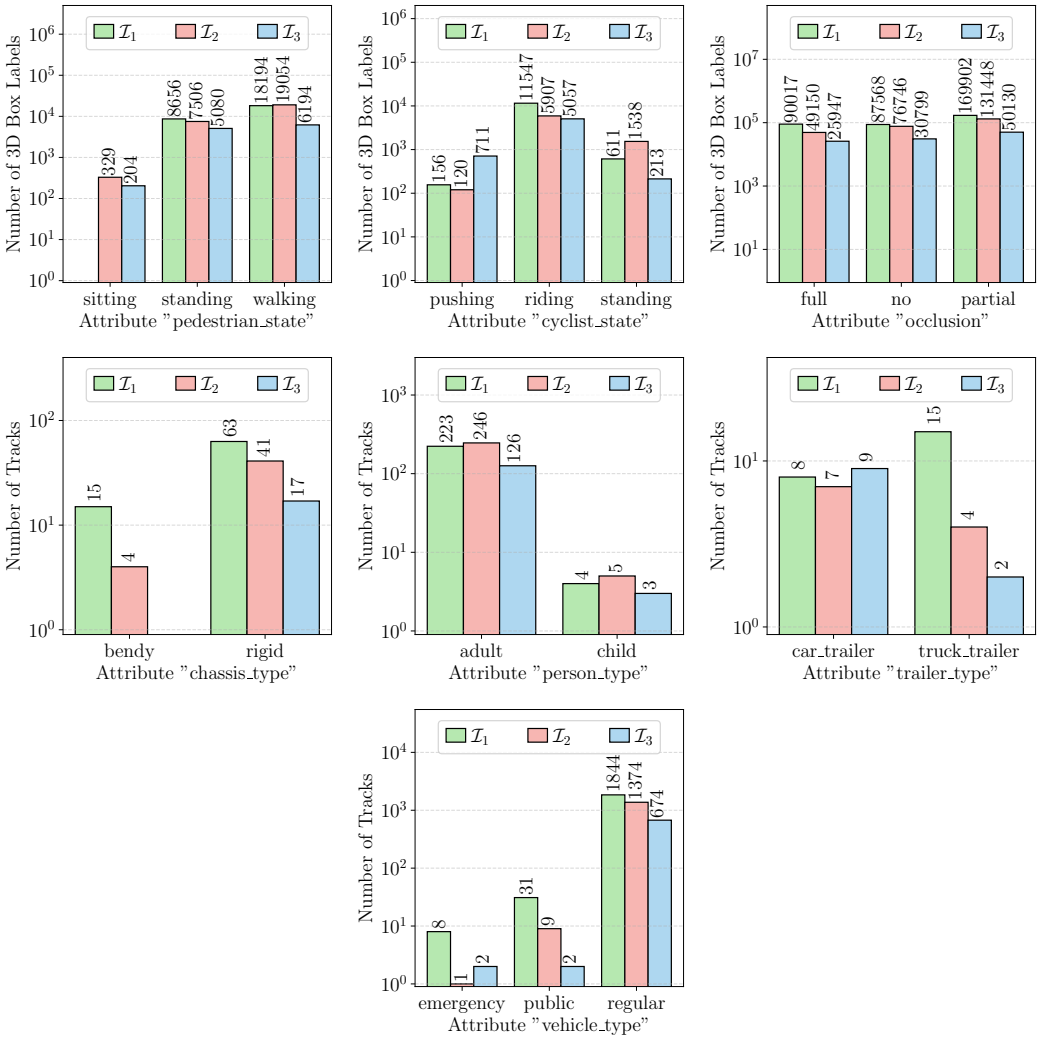


Figure 12: Visualization of all attributes and the frequency of their subcategories. We abbreviate Intersection 1, Intersection 2, and Intersection 3 as $\mathcal{I}_1$, $\mathcal{I}_2$, and $\mathcal{I}_3$, respectively. The attributes pedestrian_state, cyclist_state, and occlusion are frame-based. All other attribute types are track-based.

Model training time. Since our dataset enables focusing on different intersections across training and testing, we use the $\text{SIS}_{1/2\text{vs.}3}$ dataset split and the overall best-performing CoBEVT [28] model for training time estimates.

The $\text{SIS}_{1/2\text{vs.}3}$ split includes:

- Training set: 13 sequences from Intersection 1 and 7 sequences from Intersection 2 (total: 20 sequences).
- Validation set: 4 sequences from Intersection 1 and 3 sequences from Intersection 2 (total: 7 sequences).
- Test set: All 7 sequences from Intersection 3.

Training is parallelized using 16 PyTorch data workers (equal to the number of available CPU cores) for efficient data loading and augmentation. As a representative example, training the CoBEVT model on the $\text{SIS}_{1/2\text{vs.}3}$ split requires approximately 18 hours to complete 60 epochs.

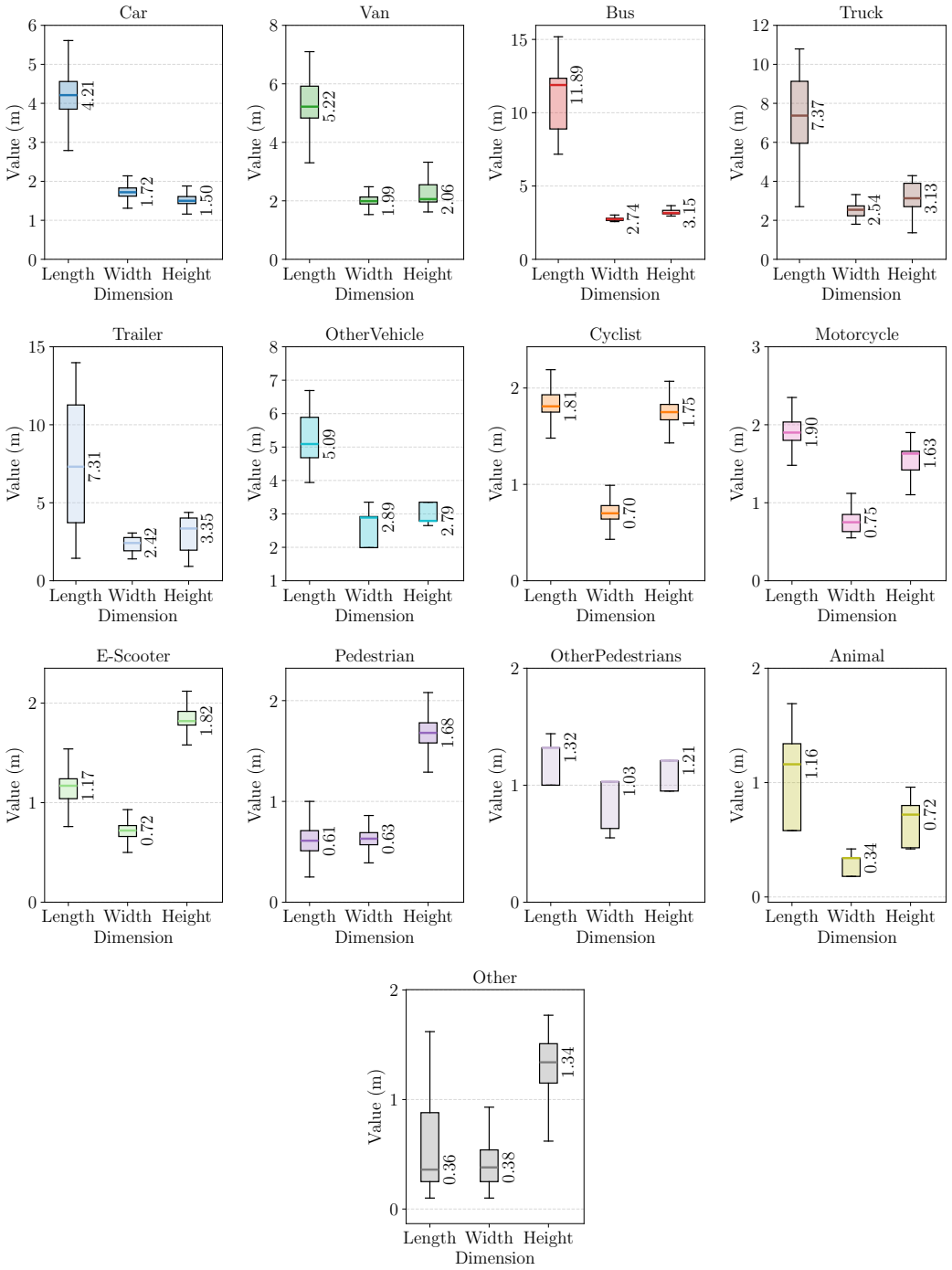


Figure 13: Distribution of object dimensions (length, width, height) for each dataset class. Each box plot summarizes the statistical spread of object dimensions per class, highlighting inter-class variation.

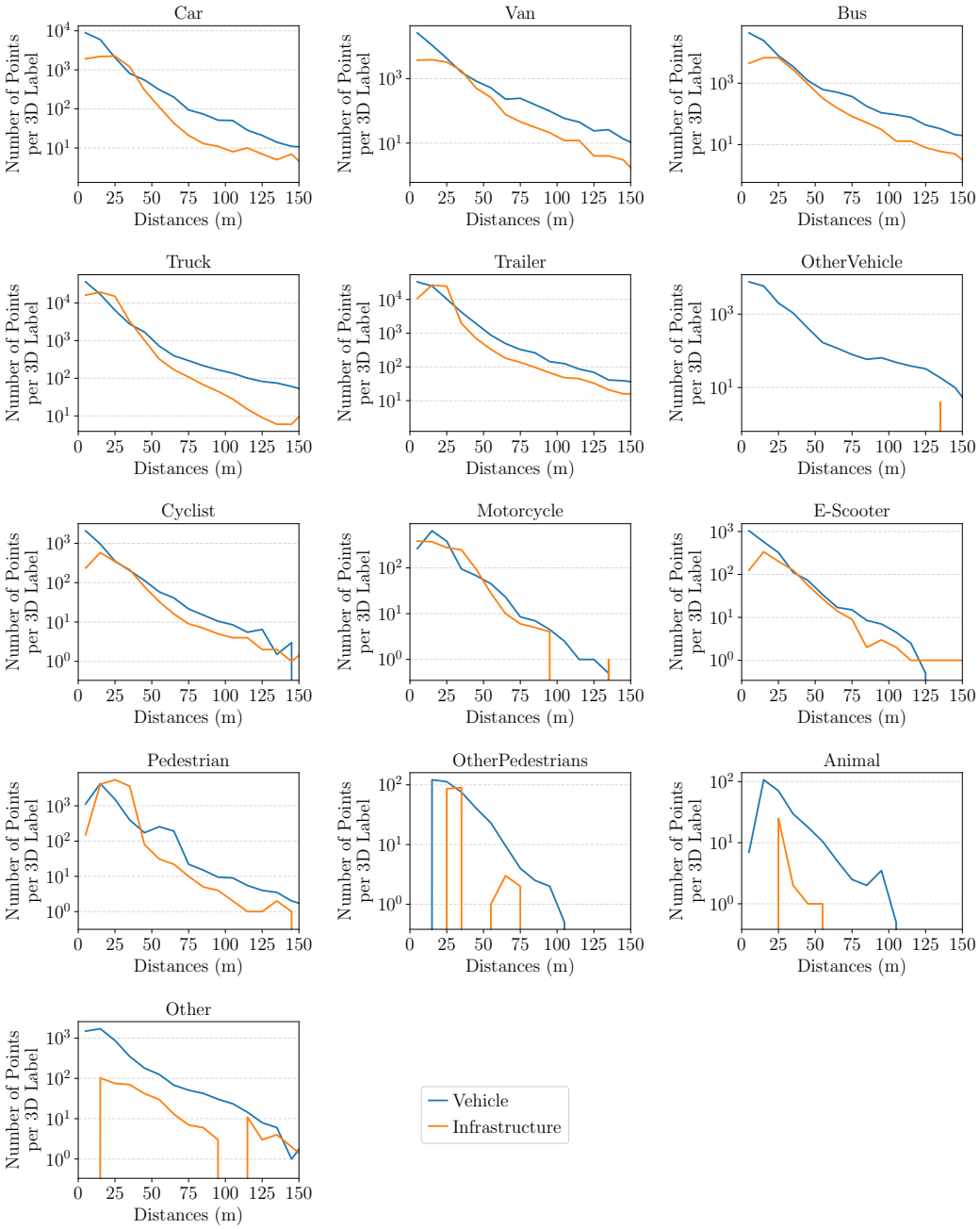


Figure 14: Comparison of the number of LiDAR points per 3D annotation across various distances. The infrastructure data is based on the fused point cloud in the local intersection origins. The vehicle values represent the average across both vehicle agents with respect to their vehicle coordinate frames. Specifically for the infrastructure plots, some classes exhibit irregular point density trends. The reasons are the static nature of the sensors and the sparse distribution of annotated instances across certain distance intervals.

C.3 Further benchmark results

We further present the results of multiple benchmark methods for a single EIS split. All intermediate fusion models perform better within known intersections than on the $\text{SIS}_{1/2\text{vs}.3}$ split (Table 2). While AttFuse [30] is the best-performing method when generalizing to an unknown intersection, V2X-ViT [29] attains the highest performance for the EIS split evaluation in Table 6.

Table 6: Evaluation of SOTA algorithms using AP metrics on a single random EIS split.

Method	$\textbf{AP}_{\textit{Veh}}$		$\textbf{AP}_{\textit{HVeh}}$		$\textbf{AP}_{\textit{Ped}}$		$\textbf{AP}_{\textit{TWheel}}$		**mAP**	
IoU	0.3	0.5	0.3	0.5	0.3	0.5	0.3	0.5	0.3	0.5
F-Cooper [6]	70.4	62.5	42.6	32.2	11.5	5.1	23.1	20.1	36.9	30.0
AttFuse [30]	65.8	56.9	48.0	36.8	11.1	5.9	19.1	15.7	36.0	28.8
V2X-ViT [29]	73.2	65.0	48.7	38.6	17.2	8.7	28.4	23.0	41.9	33.8
Where2Comm [13]	67.9	59.5	40.0	31.4	12.3	6.3	17.8	14.8	34.5	28.0
CoBEVT [28]	71.8	63.4	46.6	35.6	18.5	9.4	29.5	22.5	41.7	33.2

D Limitations

The presented dataset comprises three intersections in Ingolstadt, Germany, offering a more diverse setting than existing real-world cooperative perception benchmarks with multiple vehicles and multiple infrastructure poles. Ingolstadt is one of the few cities in Germany with a permanent, multi-intersection V2X infrastructure deployment at this scale, making it uniquely suited for collecting a dataset of this complexity. The selected locations were deliberately chosen along major arterial roads that reflect common traffic dynamics, infrastructure layouts, and occlusion patterns typically observed in many European metropolitan areas. While this contributes a step forward in promoting generalization challenges in SOTA cooperative perception algorithms, further extensions to intersections across a wider range of cities and urban topologies could support broader applicability and robustness. Future research could also focus on capturing data under adverse weather conditions such as rain, fog, or snow to improve environmental diversity. With a total sequence length of approximately 11 minutes for each agent, UrbanIng-V2X achieves a per-agent duration comparable to other V2X datasets listed in Table 1, though it remains smaller than SOTA single-agent autonomous driving datasets. Even though the data collection process and objectives of cooperative perception datasets differ fundamentally from single-agent recordings, a long-term goal for the field is to scale their raw recording durations toward the levels of SOTA single-agent autonomous driving datasets. Annotations were performed using LiDAR data to ensure high-precision depth estimation. Objects visible exclusively in camera sensors, for example, at great distances without corresponding LiDAR points, may not be annotated. Despite extensive quality assurance, including rounds of manual annotation review, the procedure itself inherently carries a risk of human error.

E Societal impact

Cooperative perception offers significant potential to improve situational awareness and safety in autonomous systems, particularly within complex urban environments. By enabling vehicles to share sensor data and jointly interpret surroundings, it addresses key limitations of isolated single-agent autonomy. However, this shift toward multi-agent cooperation introduces new challenges, including the need for a reliable and secure communication infrastructure [23]. Beyond technical concerns, these systems may pose broader risks to personal privacy and the autonomy of individual drivers, as increased connectivity could enable persistent monitoring, centralized control, or unintended surveillance. Moreover, while cooperative perception could yield substantial benefits in transportation safety, efficiency, and comfort, its societal value depends on the equitable deployment of autonomous technologies. Without deliberate policy and investment, these technologies risk deepening existing disparities by primarily benefiting higher-income populations [16]. As autonomous vehicles are projected to account for up to 30 percent of urban traffic by 2030, technology [26], with connectivity identified as a key enabler, it is critical that their development is guided by supportive regulatory frameworks that safeguard the broader public interest.

F Dataset visualization

This section provides a detailed overview of our dataset. The trajectories of all intersection sequences are visualized in Section F.1, F.2, F.3. In addition, for each intersection we show one representative frame from all sensor perspectives in Section F.4, F.5, and F.6.

F.1 Intersection 1 trajectory visualization

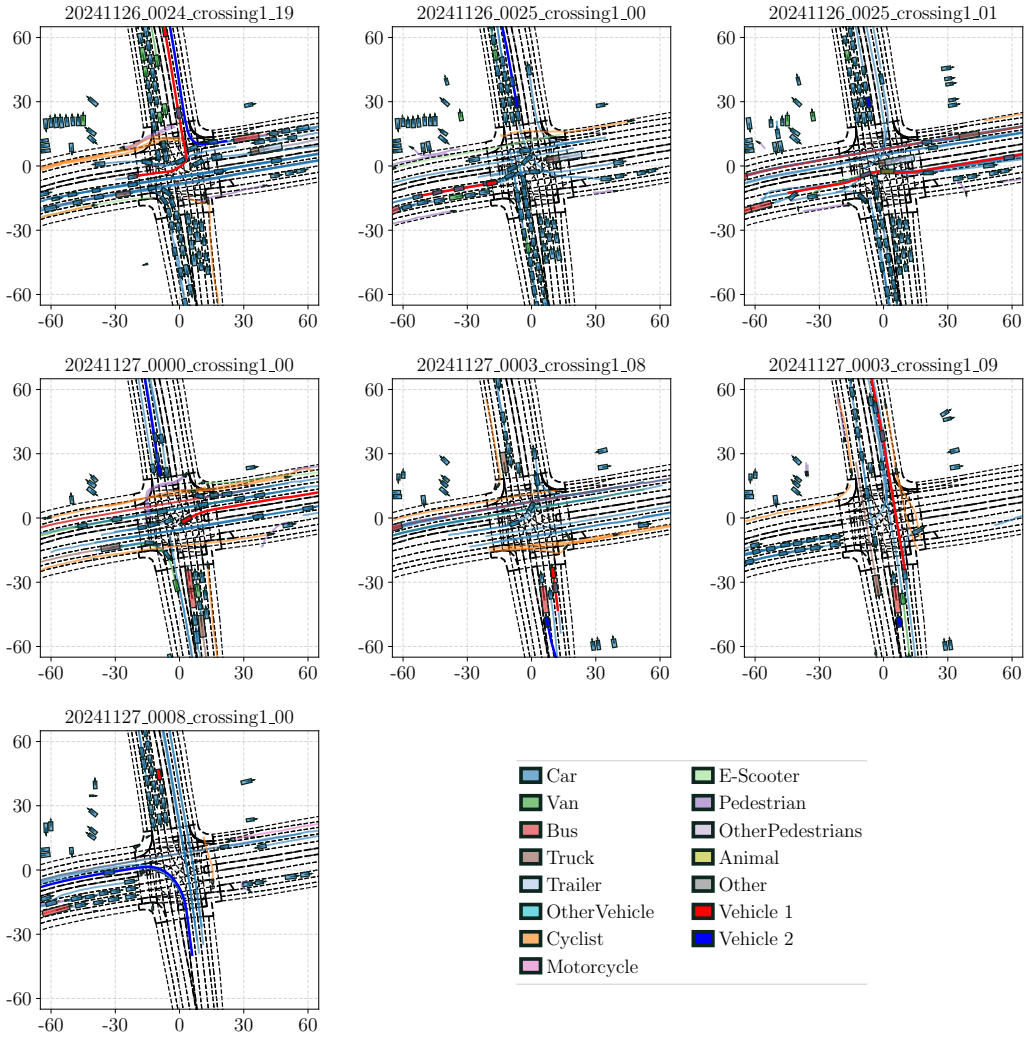


Figure 15: Visualization of trajectories at Intersection 1 across sequences 1-7. Each subplot shows the trajectories of all annotated object classes for a sequence. The sequence names correspond to the original folder names in the dataset.

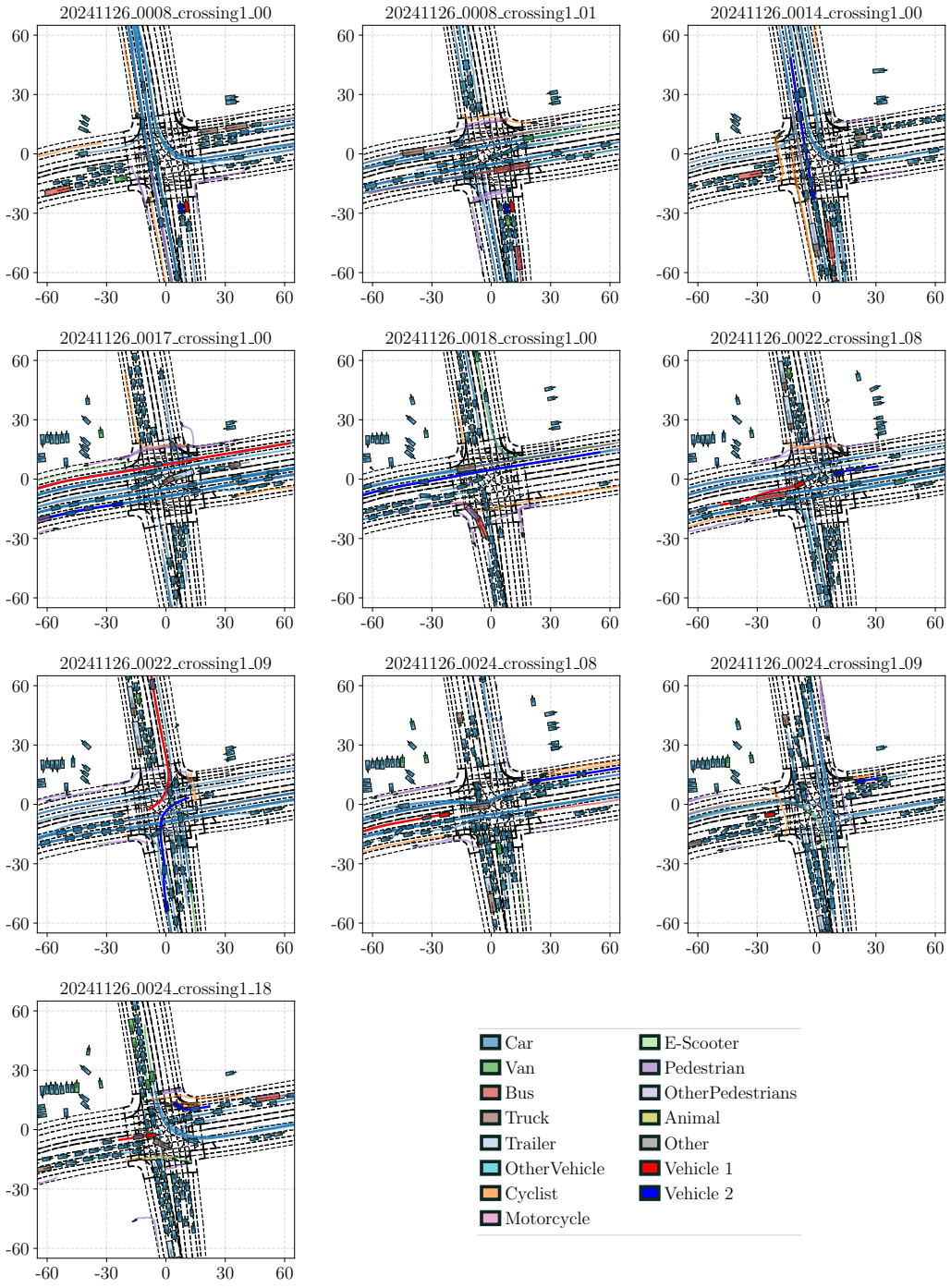


Figure 16: Visualization of trajectories at Intersection 1 across sequences 8-17. Each subplot shows the trajectories of all annotated object classes for a sequence. The sequence names correspond to the original filenames in the dataset.

F.2 Intersection 2 trajectory visualization

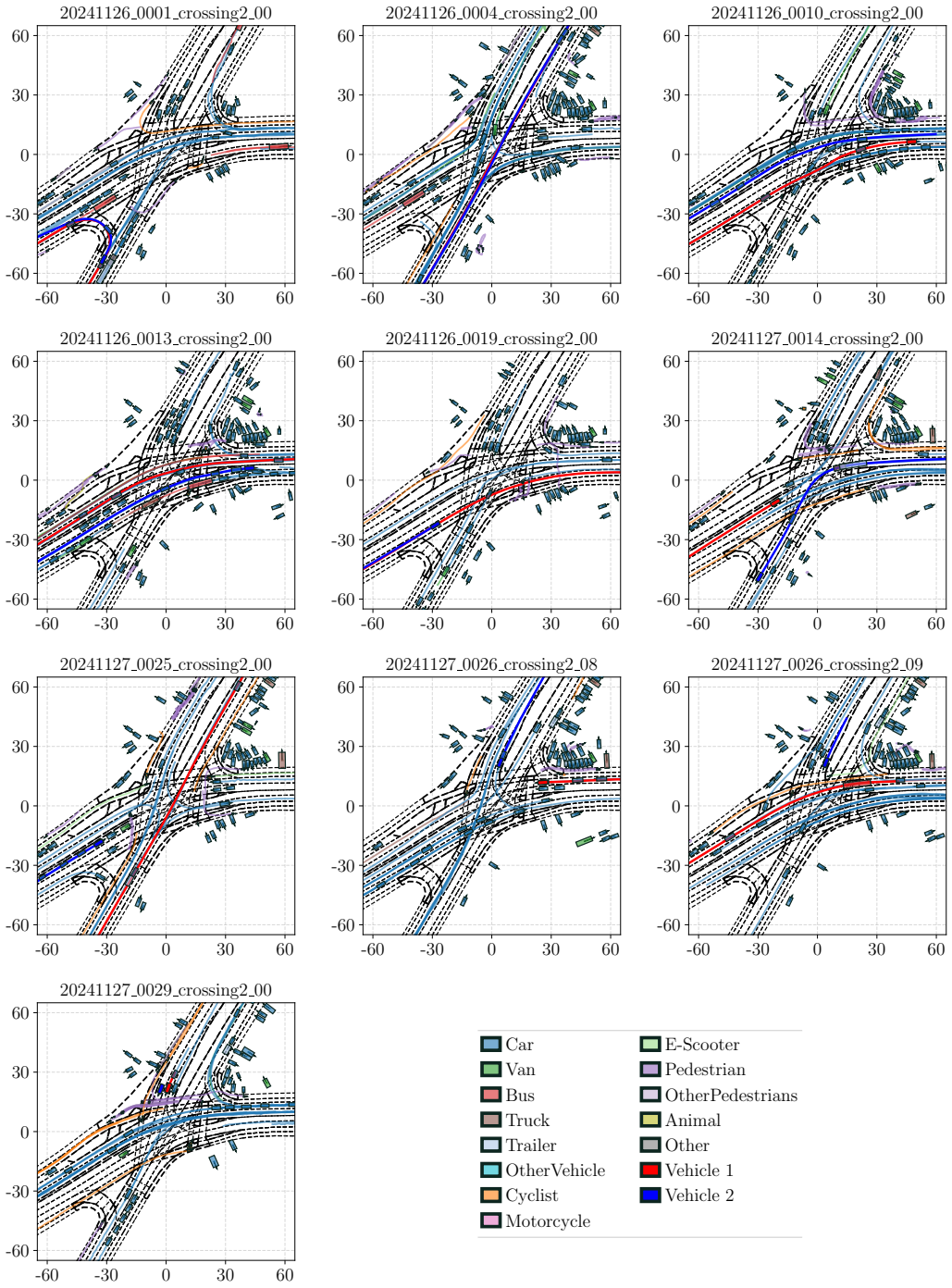


Figure 17: Visualization of trajectories at Intersection 2 across all sequences. Each subplot shows the trajectories of all annotated object classes for a sequence. The sequence names correspond to the original filenames in the dataset.

F.3 Intersection 3 trajectory visualization

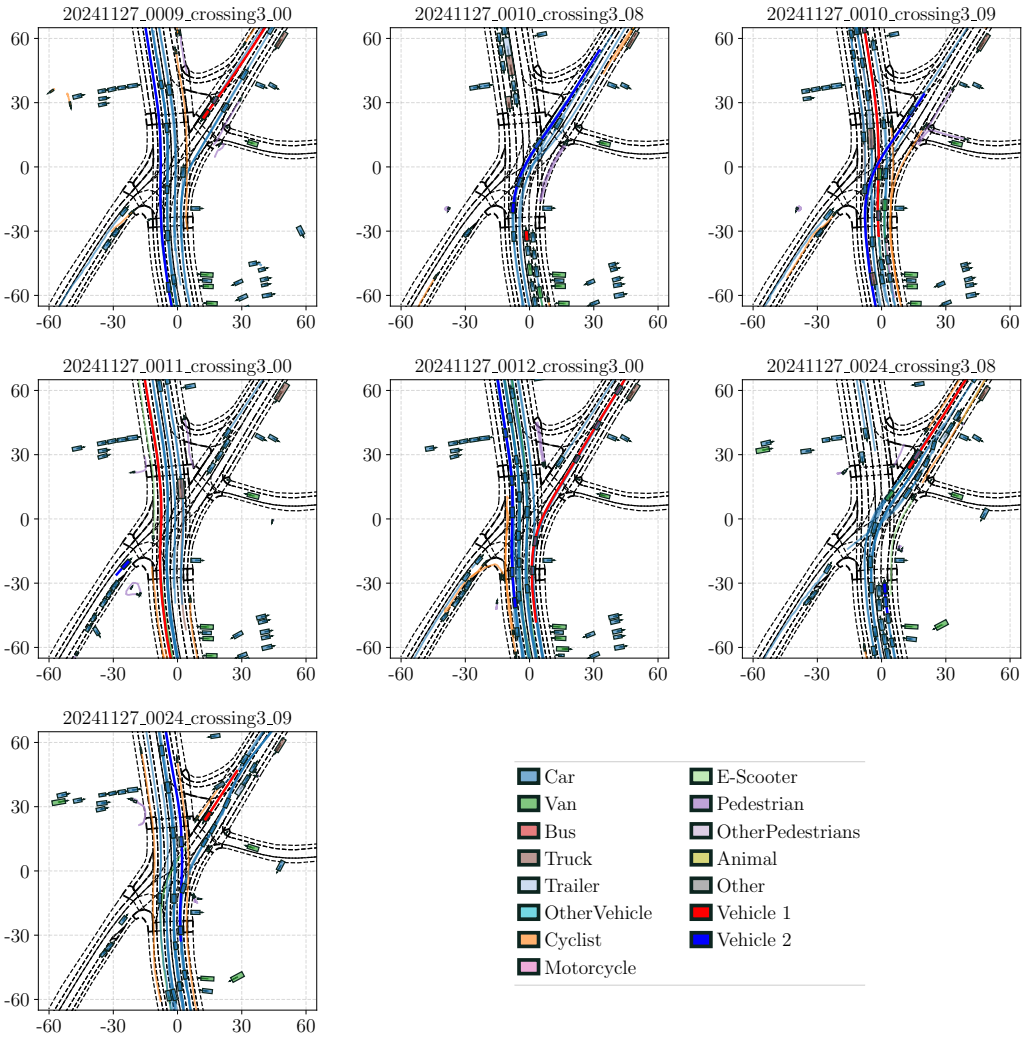


Figure 18: Visualization of trajectories at Intersection 3 across all sequences. Each subplot shows the trajectories of all annotated object classes for a sequence. The sequence names correspond to the original filenames in the dataset.

F.4 Intersection 1 multi-modal data visualization

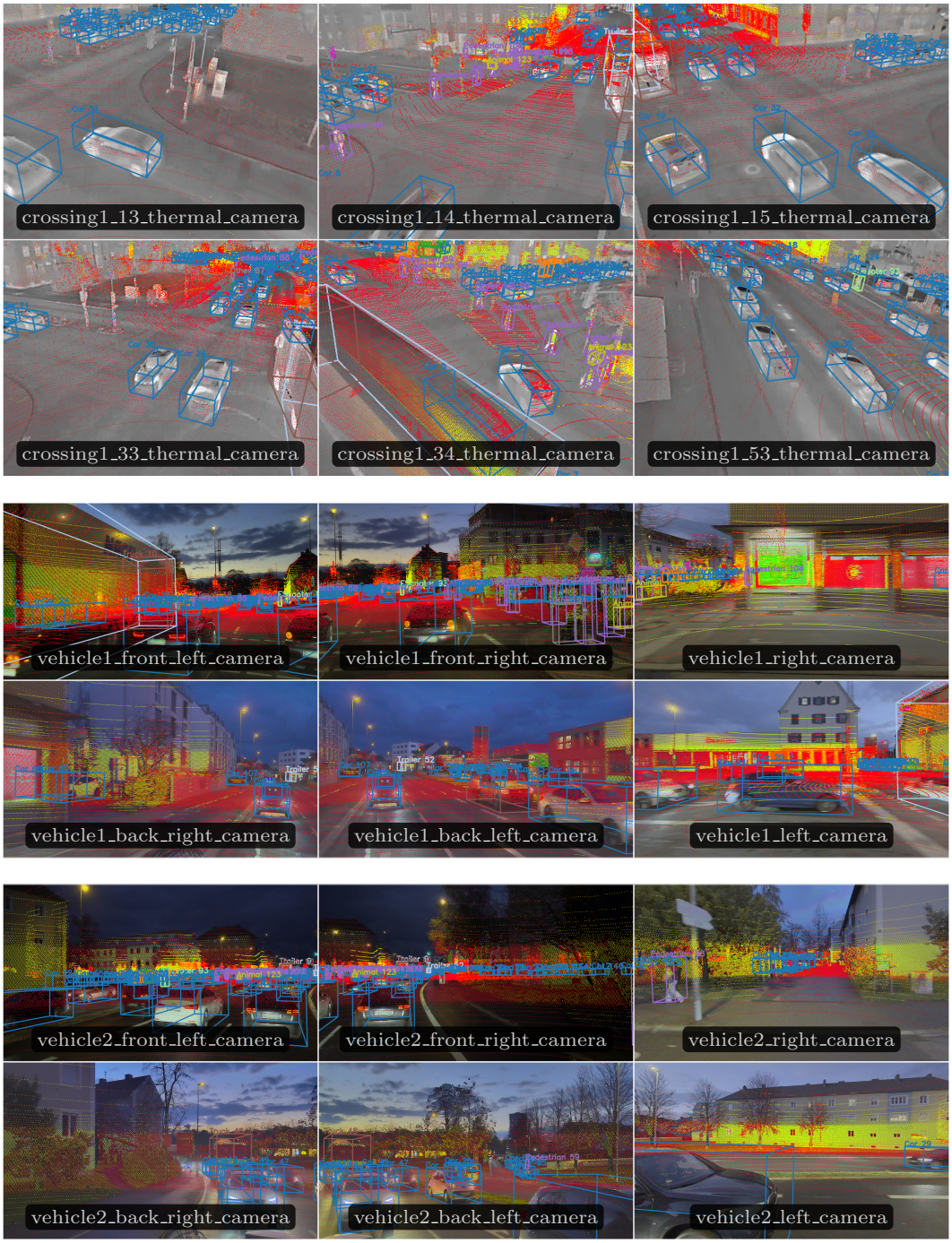


Figure 19: Multi-modal visualizations of Intersection 1 at a single timestamp, showing data from infrastructure thermal cameras (top), vehicle 1 RGB cameras (middle), and vehicle 2 RGB cameras (bottom).

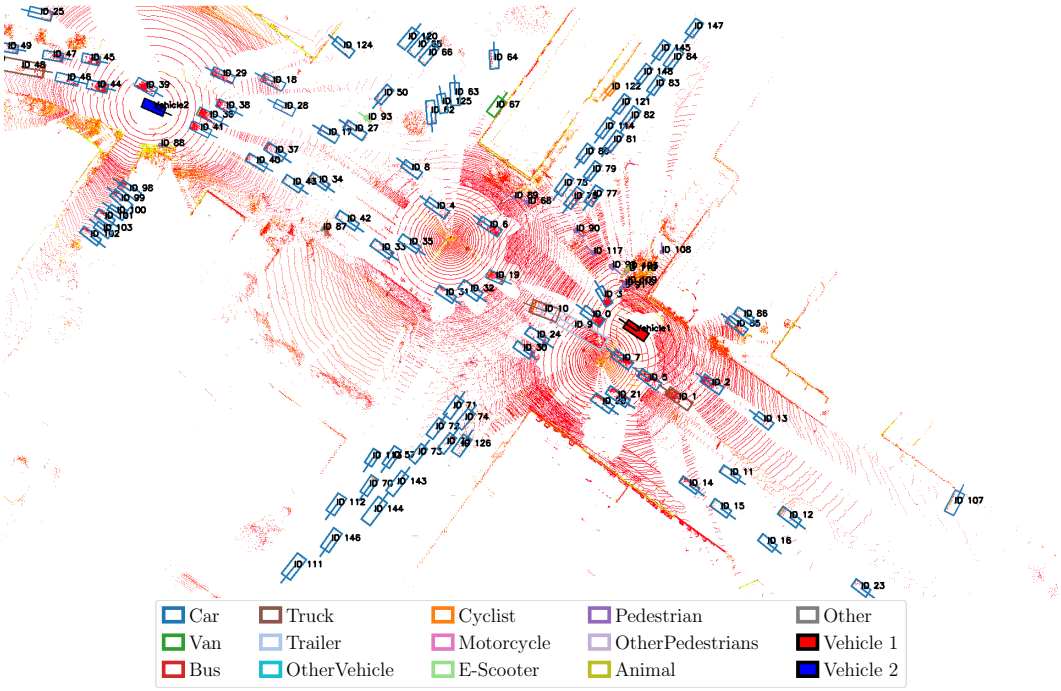


Figure 20: Visualization of the cooperative fused point cloud from all agents in Intersection 1, along with annotations.

F.5 Intersection 2 multi-modal data visualization

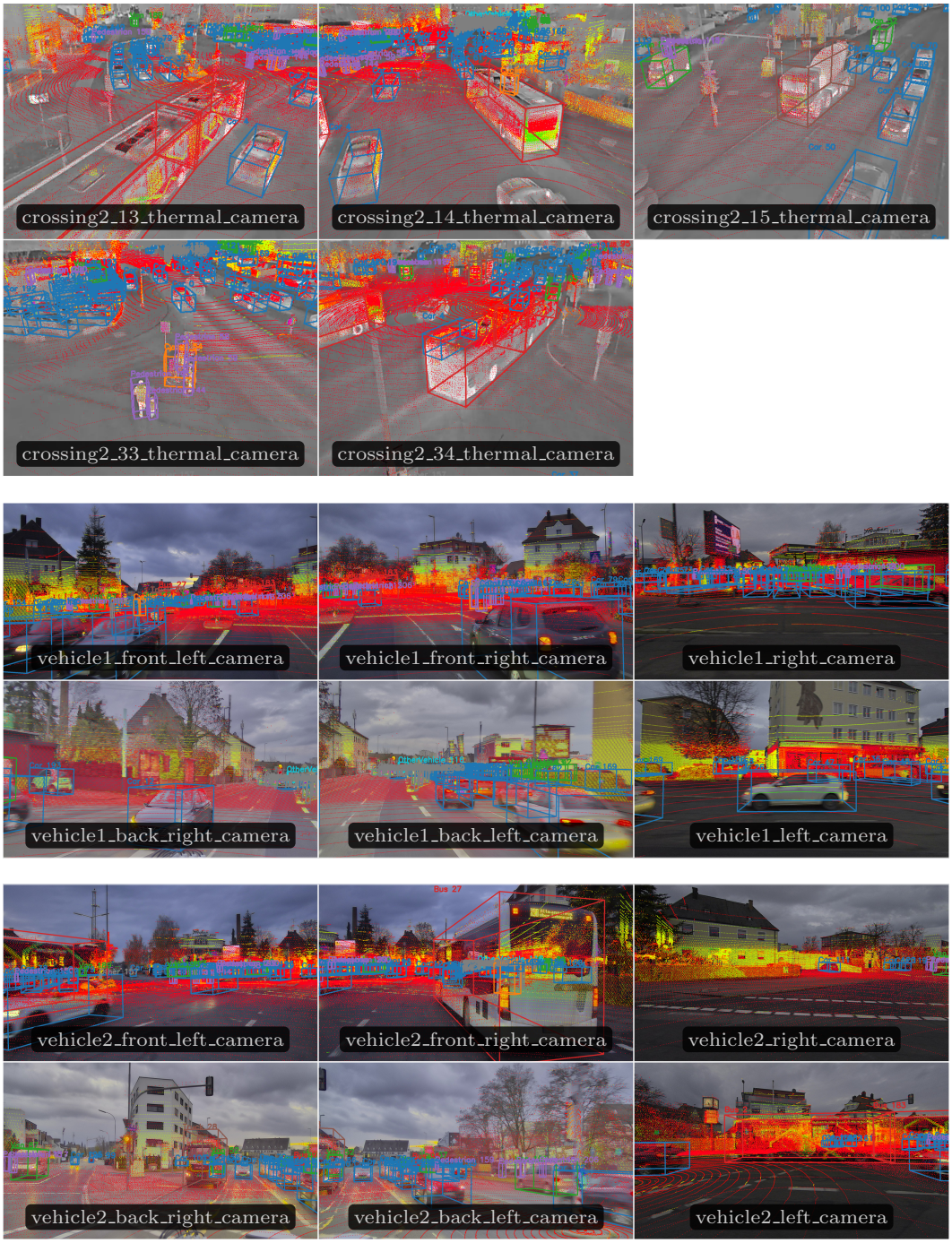


Figure 21: Multi-modal visualizations of Intersection 2 at a single timestamp, showing data from infrastructure thermal cameras (top), vehicle 1 RGB cameras (middle), and vehicle 2 RGB cameras (bottom).

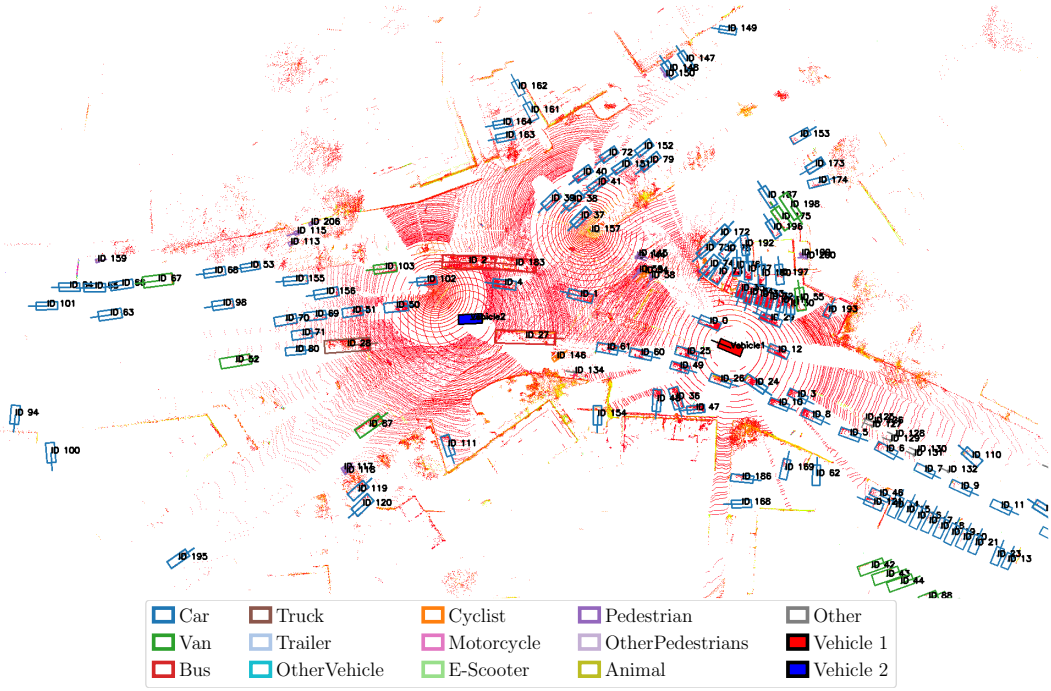


Figure 22: Visualization of the cooperative fused point cloud from all agents in Intersection 2, along with annotations.

F.6 Intersection 3 multi-modal data visualization

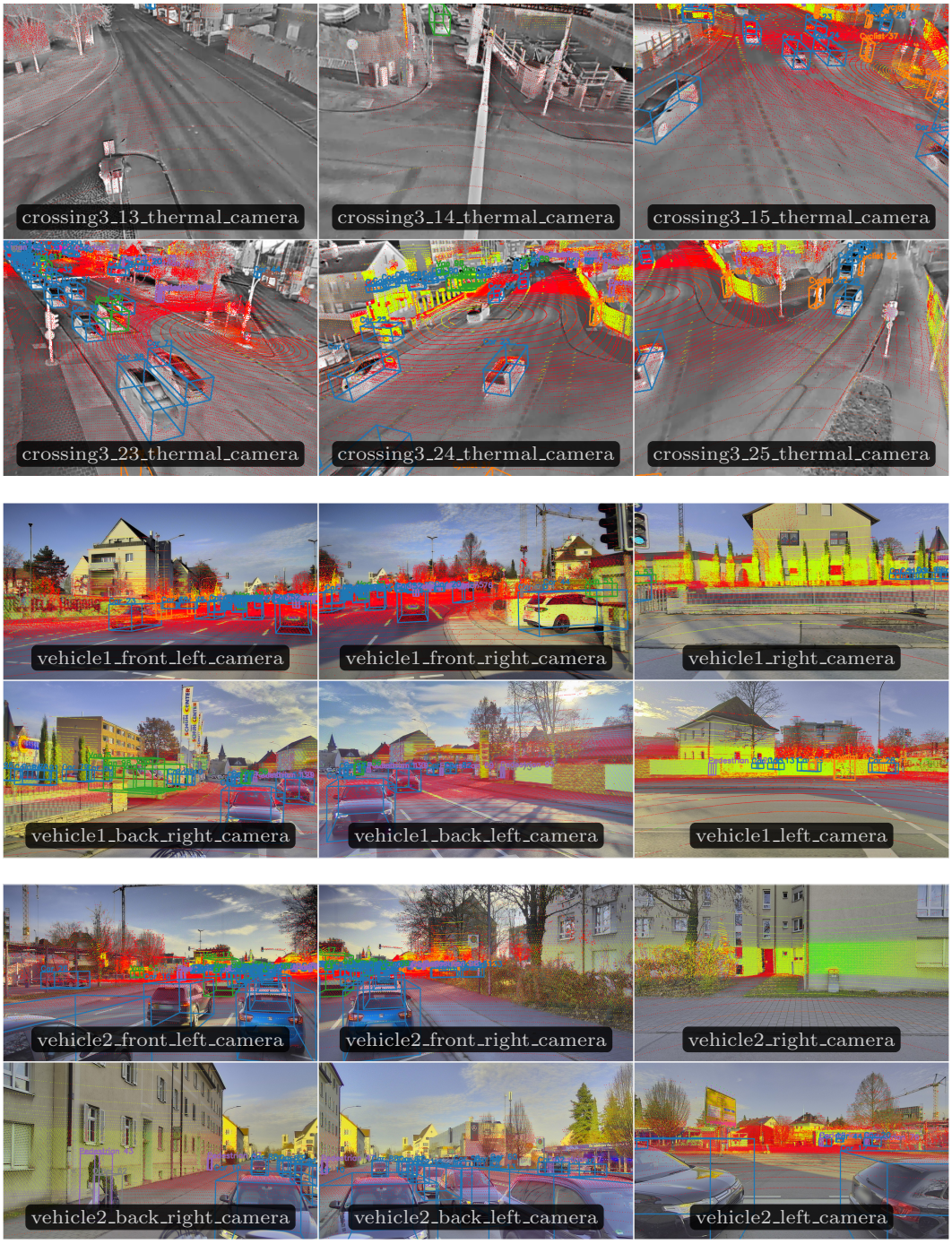


Figure 23: Multi-modal visualizations of Intersection 3 at a single timestamp, showing data from infrastructure thermal cameras (top), vehicle 1 RGB cameras (middle), and vehicle 2 RGB cameras (bottom).

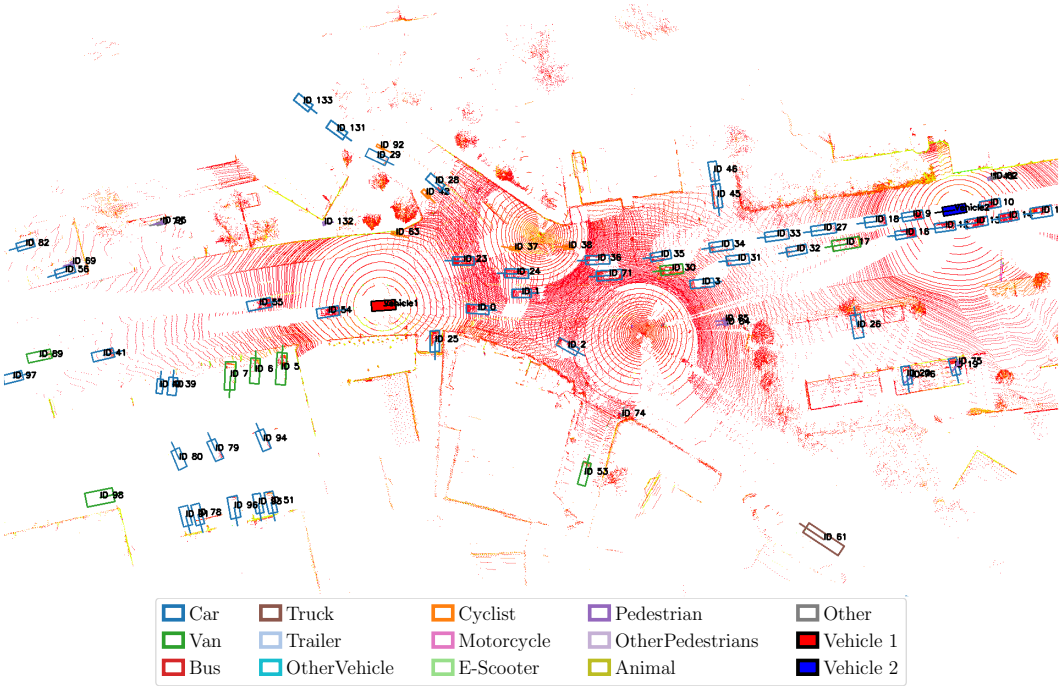


Figure 24: Visualization of the cooperative fused point cloud from all agents in Intersection 3, along with annotations.