

COOPERSCENE: Multi-Modal Cooperative Autonomy Benchmark with C-V2X Communication Characterization

Bo Wu*, Ruoshen Mo*, Justin Yue, Yanyu Zhang, Janice Nguyen, Guoyuan Wu,
Amit Roy-Chowdhury, Matthew J. Barth, and Hang Qiu†

University of California, Riverside, USA

{bwu109, rmo008, jyue009, yzhan831, jnguy172, guoyuanw, amitr, barth, hangq}@ucr.edu



Fig. 1: COOPERSCENE sample data visualization. Left: Overlay of point clouds from all agents (three connected autonomous vehicles (CAVs) in red, green, blue, and infrastructure in purple) with global 3D object labels (yellow). Middle: Individual agent sensor views in local frames. Right: Real time C-V2X throughput measured between all agent pairs.

Abstract. Cellular vehicle-to-everything (C-V2X) enables cooperative perception, prediction, and planning beyond the field of view of individual agents. However, existing datasets often overlook the complexities of real-world deployment, such as limited communication bandwidth and its dynamics, heterogeneous sensing modalities, and scalability beyond a single cooperative partner. In this paper, we introduce COOPERSCENE, a high-fidelity cooperative autonomy dataset with *real-world C-V2X communication characterization*. The dataset is organized into diverse *scenes*, including intersections, highway ramps, and parking lots. These scenes involve three connected and autonomous vehicles (CAVs) and one infrastructure roadside unit (RSU), all equipped with multi-modal sensors and commercial off-the-shelf C-V2X communication radios. All scenes are annotated with *globally consistent* 3D labels at 10 Hz, totaling *344K objects across 59K frames*, underpinned by *tight sensor- and agent-synchronization*, *centimeter-level localization and spatial alignment*, *precise cross-modality calibration*, and *3GPP-standard-compliant C-V2X communication*. COOPERSCENE establishes a rigorous benchmark for evaluating multi-agent scaling and actual performance in real-world deployable settings. Project website for data and benchmark: <https://cis1.ucr.edu/CooperScene>.

Keywords: Cooperative perception · C-V2X · Autonomous driving

* Equal contribution. † Corresponding author.

1 Introduction

Today’s autonomous vehicles (AVs) [5, 7] rely on a combination of onboard multimodal sensors to operate safely and efficiently. However, these autonomous agents are deployed as solitary entities that often suffer from incomplete situation awareness, making unilateral decisions, and stranding in long-tail events, waiting for remote operation [48]. As AVs become increasingly deployed, these limitations underscore the urgent need for a paradigm shift from *individual to cooperative autonomy* [13, 36, 45].

Cellular-vehicle-to-everything (C-V2X) technology breaks the solitary barrier by enabling vehicles to communicate with other vehicles (V2V), infrastructure (V2I) and vulnerable road users (V2P). V2X connectivity is estimated to prevent up to 595,000 vehicle crashes annually in the U.S., resulting in an annual economic savings of approximately \$71 billion [12]. Recent V2X-enabled cooperative perception [46, 53, 67], prediction [47], and driving [20, 21, 60] techniques emerge as a promising solution to address individual autonomy limitations. Driven by public cooperative perception datasets and benchmarks [16, 23, 43, 54, 56], these approaches have proven capable of overcoming occlusions, extending sensor range, and improving robustness in dynamic and unstructured environments [37, 55].

Despite the success on benchmarks, many models fall short in real-world deployment, partly due to a lack of datasets that reflect realistic V2X communication dynamics. Figure 2 summarizes the network bandwidth requirements of a few publicly benchmarked models. To achieve reasonable accuracy, the amount of latent features exchanged is often *orders of magnitude* higher than what is available today in C-V2X technology [34]¹ for individual agents in the shared Intelligent Transportation Systems (ITS) spectrum.

Moreover, most existing methods assume homogeneous sensing modalities, which limits their applicability to only a subset of traffic participants: those agents equipped with similar sensors. Meanwhile, a majority of publicly available *real-world* datasets offer only *one* cooperative agent (vehicle or infrastructure) besides the ego agent (see Table 1, column *Agent*), which also limits the benchmarking of multi-agent cooperation at scale.

To address these limitations, in this paper, we introduce COOPERSCENE, a multi-modal cooperative autonomy benchmark with C-V2X communication characterization. The dataset is organized into diverse “*scenes*”, each involving *three connected and autonomous vehicles* and *one infrastructure roadside unit*, all equipped with *multi-modal sensors* and commercial off-the-shelf *C-V2X communication radios*, interacting with each other across diverse real-world

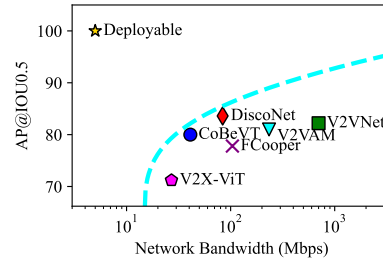


Fig. 2: Accuracy and Bandwidth of Open-source Cooperative Perception Methods Benchmarked on OPV2V [56] Dataset: A Research Gap towards Real-world Deployment.

¹ Measured by off-the-shelf C-V2X radios compliant with 3GPP Release 14 [11]

Table 1: Overview of *real-world* cooperative perception datasets in several key attributes, including sensing configuration (agents and modalities), localization and synchronization mechanisms, calibration availability, dataset scale, and supported benchmark tasks. **Green** cells indicate full support or strong capability, **yellow** denotes partial support, and **red** marks missing features. COOPERSCENE uniquely integrates C-V2X communication characterization into diverse multi-modal, multi-agent scenarios, establishing a rigorous benchmark for evaluating *deployable performance* in real-world settings.

Dataset	Agent		Modality		Scenario			Localization	Synchronization		Calibration		Statistics				Task		
	Veh.	Infra.	Cam.	LiDAR	V2X	Junc.	Ramp	Lot	Method	Sensor	Agent	Intr.	Extr.	Ped	Img	Seq	Obj	Det.	Mot.
DAIR-V2X-C [58]	1	1	✓	✓	-	✓	-	-	GPS+SLAM	-	-	✓	✓	39K	39K	100	464K	✓	-
V2V4REAL [54]	2	0	✓	✓	-	✓	-	-	GPS/IMU	-	-	✓	✓	20K	40K	41	240K	✓	-
V2X-Seq [59]	1	1	✓	✓	-	✓	-	-	GPS+SLAM	-	-	✓	✓	15K	15K	95	10K	✓	✓
TUM-Traf [69]	1	1	✓	✓	-	✓	-	-	RTK+ICP	ROS	NTP	✓	✓	2K	5K	30	30K	✓	-
COOPERSCE	3	1	✓	✓	✓	✓	✓	✓	RTK+ST-ICP	PTP	GNSS	✓	✓	59K	53K	24	344K	✓	✓

traffic scenarios, including urban intersections, freeway, local street, and parking areas, which introduce different traffic and occlusion patterns. To provide a high-fidelity benchmark, COOPERSCENE is built with *tight sensor- and agent-synchronization*, *centimeter-level localization and spatial alignment*, *precise cross-modality calibration* [61], and *3GPP-standard-compliant C-V2X communication*. All scenes are annotated at a frequency of 10 Hz with 3D bounding boxes in a globally consistent frame of reference, totaling approximately *344K labels across 59K frames*, along with synchronized pair-wise C-V2X throughput measurements. COOPERSCENE benchmarks both the performance and bandwidth of state-of-the-art cooperative perception methods to raise awareness of the research gap in both perception accuracy and communication efficiency. The summary of the contribution is as follows:

- **Unique integration of C-V2X in realistic cooperative scenes.** We introduce high-quality multi-agent interactive scenes collected from three vehicles and one infrastructure unit across diverse real-world scenarios. Each scene contains synchronized C-V2X measurements (latency, throughput, packet loss rate, and jitter) with multi-modal sensor streams (LiDAR, camera, and GNSS-RTK) and globally consistent 3D object labels.
- **Data quality refinement and validation.** We build COOPERSCENE with careful system design to guarantee high-quality data in terms of *localization*, *synchronization*, and *calibration*, using spatial-temporal ICP (ST-ICP), Precision Time Protocol (PTP), and motion capture systems and eye-hand calibration. We introduce a data validation procedure verifying the efficacy of these methods and the quality of the data collected.
- **Deployable cooperative autonomy benchmark.** Using synchronous C-V2X measurements, we provide the actual *deployable performance* benchmark for both detection and motion prediction tasks. In contrast to huge performance gains with infeasible communication assumptions, our practical benchmark raises awareness of the significant challenges in balancing model efficacy and communication efficiency.

COOPERSCENE serves as a cross-disciplinary benchmark to bridge the vision, robotics, and networking communities, paving the way towards realistic field-deployable cooperative autonomy.

2 Related Work

Cooperative Perception Datasets. Table 1 summarizes *real-world* cooperative perception datasets in contrast with COOPERSCENE. V2V4REAL [54] provides real-world cooperative perception benchmarks featuring multi-modal data from two vehicles. While it demonstrates the clear benefits of vehicle-to-vehicle (V2V) collaboration, it lacks infrastructure-side sensing. In contrast, DAIR-V2X-C [58] and V2X-Seq [59] incorporate a roadside unit (RSU) but pair it with only a single vehicle, thereby limiting cooperation to dual-agent settings. Furthermore, projecting the 3D point cloud into the 2D camera frame in these datasets reveals distinct spatial misalignments, where corresponding objects appear visibly shifted between the two modalities. Crucially, all three datasets report a large inter-agent synchronization error of 30–50 ms without providing a rigorous validation procedure. TUM-Traf [68] introduces a relatively smaller vehicle–infrastructure dataset but with high-precision localization; it incorporates an RSU with multi-modal sensors and refined poses via GNSS RTK and the Iterative Closest Point (ICP) algorithm [14]—a geometric registration method that iteratively minimizes the distance between two point clouds to determine their optimal alignment. Nevertheless, it remains limited to dual-agent settings.

While these datasets have significantly advanced cooperative perception research, several open challenges remain. Most benchmarks involve only one or two participating vehicles, limiting the study of multi-agent cooperation. Additionally, the precision of calibration, synchronization, and localization in these datasets is insufficient, potentially causing temporal drift and inter-agent and cross-modality misalignment. Recent efforts have begun to address multi-agent scaling: UrbanIng-V2X [41] collects data from multiple vehicles and infrastructure units across several intersections, substantially increasing the number of cooperative agents, though its multi-modal sensing remains limited. Most importantly, none of the existing datasets record real-world C-V2X throughput, leaving open questions about actual *achievable cooperative gains* due to latency, bandwidth, and packet loss in practical deployment settings.

COOPERSCENE addresses these gaps by providing three vehicles and one infrastructure in the same scene, enabling richer multi-agent interactions with realistic C-V2X communication characterization. The dataset includes multi-modal sensing (LiDAR, camera, GNSS/IMU), precise time synchronization via PTP and hardware triggering, and fully calibrated intrinsic and extrinsic parameters for all sensors. COOPERSCENE is, to our knowledge, the first real-world dataset to provide synchronized C-V2X communication traces alongside multi-agent multi-modal sensing, allowing joint analysis of perception quality and real communication performance. The dataset supports cooperative 3D detection and motion prediction, allowing comprehensive evaluation and benchmark.

Cooperative Perception Methods. Existing cooperative perception methods can be broadly categorized by the type of feature fusion and the underlying network architecture. Early fusion methods [18, 37, 38, 64, 65] directly share raw sensor data, such as point clouds or images, among connected agents, but they have substantial communication overhead and require strict synchronization.

Late fusion [39], on the other hand, aggregates the final object-level detections, which reduces communication cost but sacrifices spatial reasoning. Intermediate fusion [31, 40, 46, 53, 67] approaches strikes a middle ground by sharing compressed or encoded feature maps, enabling efficient collaboration with less bandwidth requirements. For example, V2VNet [46] represents vehicles as nodes and dynamically learns spatial message passing through edges, enabling agents to reason about occluded regions using neighboring information. V2X-ViT [55] introduces a hybrid transformer that alternates between intra-agent self-attention and inter-agent cross-attention to capture both local and global dependencies across agents. CoBEVT [53] further leverages Swin Transformers for multi-scale BEV fusion, achieving fine-grained spatial alignment. V2VAM [30] integrates intra-agent and inter-agent attention to dynamically weigh feature importance under lossy communication. These transformer-based frameworks improve robustness and scalability, at the cost of higher compute and data sharing volume.

The volume of these intermediate representations, while smaller than the raw data, remains well beyond what vehicular networks can support in real time. Figure 2 summarizes the performance and required bandwidth from these approaches [17, 30, 31, 46, 53, 55]. To address this bottleneck, recent communication-efficient methods [25, 35] explicitly optimize the bandwidth demand through feature selection: ERMVP [62] compresses shared features via filter-and-merge sampling, while CoSDH [52] performs supply-demand-aware region selection within a hybrid fusion scheme. However, what exacerbates the problem is the stringent decision deadline in these safety-critical autonomous systems. Extra latency incurred by data compression [15, 53], selection [25, 35], or codebook retrieval [26] further ages the stale data and risks missing the decision deadline.

Beyond perception, methods such as CMP [47], UniV2X [60], Coopernaut [20], and CoopReflect [21] extend cooperation to prediction and end-to-end driving. They jointly learn perception and planning using shared multi-agent observations, demonstrating the potential of cooperative gains in downstream autonomy tasks.

3 COOPERSCENE Dataset

In this section, we present COOPERSCENE, a multi-agent multi-modal dataset with C-V2X communication characterization. COOPERSCENE is collected across diverse real-world traffic scenarios. Specifically, the dataset contains 16 intersection scenes, 4 highway ramp scenes, and 4 parking lot scenes. The dataset features four participating agents: three vehicles equipped with **Cooperative Multi-Modal Sensing** platform (CMS) and one instrumented intersection infrastructure system. COOPERSCENE comprises 59K synchronized LiDAR frames and 53K image frames, annotated with 344K 3D bounding boxes, alongside GNSS, IMU, and C-V2X network traces. We provide details in sensing and communication platform setup, synchronization and calibration procedures, spatial alignment methodology, and data collection and annotation processes in this section.

3.1 Platform Setup

Cooperative Multi-Modal Sensing (CMS). To collect multi-modal sensor data, we design and build a CMS platform, a networked open-source portable universal sensing hub [51]. CMS provides an extensible platform with streamlined integration of multi-modal sensors, abstracting away the intricacies of calibration and synchronization, and it is easily reproducible and replicable to jump-start multi-purpose robot and multi-agent research.

Figure 3 shows CMS architecture. Specifically, the various sensors are connected to a central ROS node (*e.g.*, running on a PC or laptop) via a power-over-ethernet (PoE) switch. The PoE switch not only redirects the sensor stream via the ROS bus, but also provides power to various sensors. Due to the massive amount of data live-streamed, a switch with Small Form-factor Pluggable (SFP+) port is used to support data transfer rates of up to 10 Gbps. The GNSS module provides

microsecond-level accuracy pulse-per-second (PPS) signals to synchronize all sensors (§3.2). The transmitting/receiving (TX/RX) module enables CMS platforms to communicate with other platforms using vehicle-to-everything (V2X) communication technology. For COOPERSCENE data collection, each CMS platform is equipped with an Ouster OS1-128 LiDAR [4], a Lucid Triton 5.4 MP GigE camera [2], a TX/RX module (Cohda MK6 [1]) with cellular, WiFi, C-V2X transceivers and GNSS receivers, and a high-resolution Xsens MTi-680 RTK GNSS/INS [8] module. The CMS platform can be mounted on the vehicle roof.

Infrastructure Instrumentation. To capture a complementary infrastructure perspective, we deploy a similar CMS setup (except camera) at a busy intersection. One LiDAR sensor is mounted on a traffic light pole, and a GNSS-disciplined MK6 unit synchronizes the LiDAR clocks and communicates with vehicles. The consistent setup allows us to use the same data collection and processing procedures as those used by the CMS platform on the vehicle.

3.2 Sensor and Agent Synchronization

Synchronization is critical for data integrity and consistency, as asynchronous data mismatches in time domain and confuses downstream processes. However, depending on the available synchronization support from the sensors, the task

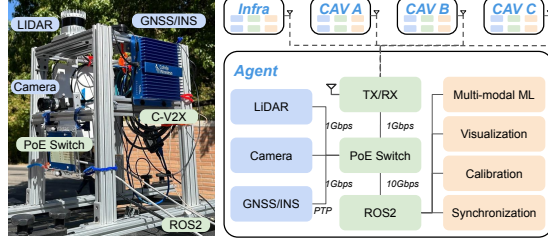


Fig. 3: CMS Overview. CMS integrates LiDAR, camera, GNSS with a power-over-ethernet (PoE) switch, which forwards the data to a central ROS node (running on a laptop). The laptop and sensors are synchronized with GNSS time, and all intrinsic and extrinsic parameters are calibrated for all sensors. TX/RX module communicates with other CMS platforms on both vehicles and the infrastructure.

can become non-trivial. We first tackle *onboard* synchronization and then discuss *cross-platform* synchronization.

Onboard Sensor Synchronization. Although ROS provides built-in time synchronization by timestamping sensor data when the central node receives published messages, it is inaccurate for multi-modal sensors due to network overhead and operating system scheduling. Moreover, each sensor has its own local oscillator offset, leading to asynchronous data collection. To address this issue, we use PTP [9] to synchronize sensor clocks and apply hardware triggering to ensure LiDAR and camera data are captured simultaneously. In CMS platform, Cohda MK6 acts as the PTP master and disciplines all other devices, including the LiDAR, camera, and the laptop. Beyond synchronizing the local clocks, to further guarantee synchronized capture between the LiDAR and camera, we use a General-Purpose Input/Output (GPIO) pin from the LiDAR to generate a stable 10 Hz synchronization pulse that triggers the camera to capture images. Specifically, the camera captures an image upon detecting a rising edge of the pulse, which is generated at the same time the LiDAR begins data collection.

Cross-platform Agent Synchronization. With onboard sensors synchronized, the next challenge is to align time across multiple platforms. Achieving cross-platform synchronization requires a common global time reference. Prior work has explored both centralized signals (*e.g.*, WiFi [24], cellular [49]) and distributed protocols [22, 33] in various contexts. For outdoor autonomous driving scenarios, where GNSS signals are available, we use the GNSS signal to discipline the Cohda MK6. Cohda MK6 then serves as a grandmaster clock to synchronize the local clocks of onboard sensors with PTP support. Because each agent’s grandmaster clock is locked to the same global GNSS time, all sensors across all platforms achieve accurate global time alignment.

3.3 Spatial Alignment

GNSS/IMU errors and noises in real-world systems often lead to spatial misalignment among LiDAR frames collected from different agents. ICP [14] is widely used to align two point clouds to mitigate this problem. However, standard ICP cannot be directly applied to the multi-agent scenarios in COOPERSCENE. To address this, we propose a spatial-temporal ICP approach, as illustrated in Fig. 4. First, we use GNSS positions and IMU orientations to project all LiDAR frames from each agent’s local coordinate system into a global East-North-Up (ENU) frame centered at the intersection. This step computes the transformation matrix T_{local}^{global} . Next, for each vehicle, we search its entire sequence to find the frame with the maximum

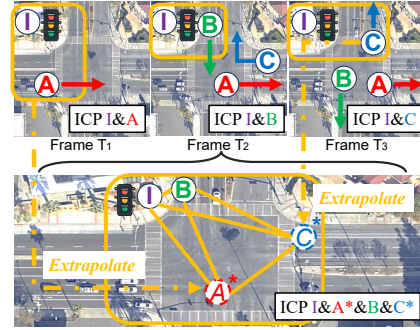


Fig. 4: Multi-agent Spatial-temporal Pose Alignment

LiDAR overlap with the infrastructure’s LiDAR. We run ICP on this single, optimal frame pair to find a coarse adjustment transform $T_{PairwiseICP}$. Finally, given the high accuracy and temporal stability of GNSS-RTK positioning, the relative distance changes within the same agent remain consistent across frames. We therefore extrapolate the adjustment found in each vehicle-infrastructure pair to a frame where all agents are present and closest to each other; the frame that minimizes the total inter-agent distance. In this extrapolated “joint” frame, we perform another round of ICP to simultaneously align all vehicles and the infrastructure, obtaining the global transformation $T_{JointICP}$. Each vehicle’s final spatial correction matrix is computed as: $T = T_{JointICP} \times T_{PairwiseICP} \times T_{local}^{global}$. After spatial-temporal ICP alignment, all vehicles are aligned with the infrastructure and mutually aligned with one another.

3.4 Sensor Calibration

To better leverage the complementary nature of different sensing modalities, one key insight in multi-modal sensing is to use the returns of different sensors from the same object to enhance perception accuracy and confidence. The task of sensor calibration is to figure out the intrinsic and extrinsic parameters of all sensors, such that the downstream processing pipeline can understand the *correspondence* of the data from each modality. The device intrinsic (*e.g.* camera focal length and principal point) is unique to each sensor. The calibration methods to obtain these intrinsics are well-studied [44, 66]. In this section, we focus on extrinsic (relative pose) calibration [61] between multi-modal sensors. Existing methods typically fall into two categories: target-based approaches [63] that rely on predefined geometries (*e.g.*, checkerboards or AprilTags), and targetless, data-driven approaches [27, 61]. However, these conventional techniques either focus on a single modality pair or fail to achieve sufficient alignment precision.

To get groundtruth-level pair-wise calibration, one approach is to use a motion capture system (MoCap) [3, 6], putting multiple markers on each device to get the pose estimate. While MoCap gives millimeter-level accuracy, the limitation of this approach is that the marker placement outside the sensor cannot represent the optical model inside the sensor, hence making it difficult to calculate the origin of each sensor’s coordinate system. The issue can be rectified by solving eye-hand calibration [42]. For a particular sensor, we formulate the eye-hand problem as solving for X in $AX = XB$, where A and B are the relative motions between the markers’ poses in MoCap ($T_{markers}^{mocap}$) and the transforms from a fiducial marker to the sensor. Eye-hand then yields $T_{sensor}^{markers}$, the transform from a sensor’s optical center to the markers, and we perform the following calculation to find the true pose of the sensor in MoCap: $T_{sensor}^{mocap} = T_{markers}^{mocap} * T_{sensor}^{markers}$.

After finding the true sensor poses, we can accurately calculate the extrinsics between sensors. Given sensor poses T_{S_1} and T_{S_2} for sensors S_1 and S_2 , the extrinsic from S_1 to S_2 is $T_{S_1}^{S_2} = (T_{S_2})^{-1} * T_{S_1}$. With the extrinsics between sensors, we can also leverage a few modality-specific pair-wise calibrations to form a chain of relative poses. One possible chain of extrinsic parameters is as follows: 1) GNSS-to-LiDAR/IMU extrinsic (T_G^L). 2) LiDAR/IMU-to-camera

extrinsic (T_L^C). With this chain, one can calculate any pair-wise transformation matrices. For instance, to find GNSS and camera correspondence, we only need to calculate $T_G^C = T_G^L * T_L^C$.

3.5 Data Acquisition

Scene Configuration and Data Collection. By leveraging CMS platforms and the instrumented road-side unit, we can now stage diverse vehicle encountering scenarios, *e.g.* at intersections, highway ramps, and parking lots. In these interactive scenarios, we record multimodal data, including LiDAR, images, GNSS, and C-V2X network traces with busy background traffic. The data is collected in multiple continuous takes. Each take spans from the vehicles’ departure at various starting points to their convergence at a designated encounter zone, concluding once all vehicles exit the area. To ensure scenario diversity, we vary the initial positions and driving routes across takes, while coordinating the vehicles to arrive at the target zone simultaneously.

Annotation. After data collection, each sensor frame is assigned a unique frame identifier based on its *globally-synchronized* sensor timestamp. Frames across different agents are synchronized using these identifiers. In cases of frame loss, the most recent valid frame is reused to ensure that all frame sequences remain temporally consistent and of equal length. To reduce manual annotation effort, we develop a semi-automatic labeling pipeline that integrates an object detection backbone [32], an object tracking module [50], and a late fusion module. Specifically, the detection backbone first generates coarse bounding boxes² for each agent’s data. The tracking module then refines the bounding box positions and orientations. Then all annotations from each agent are transformed in globally consistent ENU frame, where the late fusion module is applied to refine box poses and remove redundant boxes based on the Intersection-over-Union (IoU) of detected objects. The fusion process is performed in two stages: in the first stage, overlapping boxes with high IoU values are averaged to merge their spatial information; in the second stage, overlapping boxes are further filtered by retaining only those with the highest confidence scores. Finally, the tracking module is reapplied to refine object positions and orientations, producing preliminary labels. These preliminary labels are then manually verified and refined by trained professional annotators using [29], producing the ground-truth labels.

4 COOPERSCENE Benchmark

In this section, we present a comprehensive benchmark evaluation on COOPERSCENE, focusing on object detection and motion prediction. In the benchmark, the dataset is divided into train, validate, and test subsets with a ratio of 5:2:3. We conduct all experiments on a unified cooperative perception pipeline developed on top of MMDet3D [19] and OpenCOOD [56]. Our unified pipeline integrates

² The initial release focuses on the *Car* class, future releases will expand object classes.

a configurable C-V2X network simulator and provides a consistent *multi-ego* training and evaluation protocol. We retrained all perception and prediction baseline models on the COOPERSCENE dataset to allow them to adapt to the characteristics and sensor configurations, avoiding domain transfer gaps.

Metrics. For perception, we adopt the mean Average Precision (mAP) as the primary metric, with 2D (BEV) and 3D IoU thresholds of (0.3, 0.5, 0.7). For motion, we use Minimum Average Displacement Error (minADE), and Minimum Final Displacement Error (minFDE) evaluated at different horizons (1s, 3s, 5s). In our *multi-ego* pipeline, we iteratively treat each of the three vehicles as the ego vehicle, with the remaining vehicles and the infrastructure as cooperators. The evaluation reports results averaged over all ego vehicles.

Baselines. To build a comprehensive benchmark on COOPERSCENE, we evaluate state-of-the-art cooperative perception models, including graph-based, transformer-based fusion algorithms and recent communication-efficient methods. All models are retrained using consistent parameters, including voxelization size and detection range for fair comparison. Furthermore, we benchmark the cooperative motion prediction performance following the perception–prediction (P&P) setting. The baselines include:

- **V2VNet** [46]: A graph-based intermediate fusion model. Vehicles share encoded features through learnable edge connections.
- **V2VAM** [30]: An attention-based transformer model, which captures intra-agent and inter-agent attention, dynamically weights agent contributions.
- **V2X-ViT** [55]: A vision transformer-based BEV fusion architecture composed of multi-agent self-attention and multi-scale window self-attention layers.
- **CoBEVT** [53]: A Swin-Transformer-based BEV fusion architecture that applies multi-scale window attention for spatial feature alignment.
- **CMP** [47]: A cooperative motion prediction framework built on top of the CoBEVT perception backbone, which aggregates predictions from multiple vehicles using a transformer-based motion decoder.
- **ERMVP** [62]: A communication-efficient and collaboration-robust model that compresses shared features via filter-and-merge sampling and calibrates pose induced misalignment with consensus sparse features.
- **CoSDH** [52]: A communication-efficient model using supply–demand-aware region selection and an intermediate–late hybrid fusion scheme to balance accuracy and bandwidth.

Detection Benchmark Results. Table 2 summarizes the perception benchmark. We benchmark each baseline over five cooperative settings; **1) V+I**: with only infra, **2) V+V**: with only one peer vehicle, **3) V+V+I**: with one vehicle and infra, **4) V+2V**: with two peer vehicles, and **5) V+2V+I**: with all three cooperative agents. In addition to showing the mAP under C-V2X bandwidth constraints, we also benchmark the performance with an unlimited network assumption (in lightgray color) as done in other existing benchmarks [54, 69], to show the original benefits of fusing rich features from multiple vantage points.

Overall, all cooperative perception models benefit from information sharing among vehicles and infrastructure. As the number of cooperative agents grows,

Table 2: Benchmark of cooperative perception under C-V2X and **unlimited** (infeasible) network, with increasing number of agents, in terms of accuracy (mAP), total data sharing size per frame, and the latency if all data were to be transmitted over C-V2X.

Model	Agents	mAP (C-V2X)			mAP (Unlimited)			Sharing	Latency (ms)
		@0.3	@0.5	@0.7	@0.3	@0.5	@0.7	Size (MB) under C-V2X	
V2VNet [46]	V+I	0.43	0.33	0.20	0.55	0.42	0.23	10.0	64,349
	V+V	0.35	0.25	0.15	0.66	0.55	0.34	10.0	64,038
	V+V+I	0.36	0.27	0.15	0.70	0.60	0.37	20.0	76,879
	V+2V	0.32	0.23	0.13	0.79	0.74	0.47	20.0	72,773
	V+2V+I	0.32	0.23	0.13	0.81	0.75	0.50	30.0	107,432
V2X-ViT [55]	V+I	0.50	0.38	0.20	0.54	0.41	0.21	0.338	2,211
	V+V	0.53	0.43	0.26	0.65	0.55	0.35	0.338	1,305
	V+V+I	0.56	0.45	0.26	0.70	0.59	0.36	0.667	9,215
	V+2V	0.60	0.53	0.33	0.79	0.73	0.49	0.667	7,022
	V+2V+I	0.60	0.52	0.33	0.80	0.74	0.50	1.014	9,976
V2VAM [30]	V+I	0.53	0.40	0.22	0.57	0.44	0.24	0.423	2,264
	V+V	0.53	0.43	0.26	0.68	0.56	0.37	0.423	2,253
	V+V+I	0.58	0.47	0.27	0.74	0.63	0.39	0.846	10,301
	V+2V	0.61	0.54	0.34	0.82	0.76	0.53	0.846	8,117
	V+2V+I	0.62	0.54	0.34	0.84	0.78	0.53	1.269	11,302
CoBEVT [53]	V+I	0.48	0.39	0.20	0.52	0.43	0.21	0.500	2,707
	V+V	0.50	0.41	0.26	0.64	0.55	0.37	0.500	2,393
	V+V+I	0.54	0.44	0.25	0.70	0.60	0.37	1.000	11,059
	V+2V	0.57	0.49	0.33	0.80	0.73	0.54	1.000	8,987
	V+2V+I	0.59	0.50	0.32	0.82	0.75	0.54	1.500	12,381
ERMVP [62]	V+I	0.61	0.46	0.24	0.65	0.51	0.26	0.346	1,730
	V+V	0.61	0.49	0.30	0.76	0.63	0.42	0.346	1,730
	V+V+I	0.63	0.51	0.31	0.79	0.67	0.43	0.692	3,460
	V+2V	0.66	0.57	0.37	0.87	0.80	0.57	0.692	3,460
	V+2V+I	0.64	0.56	0.36	0.87	0.81	0.57	1.038	5,190
CoSDH [52]	V+I	0.57	0.43	0.23	0.58	0.44	0.24	0.0595	298
	V+V	0.63	0.47	0.29	0.67	0.54	0.35	0.0595	298
	V+V+I	0.68	0.53	0.32	0.72	0.62	0.39	0.1190	595
	V+2V	0.74	0.61	0.40	0.79	0.72	0.51	0.1190	595
	V+2V+I	0.75	0.61	0.41	0.81	0.73	0.52	0.1785	893

almost all (except V2VNet) methods achieve increasingly better detection performance than non-cooperative settings (Table 3), under both C-V2X and unlimited settings, across different IoU thresholds. However, the mAPs under realistic C-V2X bandwidth are upto 49% lower than the grayed-out numbers under an unlimited network. This occurs because the required data transmission size exceeds C-V2X network capabilities, which results in late data delivery. The drastic difference shows the gap between idealized perception and bandwidth-constrained real-world deployment.

Secondly, for *non-communication-efficient* methods (the top four), due to the mismatch between the sharing volume and the limited C-V2X network capacity, cooperation can only happen *very infrequently*, ranging from every 2 to 100s of seconds, while hundreds of frames in between the interval can not benefit from cooperation. Consequently, agents receive *outdated, stale* data, which fails to improve, but can actively harm, the perception accuracy. Furthermore, adding more participants only exacerbates channel contention and increases latency. Hence, the benefit of an additional vantage point may effectively degrade end-to-end performance due to the increased staleness (*e.g.* for V2VNet, and CoBEVT at V+2V+I, where sharing size is large). In contrast, the communication-aware baselines ERMVP [62] and CoSDH [52] suffer less from this contention. They transmit only the most informative features within a bandwidth budget,

Table 3: Single agent perception performance using different backbones.

Agent	Method	mAP@0.3	mAP@0.5	mAP@0.7
Infra	PointPillars [28]	0.24	0.09	0.04
	SECOND [57]	0.20	0.11	0.06
CAV A	PointPillars [28]	0.37	0.20	0.08
	SECOND [57]	0.36	0.23	0.11
CAV B	PointPillars [28]	0.38	0.20	0.08
	SECOND [57]	0.35	0.24	0.13
CAV C	PointPillars [28]	0.34	0.18	0.07
	SECOND [57]	0.34	0.22	0.11

Table 4: Multi-modal multi-agent cooperative perception benchmark.

Agent	Modality	mAP(BEV)			mAP(3D)		
		@0.3	@0.5	@0.7	@0.3	@0.5	@0.7
V	LiDAR	0.78	0.56	0.31	0.70	0.38	0.21
	LiDAR+Cam	0.80	0.57	0.33	0.72	0.42	0.23
V+V	LiDAR	0.81	0.65	0.39	0.75	0.54	0.22
	LiDAR+Cam	0.80	0.65	0.41	0.74	0.54	0.25
V+2V	LiDAR	0.90	0.84	0.56	0.87	0.73	0.30
	LiDAR+Cam	0.90	0.83	0.61	0.87	0.74	0.35

substantially reducing the per-frame sharing size. While they achieve the strongest deployable accuracy under C-V2X among all baselines, the latency does not scale gracefully and is still too long for vehicles to make safety-critical reactions. This result corroborates the value of bandwidth reduction research [25, 35] in cooperative perception, motivating future research to further resolve the *mismatch* between sharing size and network capacity.

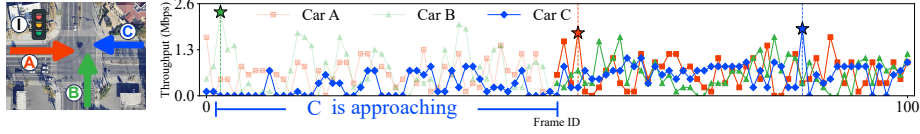
Multi-modal Cooperative Perception. To benchmark multi-modal cooperative perception, we developed a multi-modal cooperative framework based on top of BEVFusion [32]. Table 4 compares the performance under LiDAR-only and LiDAR+Camera configurations. The results show that when cooperative perception is combined with multi-modal sensing, the performance gain becomes more significant, especially at *higher IoU thresholds*, where the model benefits more from the complementary modality. This indicates that both inter-agent collaboration and multi-modal feature fusion contribute to perception understanding.

Motion Prediction Benchmark Results. To further evaluate the capability of COOPERSCENE in supporting downstream tasks, we benchmark the cooperative motion prediction performance of V2VNet [46] and CMP [47] following the perception-prediction (P&P) setting. In this setting, cooperative perception first aggregates multi-agent features, which are then used by the motion prediction model to predict future trajectories. In the P&P setting, the communication cost is dominated by the perception feature exchange, while the data exchange in the subsequent prediction module (*e.g.* in CMP) is negligible. Therefore, we omit the communication characteristics, and refer to those reported in Table 2 under different settings. The quantitative results are summarized in Table 5. Overall, cooperative perception improves motion prediction accuracy, especially for *longer prediction horizons* for both models. Between the two models, CMP outperforms V2VNet across all evaluation horizons. This performance gap highlights the advantage of a transformer-based motion decoder in modeling complex multi-agent interactions. V2VNet’s short-horizon results are interesting, as it is counterintuitive. We can only suspect that the additional information from peer vehicles disturbs the intermediate fusion for short-horizon predictions. We believe there is still a huge research space into cooperative motion prediction, as well as other downstream tasks beyond detection.

C-V2X Network Performance Analysis. Figure 5 shows the variation of throughput across frames. The C-V2X network exhibits highly dynamic behavior

Table 5: COOPERSCENE benchmark on cooperative motion prediction

Model	Agent	minADE@1s↓	minADE@3s↓	minADE@5s↓	minFDE@1s↓	minFDE@3s↓	minFDE@5s↓
V2VNet [46]	V	0.4895	2.1648	4.4538	0.9307	5.7344	10.6390
	V+V	0.5424	1.5611	3.5591	0.7528	4.4820	9.1667
	V+2V	0.7610	1.3796	2.9096	1.0079	3.4762	7.3660
CMP [47]	V	0.3851	0.8421	1.4022	0.5786	1.6125	3.0712
	V+V	0.4076	0.9330	1.6416	0.6143	1.8492	3.7389
	V+2V	0.3214	0.7252	1.2893	0.4723	1.4551	2.9716

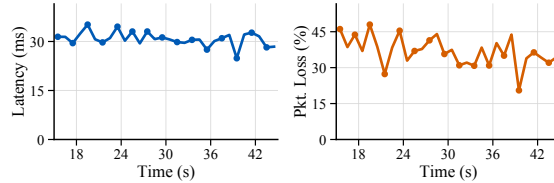
**Fig. 5:** C-V2X Communication throughput variation across different frames. Pentagram denotes the peak throughput.

due to vehicle mobility and channel contention among multiple agents. As a result, no single agent can occupy the channel continuously, and the effective throughput fluctuates significantly over time. These observations highlight the need for careful spectrum sharing design and motivate the use of more adaptive cooperation strategies [35]. In addition, we empirically evaluated the feasibility of transmitting cooperative perception data using the baselines. The data size and corresponding transmission time are summarized in Table 2. The data were segmented into 2100 bytes chunks and transmitted sequentially over the C-V2X radio, with each packet sent at a 1 ms interval³. All models require more than 100 ms to deliver a single frame, resulting in outdated data and providing no significant improvement in perception performance.

While bandwidth is the dominant bottleneck, other communication characteristics also affect end-to-end performance. To isolate their impact, we remove the bottleneck by conducting an oracle-style experiment: each CAV transmits only the down-sampled LiDAR points bounded by ground-truth boxes, reducing the data volume below the C-V2X capacity. As shown in Table 6, even when bandwidth is

Table 6: Evaluation under different network settings

Method	AP ₃₀	AP ₅₀	AP ₇₀	Pkt. Loss (%)	Latency (ms)	Tput (Mbps)
Local	0.18	0.15	0.04	—	—	—
C-V2X	0.21	0.19	0.05	38.19	31.46	0.72
Unlimited	0.24	0.21	0.06	0	<1	1.25

**Fig. 6:** Latency and packet loss rate under C-V2X

sufficient in both C-V2X and unlimited, packet loss and latency still degrade detection accuracy relative to the unlimited setting. Figure 6 further shows that such packet loss and latency fluctuate over time and are inevitable in real-world

³ Corresponding to the minimum time resource in C-V2X standard [10]

channels due to vehicle mobility, interference, and channel contention. This indicates that resolving the bandwidth gap alone is insufficient; robustness to packet loss and latency is equally important for deployable cooperative perception.

Overall, the results reveal a substantial gap between the assumptions made in existing works and the actual operational characteristics of real-world C-V2X networks, underscoring the importance of network-aware cooperative perception design, which is a central goal of COOPERSCENE.

5 Data Quality Validation

Synchronization. Figure 7 presents an example synchronization sequence from the COOPERSCENE dataset. Figure 7a shows onboard LiDAR and camera sequences capturing pedestrians crossing the intersection before and after synchronization. With PTP and hardware triggering synchronization, the LiDAR and camera captures the motion of the pedestrian at the same time compared to asynchronous data. Figure 7b displays camera images from multiple vehicles capturing a pedestrian walking along the pathway, demonstrating consistent motion across agents and confirming effective cross-platform synchronization.

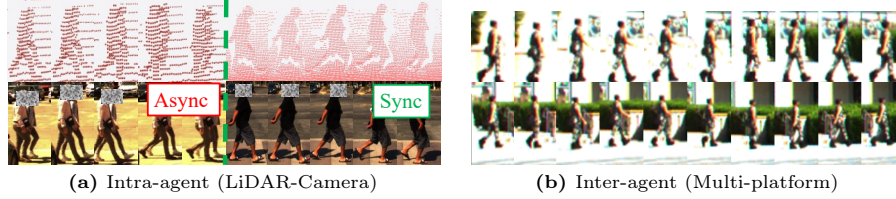


Fig. 7: Visualization of synchronized sequences in COOPERSCENE.

Spatial Alignment. We quantify spatial alignment accuracy using the root mean square error (RMSE) of overlapping areas between synchronized LiDAR frames for each agent pair. Specifically, we run ICP once more on the already aligned frames, and compute the RMSE of the Euclidean distances between the matched point correspondences, measuring the misalignment between agents after convergence. Figure 8a shows the RMSE of COOPERSCENE under different spatial alignment methods. No ICP indicates that LiDAR frames are projected into the ENU coordinate system described in §3.3 without further correction. Pairwise ICP aligns each vehicle to the infrastructure after ENU projection. Spatial-temporal ICP, which considers all agents jointly, substantially improves alignment accuracy,

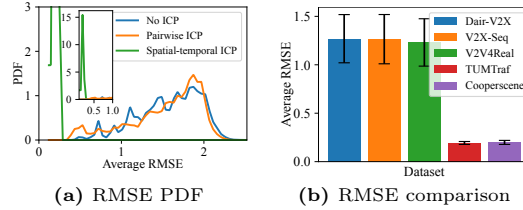


Fig. 8: Spatial alignment error comparison between different methods (a), and among other datasets (b).

Figure 8a shows the RMSE of COOPERSCENE under different spatial alignment methods. No ICP indicates that LiDAR frames are projected into the ENU coordinate system described in §3.3 without further correction. Pairwise ICP aligns each vehicle to the infrastructure after ENU projection. Spatial-temporal ICP, which considers all agents jointly, substantially improves alignment accuracy,

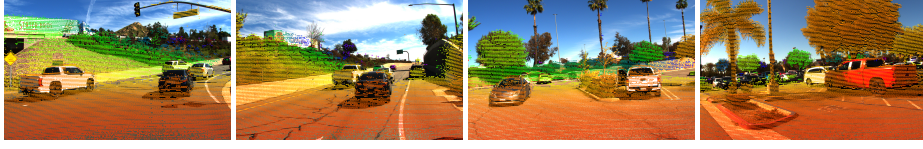


Fig. 9: LiDAR-to-camera overlays using calibrated extrinsic.

achieving an RMSE of 0.2 m for most frames. Figure 8b compares the RMSE of COOPERSCENE with other real-world datasets. COOPERSCENE achieves the lowest average RMSE among all datasets except TumTraf. TumTraf aligns a single vehicle, whereas COOPERSCENE must jointly align multiple agents. Furthermore, TumTraf applies ICP every 10 frames, while COOPERSCENE achieves comparable performance with difference in only 0.01 m by applying ICP on just a single frame per vehicle, highlighting the efficiency and robustness of our spatial alignment.

Calibration. Figure 9 shows the results of the extrinsic calibration as LiDAR-Camera overlays, where the LiDAR’s point cloud aligns with objects such as buildings, cars, trees, and poles in the image. We provide a qualitative analysis of our calibration by confirming accurate extrinsics through reprojection analysis. Specifically, we use the fiducial tag’s pose $\mathcal{T}_{tag}^{mocap}$ and the sensor’s pose $\mathcal{T}_{sensor}^{mocap}$ in the MoCap frame. By transforming and projecting the fiducial tag’s corners into a sensor’s frame with $\mathcal{T}_{tag}^{sensor} = (\mathcal{T}_{sensor}^{mocap})^{-1} * \mathcal{T}_{tag}^{mocap}$, we can then project the corners into a sensor’s image as shown by Figure 10. For both sensors, the reprojection error is minimized after applying eye-hand calibration.

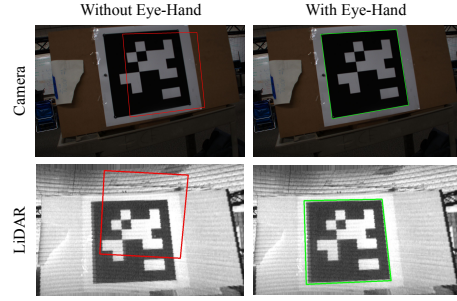


Fig. 10: Reprojection Qualitative Analysis

6 Conclusion

In this paper, we introduce COOPERSCENE, a real-world multi-agent, multi-modal cooperative autonomy dataset with synchronized real-world C-V2X communication characterization. COOPERSCENE provides tightly aligned multi-modal multi-agent data via high precision localization, synchronization, and calibration. The benchmark enables the research community to jointly evaluate perception performance and the practical constraints imposed by real-world vehicular networks. Evaluation results highlight a key insight: the bandwidth demands of state-of-the-art cooperative perception models far exceed the highly variable and limited throughput achievable by current C-V2X. This gap emphasizes the need for communication-efficient, network-aware cooperative perception approaches. COOPERSCENE bridges this gap, provides a unified dataset to advance cooperative autonomy toward scalable, reliable, and deployable real-world systems.

Acknowledgement. This work was conducted at the Collaborative Intelligence Systems Lab (CISL) at UC Riverside. CISL research is supported in part by NSF (ITE-2453817), OUSD (R&E)/RT&L (Center of Excellence W911NF-20-2-0267), and USDOT (Carnation 69A3552348324, NCST 69A3552348319). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the funding agencies. We sincerely thank the college staff led by Victor Hill, as well as the university library staff led by Barbara Neda, for their invaluable support in data maintenance and release. We also express our gratitude to Haoge Zhou, Zhaoze Sun, Chuheng Wei, Ayoub Elazami Elidrissi, Dylon Wong, Joshua Ha, Shubham Derhgawen, Anatulya Nandi, Jonathan Setiabudi, Athena Nelson, Shashwat Jha, Jianpeng Yao, Terek Johnson, Rakshith Mahishi, Divyank Shah, Bhargav HS, Joong Ho Kim, Mason Audet, Shilpa Mukhopadhyay, Xuanpeng Zhao for their effort and participation in data collection and annotation.

References

1. Cohda Wireless MK6. <https://www.cohdawireless.com/solutions/mk6/>, accessed: 2025-11-10
2. Lucid Triton Gig-e Camera. <https://thinklucid.com/product/triton-5-mp-imx490/>, accessed: 2025-11-10
3. OptiTrack Motion Capture System. <https://optitrack.com/>, accessed: 2025-11-10
4. Ouster. <https://ouster.com>, accessed: 2025-11-10
5. Tesla Robotaxi. <https://www.tesla.com/robotaxi>, accessed: 2025-11-10
6. Vicon Motion Capture System. <https://www.vicon.com/>, accessed: 2025-11-10
7. Waymo. <https://waymo.com>, accessed: 2025-11-10
8. XSense MTi-680 RTK GNSS. <https://www.movella.com/sensor-modules/xsens-mti-680-rtk-gnss-ins>, accessed: 2025-11-10
9. Ieee standard for a precision clock synchronization protocol for networked measurement and control systems. IEEE Std 1588-2019 (Revision of IEEE Std 1588-2008) pp. 1–499 (2020). <https://doi.org/10.1109/IEEESTD.2020.9120376>
10. 3GPP: Physical layer procedures. Technical Specification (TS) 36.213 (2021), version 14.17.0
11. 3GPP, E.: Digital cellular telecommunications system (phase 2+) (gsm); universal mobile telecommunications system (umts); lte; 5g; release description; release 14 (3gpp tr 21.914 version 14.0.0) — etsi tr 121 914 v14.0.0. Tech. Rep. TR 121 914 V14.0.0, ETSI (Jun 2018), https://www.etsi.org/deliver/etsi_tr/121900_121999/121914/14.00.00_60/tr_121914v140000p.pdf, published June 2018. Accessed: 2025-11-10
12. Administration, N.H.T.S., et al.: Department of transportation (dot), "federal motor vehicle safety standards; v2v communications", notice of proposed rulemaking (nprm). Tech. rep. (2016)
13. Automation, C.D.: SAE J3216: Taxonomy and definitions for terms related to cooperative driving automation for on-road motor vehicles. SAE International (2020)
14. Besl, P., McKay, N.D.: A method for registration of 3-d shapes. IEEE Transactions on Pattern Analysis and Machine Intelligence **14**(2), 239–256 (1992). <https://doi.org/10.1109/34.121791>

15. Brettle, F., et al.: Google/draco: a library for compressing and decompressing 3d geometric meshes and point clouds. <https://github.com/google/draco> (2018), accessed: 2025-11-10
16. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuscenes: A multimodal dataset for autonomous driving. arXiv:1903.11027 (2019)
17. Chen, Q., Ma, X., Tang, S., Guo, J., Yang, Q., Fu, S.: F-cooper: Feature based cooperative perception for autonomous vehicle edge computing system using 3d point clouds. In: Proceedings of the 4th ACM/IEEE Symposium on Edge Computing. p. 88–100. SEC '19, Association for Computing Machinery, New York, NY, USA (2019)
18. Chen, Q., Tang, S., Yang, Q., Fu, S.: Cooper: Cooperative perception for connected autonomous vehicles based on 3d point clouds. In: 2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS). pp. 514–524 (2019). <https://doi.org/10.1109/ICDCS.2019.00058>
19. Contributors, M.: MMDetection3D: OpenMMLab next-generation platform for general 3D object detection. <https://github.com/open-mmlab/mmdetection3d> (2020), accessed: 2025-11-10
20. Cui, J., Qiu, H., Chen, D., Stone, P., Zhu, Y.: Coopernaut: End-to-end driving with cooperative perception for networked vehicles. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). p. 17231–17241. IEEE (Jun 2022). <https://doi.org/10.1109/cvpr52688.2022.01674>, <http://dx.doi.org/10.1109/CVPR52688.2022.01674>
21. Cui, J., Tang, C., Holtz, J., Nguyen, J., Allievi, A.G., Qiu, H., Stone, P.: Coopreflect: Towards natural language communication for cooperative autonomous driving via multi-agent learning. In: Proceedings of the 25th International Conference on Autonomous Agents and Multiagent Systems. AAMAS '26 (2026), <https://arxiv.org/abs/2505.18334>, oral Presentation
22. Elson, J., Girod, L., Estrin, D.: Fine-Grained network time synchronization using reference broadcasts. In: 5th Symposium on Operating Systems Design and Implementation (OSDI 02). USENIX Association, Boston, MA (Dec 2002)
23. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: The kitti dataset. International Journal of Robotics Research (IJRR) (2013)
24. Hao, T., Zhou, R., Xing, G., Mutka, M.: Wizsync: Exploiting wi-fi infrastructure for clock synchronization in wireless sensor networks. In: 2011 IEEE 32nd Real-Time Systems Symposium. pp. 149–158 (2011)
25. Hu, Y., Fang, S., Lei, Z., Zhong, Y., Chen, S.: Where2comm: Communication-efficient collaborative perception via spatial confidence maps. Advances in neural information processing systems (2022)
26. Hu, Y., Peng, J., Liu, S., Ge, J., Liu, S., Chen, S.: Communication-efficient collaborative perception via information filling with codebook. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15481–15490 (2024)
27. Iyer, G., Ram, R.K., Murthy, J.K., Krishna, K.M.: Calibnet: Geometrically supervised extrinsic calibration using 3d spatial transformer networks. In: 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE (Oct 2018). <https://doi.org/10.1109/iros.2018.8593693>, <http://dx.doi.org/10.1109/IROS.2018.8593693>
28. Lang, A.H., Vora, S., Caesar, H., Zhou, L., Yang, J., Beijbom, O.: Pointpillars: Fast encoders for object detection from point clouds. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12697–12705 (2019)

29. Li, E., Wang, S., Li, C., Li, D., Wu, X., Hao, Q.: Sustech points: A portable 3d point cloud interactive annotation platform system. In: 2020 IEEE Intelligent Vehicles Symposium (IV). pp. 1108–1115 (2020). <https://doi.org/10.1109/IV47402.2020.9304562>
30. Li, J., Xu, R., Liu, X., Ma, J., Chi, Z., Ma, J., Yu, H.: Learning for vehicle-to-vehicle cooperative perception under lossy communication. *IEEE Transactions on Intelligent Vehicles* **8**(4), 2650–2660 (Apr 2023). <https://doi.org/10.1109/tiv.2023.3260040>, <http://dx.doi.org/10.1109/TIV.2023.3260040>
31. Li, Y., Ren, S., Wu, P., Chen, S., Feng, C., Zhang, W.: Learning distilled collaboration graph for multi-agent perception. *Advances in Neural Information Processing Systems* **34**, 29541–29552 (2021)
32. Liu, Z., Tang, H., Amini, A., Yang, X., Mao, H., Rus, D.L., Han, S.: Bevfusion: Multi-task multi-sensor fusion with unified bird’s-eye view representation. In: 2023 IEEE International Conference on Robotics and Automation (ICRA). pp. 2774–2781 (2023). <https://doi.org/10.1109/ICRA48891.2023.10160968>
33. Maróti, M., Kusy, B., Simon, G., Lédeczi, A.: The flooding time synchronization protocol. p. 39–49. *SenSys ’04*, Association for Computing Machinery, New York, NY, USA (2004)
34. Mo, R., Wu, B., Tan, Z., Qiu, H.: See-v2x: C-v2x direct communication dataset: An application-centric approach. In: *Proceedings of the 23rd ACM Conference on Embedded Networked Sensor Systems*. p. 305–311. Association for Computing Machinery, New York, NY, USA (2025), <https://doi.org/10.1145/3715014.3722077>
35. Mukhopadhyay, S., Roy-Chowdhury, A., Qiu, H.: Coopertrim: Adaptive data selection for uncertainty-aware cooperative perception. In: *The Fourteenth International Conference on Learning Representations. ICLR ’26* (2026), <https://openreview.net/forum?id=8NgKNuHRiH>
36. National Academies of Sciences, Engineering, and Medicine and others: Business models to facilitate deployment of connected vehicle infrastructure to support automated vehicle operations. National Academics (2020)
37. Qiu, H., Ahmad, F., Bai, F., Gruteser, M., Govindan, R.: Avr: Augmented vehicular reality. In: *Proceedings of the 16th Annual International Conference on Mobile Systems, Applications, and Services*. p. 81–95. *MobiSys ’18*, New York, NY, USA (2018)
38. Qiu, H., Huang, P., Asavisanu, N., Liu, X., Psounis, K., Govindan, R.: Autocast: Scalable infrastructure-less cooperative perception for distributed collaborative driving. In: *Proceedings of the 20th Annual International Conference on Mobile Systems, Applications, and Services. MobiSys ’22* (December 2022)
39. Rawashdeh, Z.Y., Wang, Z.: Collaborative automated driving: A machine learning-based method to enhance the accuracy of shared information. In: 2018 21st International Conference on Intelligent Transportation Systems (ITSC). pp. 3961–3966 (2018). <https://doi.org/10.1109/ITSC.2018.8569832>
40. Ren, H., Zhang, W., Shi, S., Zhang, X., Zhang, L., Zhang, Y.: Unisense: Spatial-uncertainty-aware collaborative sensing for autonomous driving. In: *Proceedings of the 23rd Annual International Conference on Mobile Systems, Applications, and Services. MobiSys ’25* (2025)
41. Sekaran, K.C., Geisler, M., Rökle, D., Mohan, A., Cremers, D., Utschick, W., Botsch, M., Huber, W., Schön, T.: Urbaning-v2x: A large-scale multi-vehicle, multi-infrastructure dataset across multiple intersections for cooperative perception (2025), <https://arxiv.org/abs/2510.23478>

42. Strobl, K.H., Hirzinger, G.: Optimal hand-eye calibration. In: 2006 IEEE/RSJ International Conference on Intelligent Robots and Systems. pp. 4647–4653 (2006). <https://doi.org/10.1109/IRoS.2006.282250>
43. Sun, P., Kretschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B., Vasudevan, V., Han, W., Ngiam, J., Zhao, H., Timofeev, A., Ettinger, S., Krivokon, M., Gao, A., Joshi, A., Zhang, Y., Shlens, J., Chen, Z., Anguelov, D.: Scalability in perception for autonomous driving: Waymo open dataset. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020)
44. Tsai, R.: A versatile camera calibration technique for high-accuracy 3d machine vision metrology using off-the-shelf tv cameras and lenses. *IEEE Journal on Robotics and Automation* **3**(4), 323–344 (1987). <https://doi.org/10.1109/JRA.1987.1087109>
45. USDOT: Preparing for the future of transportation: Automated vehicles 3.0. <https://www.transportation.gov/av/3> (2018), accessed: 2025-11-10
46. Wang, T.H., Manivasagam, S., Liang, M., Bin, Y., Zeng, W., Tu, J., Urtasun, R.: V2VNet: Vehicle-to-vehicle communication for joint perception and prediction. In: ECCV (2020)
47. Wang, Z., Wang, Y., Wu, Z., Ma, H., Li, Z., Qiu, H., Li, J.: Cmp: Cooperative motion prediction with multi-agent communication. *IEEE Robotics and Automation Letters* (2025)
48. Waymo: Fleet response: Lending a helpful hand to Waymo’s autonomously driven vehicles. <https://waymo.com/blog/2024/05/fleet-response>, accessed: 2025-11-10
49. Weiss, M.: Telecom requirements for time and frequency synchronization. In: National Institute of Standards and Technology (NIST), USA,[Online]: www.gps.gov/cgsic/meetings/2012/weiss1.pdf (2012), accessed: 2025-11-10
50. Weng, X., Wang, J., Held, D., Kitani, K.: 3d multi-object tracking: A baseline and new evaluation metrics. In: 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 10359–10366 (2020). <https://doi.org/10.1109/IROS45743.2020.9341164>
51. Wu, B., Li, J., Mo, R., Yue, J., Bharadia, D., Qiu, H.: Demo Abstract: Cooperative Multi-modal Sensing, p. 712–713. Association for Computing Machinery, New York, NY, USA (2025), <https://doi.org/10.1145/3715014.3724372>
52. Xu, J., Zhang, Y., Cai, Z., Huang, D.: Cosdh: Communication-efficient collaborative perception via supply-demand awareness and intermediate-late hybridization (2025), <https://arxiv.org/abs/2503.03430>
53. Xu, R., Tu, Z., Xiang, H., Shao, W., Zhou, B., Ma, J.: Cobevt: Cooperative bird’s eye view semantic segmentation with sparse transformers. In: Conference on Robot Learning (CoRL) (2022)
54. Xu, R., Xia, X., Li, J., Li, H., Zhang, S., Tu, Z., Meng, Z., Xiang, H., Dong, X., Song, R., Yu, H., Zhou, B., Ma, J.: V2v4real: A real-world large-scale dataset for vehicle-to-vehicle cooperative perception. In: The IEEE/CVF Computer Vision and Pattern Recognition Conference (2023)
55. Xu, R., Xiang, H., Tu, Z., Xia, X., Yang, M.H., Ma, J.: V2x-vit: Vehicle-to-everything cooperative perception with vision transformer. In: Proceedings of the European Conference on Computer Vision (2022)
56. Xu, R., Xiang, H., Xia, X., Han, X., Li, J., Ma, J.: Opv2v: An open benchmark dataset and fusion pipeline for perception with vehicle-to-vehicle communication. In: 2022 International Conference on Robotics and Automation (ICRA). pp. 2583–2589. IEEE (2022)

57. Yan, Y., Mao, Y., Li, B.: Second: Sparsely embedded convolutional detection. *Sensors* **18**(10), 3337 (2018)
58. Yu, H., Luo, Y., Shu, M., Huo, Y., Yang, Z., Shi, Y., Guo, Z., Li, H., Hu, X., Yuan, J., Nie, Z.: Dair-v2x: A large-scale dataset for vehicle-infrastructure cooperative 3d object detection. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 21361–21370 (2022)
59. Yu, H., Yang, W., Ruan, H., Yang, Z., Tang, Y., Gao, X., Hao, X., Shi, Y., Pan, Y., Sun, N., Song, J., Yuan, J., Luo, P., Nie, Z.: 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* (2023)
60. Yu, H., Yang, W., Zhong, J., Yang, Z., Fan, S., Luo, P., Nie, Z.: End-to-end autonomous driving through v2x cooperation. In: *The 39th Annual AAAI Conference on Artificial Intelligence* (2025)
61. Yuan, W., Li, J., Yue, J., Shah, D., Karydis, K., Qiu, H.: Bevcilib: Lidar-camera calibration via geometry-guided bird’s-eye view representations. In: *9th Annual Conference on Robot Learning. CoRL ’25* (2025), <https://arxiv.org/abs/2506.02587>
62. Zhang, J., Yang, K., Wang, Y., Wang, H., Sun, P., Song, L.: Ermvp: Communication-efficient and collaboration-robust multi-vehicle perception in challenging environments. In: *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 12575–12584 (2024). <https://doi.org/10.1109/CVPR52733.2024.01195>
63. Zhang, Q., Pless, R.: Extrinsic calibration of a camera and laser range finder (improves camera calibration). In: *2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (IEEE Cat. No.04CH37566). vol. 3, pp. 2301–2306 vol.3 (2004). <https://doi.org/10.1109/IROS.2004.1389752>
64. Zhang, Q., Zhang, X., Zhu, R., Bai, F., Naserian, M., Mao, Z.M.: Robust real-time multi-vehicle collaboration on asynchronous sensors. In: *Proceedings of the 29th Annual International Conference on Mobile Computing and Networking*. pp. 1–15 (2023)
65. Zhang, X., Zhang, A., Sun, J., Zhu, X., Guo, Y.E., Qian, F., Mao, Z.M.: Emp: edge-assisted multi-vehicle perception. In: *Proceedings of the 27th Annual International Conference on Mobile Computing and Networking*. p. 545–558. *MobiCom ’21*, Association for Computing Machinery, New York, NY, USA (2021). <https://doi.org/10.1145/3447993.3483242>, <https://doi.org/10.1145/3447993.3483242>
66. Zhang, Z.: A flexible new technique for camera calibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **22**(11), 1330–1334 (2000). <https://doi.org/10.1109/34.888718>
67. Zhu, R., Zhu, X., Zhang, A., Zhang, X., Sun, J., Qian, F., Qiu, H., Mao, Z.M., Lee, M.: Boosting collaborative vehicular perception on the edge with vehicle-to-vehicle communication. In: *Proceedings of the 22nd ACM Conference on Embedded Networked Sensor Systems*. p. 141–154. *SenSys ’24* (2024). <https://doi.org/10.1145/3666025.3699328>, <https://doi.org/10.1145/3666025.3699328>
68. Zimmer, W., Wardana, G.A., Sritharan, S., Zhou, X., Song, R., Knoll, A.: Tumtraf v2x cooperative perception dataset. *arXiv preprint arXiv:2403.01316* (2024)
69. Zimmer, W., Wardana, G.A., Sritharan, S., Zhou, X., Song, R., Knoll, A.C.: Tumtraf v2x cooperative perception dataset. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2024)