RESOLVE: A Multi-Resolution and Multi-Modal Dataset for Roadside Cooperative Perception

Shaozu Ding[1], Linan Song[1], Marco De Vincenzi[1,2], and Dajiang Suo[1]*

[1] The Polytechnic School, Arizona State University, USA
{sding32,lsong124,mdevinc2,dsuo2}@asu.edu
[2] Department of Computer Science and Engineering, New York University, USA

Abstract. LiDAR has increasingly been integrated into traffic cameras to expand coverage and mitigate occlusion in roadside cooperative perception. However, how unimodal and camera–LiDAR fusion architectures behave under variations in LiDAR point sparsity induced by sensor configurations and scene-dependent sensing conditions remains underexplored. We introduce RESOLVE, a large-scale real-world benchmark dataset featuring multi-resolution roadside LiDAR and synchronized camera-LiDAR sensing for systematic evaluation of unimodal and fusion-based architectures in roadside 3D detection and tracking. RESOLVE contains over 100k images and 26k point cloud frames with 220k manually annotated bounding boxes, captured at a real-world urban intersection across diverse lighting and weather conditions and spanning 10 classes of traffic participants. In particular, RESOLVE enables controlled evaluation across three LiDAR resolution levels while keeping all other sensing and environmental factors fixed. This allows fair cross-architecture comparisons under point cloud distribution shifts resulting from resolution variations, sensing distance, and training–inference resolution mismatches. Results from extensive benchmark experiments reveal insights into how multimodal fusion can compensate for LiDAR point sparsity, offering clues for designing cost-efficient roadside multimodal perception. The dataset and benchmark codes are available at https://github.com/ASU-Suo-Lab/RESOLVE.

Keywords: Roadside cooperative perception · Multi-resolution LiDAR · Multi-modal dataset · Intelligent transportation systems

1 Introduction

In the context of intelligent transportation systems (ITS) [11] and autonomous driving [24], cooperative perception refers to merging data from distributed sensors (e.g., Camera, LiDAR) to mitigate occlusions and expand sensing coverage. In parallel with vehicle-centric cooperative perception, which relies on vehicle-to-vehicle (V2V) [21] or vehicle-to-infrastructure (V2I) [3, 4] data sharing, another

* Corresponding author.

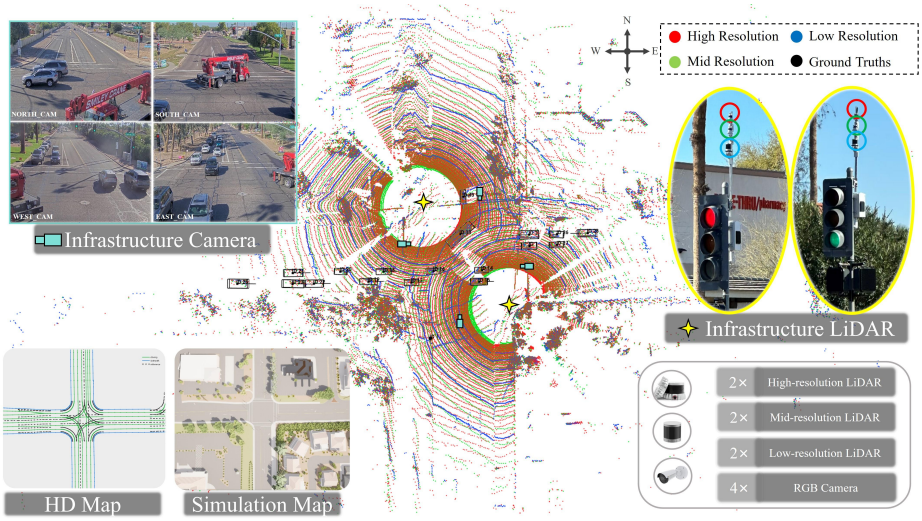


Fig. 1: This illustration provides an comprehensive overview of our RESOLVE dataset. We visualize globally fused point clouds of a representative scene, with point clouds from LiDARs of different resolutions encoded in different colors. The sensor installation locations on the infrastructure side are highlighted. Both the high-fidelity simulation map based on CARLA and the HD map is displayed.

paradigm focuses on sharing data among infrastructure-mounted sensors through infrastructure-to-infrastructure (I2I) [1] communication mechanisms. In recent years, this infrastructure-centric roadside cooperative perception has emerged as an effective approach to improving traffic safety and operational efficiency in complex urban environments, which is the focus of this paper.

However, existing roadside cooperative perception datasets [16, 41] are collected under a fixed LiDAR configuration. In real-world settings, LiDAR is an emerging technology that is incrementally integrated into existing traffic camera-based infrastructure. Within the lifecycle of a single deployment site such as an intersection, LiDAR configurations can vary due to budget constraints, hardware upgrades, procurement cycles, and installation conditions [7, 23, 39]. In addition, LiDAR measurements become sparser as objects move farther from the sensing location due to LiDAR's angular sampling characteristics [10, 18]. These variations lead to shifts in the point density and geometric representation, creating point distribution changes even when the underlying scene remains the same.

As a result, current state-of-the-art LiDAR perception [32, 46] and camera-LiDAR fusion models [33, 47] are primarily evaluated under matched sensing sparsity, since existing datasets do not provide controlled data to systematically investigate perception behavior under different LiDAR sensing sparsity. In general, it is still difficult to analyze how roadside perception performance varies across different LiDAR resolution, sensing geometry and distance, and under potential resolution mismatches between training and inference.

To address this gap, we propose RESOLVE (Fig. 1), the first real-world roadside cooperative perception dataset featuring synchronized multi-resolution LiDAR and camera data. Unlike prior datasets with fixed sensor configurations, RESOLVE allows controlled evaluation by providing multiple LiDAR resolution settings while keeping other sensing and environmental factors consistent.

Our contributions are as follows:

- We introduce RESOLVE, a large-scale multi-resolution roadside cooperative perception dataset supporting infrastructure-based 3D object detection and tracking. It includes over 20,000 synchronized frames of images from traffic cameras covering four intersection directions and point clouds captured by six LiDARs with three resolution levels at two diagonal intersection corners. The dataset also provides 220k expert-annotated 3D bounding boxes across 10 categories of traffic participants, along with simulation assets and HD maps for reproducibility and future extensions.
- We present extensive experiments with benchmarks for evaluating unimodal LiDAR architectures and multimodal LiDAR–camera fusion models under varying LiDAR resolutions, object detecting distances, and resolution mismatches between training and inference, enabling controlled studies of unimodal performance and multimodal perception gain in roadside perception.
- The benchmark results reveal several observations about multimodal cooperative perception under sparse LiDAR sensing, including how multimodal fusion may compensate for sensing sparsity and how performance varies with LiDAR resolution, sensing distance, and resolution mismatches between training and inference. This provides insights for future research in designing more robust and cost-efficient perception models under roadside sensing scenarios where sensing resolution, object heterogeneity, and distance-conditioned sparsity must be jointly considered.

2 Related Work

Multiple ITS-relevant datasets have been made public by the research community. The first category is designed to facilitate the training of ego-vehicle autonomous driving or individual smart infrastructure for traffic monitoring. In contrast, an increasing number of cooperative perception datasets have emerged, aiming to enhance sensing coverage and mitigate occlusion by enabling data sharing among vehicles, between vehicles and infrastructure, and across distributed infrastructure components. Camera, LiDAR, and Radar have been deployed on ego vehicle for multimodal data collection [6, 13, 40, 42, 67]. To improve traffic monitoring performance, infrastructure-mounted sensors have also been used to collect data near intersections [48, 61, 63, 69], campuses [45], and highways [68].

To mitigate occlusions and achieve see-through capabilities for single vehicles, sensing data has also been shared among adjacent vehicles [27, 56] and between vehicles and infrastructure [34, 41, 60, 64, 70]. Similarly, research has focused on merging data from distributed infrastructure sensors [16, 41, 53, 68] to improve sensing coverage and mitigate dead zones at the corners of individual sensors.

Table 1: A comprehensive comparison of real-world autonomous driving, V2X, and roadside datasets with our RESOLVE dataset. V: Vehicle-only, I: Infrastructure-only, V2V: Vehicle-to-Vehicle, V2I: Vehicle-to-Infrastructure, I2I: Infrastructure-to-Infrastructure, Mixed: Corridor & Intersection, LiDAR Res.: LiDAR Resolution.

Dataset	Year	Type	Scene	# of LiDAR Res.	LiDAR Res. Impact	Class	Images	LiDAR Frames	3D Boxes
KITTI [13]	2012	V	Driving	1	–	8	15k	15k	200k
nuScenes [6]	2019	V	Driving	1	–	23	1.4M	400k	1.4M
Waymo Open [42]	2019	V	Driving	2	–	4	1M	200k	12M
OmniHD-Scenes [67]	2024	V	Driving	1	–	4	–	12k	514.6k
TAD-E2E [31]	2025	V	Driving	2	–	–	–	1M	–
V2V4Real [56]	2022	V2V	Driving	1	–	5	40k	20k	240k
DAIR-V2X [64]	2022	V2I	Intersection	2	–	10	39k	39k	464k
V2X-Real [53]	2024	V2I	Intersection	2	–	10	171k	33k	1.2M
TUMTraf-V2X [70]	2024	V2I	Intersection	2	–	8	5.0k	2k	29k
HoloVIC [34]	2024	V2I	Intersection	3	–	3	–	100k	11.47M
CATS-V2V [27]	2025	V2V	Driving	1	–	10	1.26M	60k	–
UrbanIng-V2X [41]	2025	V2I	Intersection	3	–	13	81.6k	27.2k	712k
V2X-Radar [60]	2025	V2I	Intersection	1	–	5	40k	20k	350k
IPS300+ [48]	2022	I	Intersection	1	–	7	14.1k	14.1k	4.5M
Rope3D [61]	2022	I	Intersection	2	–	13	50k	–	1.5M
A9-Intersection [69]	2023	I	Intersection	1	–	10	4.8k	4.8k	57.4k
CORP [45]	2024	I	Campus	2	–	5	205k	102k	215k
H-V2X [30]	2024	I	Highway	0	–	4	1.9M	–	–
RCooper [16]	2024	I2I	Mixed	2	–	10	50k	30k	–
RoScenes [68]	2024	I2I	Highway	0	–	4	1.21M	–	21.39M
RESOLVE (Ours)	2025	I2I	Intersection	3	Yes	10	100k	26k	220k

Among those involving infrastructure-to-infrastructure sensing data sharing, RESOLVE is the first real-world multimodal dataset with multi-resolution LiDAR and controlled deployment setup for both traffic camera and LiDAR sensors, enabling resolution-isolated cross-architecture comparison under point cloud distribution shifts. A summary of these datasets is given in Table 1.

3 RESOLVE Dataset

RESOLVE is a multi-resolution, multi-modal roadside dataset for a complex urban intersection. The data covers high-speed and high-density traffic flow and includes diverse weather and lighting conditions. All annotations undergo expert processes and quality control to ensure consistency and accuracy, while low-latency synchronization across multiple sensors is achieved through a unified time base. Based on LiDAR resolutions, the dataset can be divided into three subsets: high, medium, and low, targeting the highest perception accuracy, the balance of performance and cost, and low-latency real-time applications, respectively.

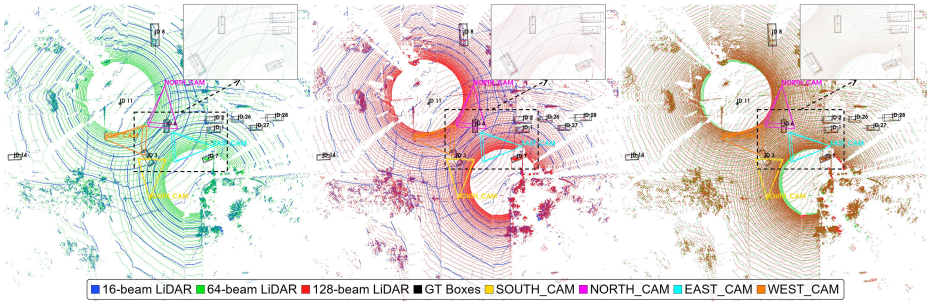
Table 2: Key Sensor Specifications in RESOLVE Dataset.

Sensor	Sensor Model	Details
LiDAR	Ouster OS1-128 (×2)	128 beams, 10 Hz, 360° horizontal FOV, −22.5° to +22.5° vertical FOV, ≤ 90 m range at ≥ 10% reflectivity.
	Ouster OS1-64 (×2)	64 beams, 10 Hz, 360° horizontal FOV, −22.5° to +22.5° vertical FOV, ≤ 90 m range at ≥ 10% reflectivity.
	RoboSense Helios-16 (×2)	16 beams, 10 Hz, 360° horizontal FOV, −15.0° to +15.0° vertical FOV, ≤ 110 m range at ≥ 10% reflectivity.
Camera	AXIS P1455-LE (×4)	RGB, 30 Hz, 1920×1080 resolution, 114° horizontal FOV, 58° vertical FOV.

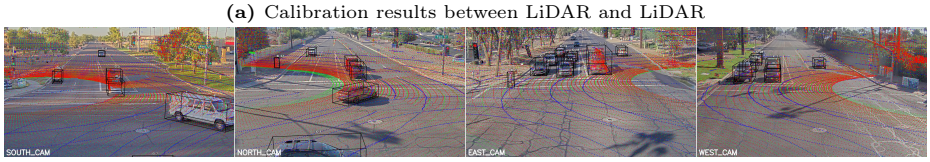
3.1 Sensor Setup

Sensor Specification. Four high-resolution network cameras are installed on traffic light poles facing the east, south, west, and north directions to provide 360° coverage. Six LiDARs are deployed in two symmetric groups, each including 16-, 64-, and 128-beam units. The two groups are mounted on pedestrian signal poles at the northwest and southeast corners. To ensure fairness in subsequent comparative experiments and reduce the impact of pose factors on point cloud density differences, the positions and orientations of each LiDAR group are kept as consistent as possible within feasible limits, except for installation height. Edge computing servers and other modules used for data collection and synchronization are installed in a roadside cabinet next to the pedestrian pole near the southeast pole. Detailed parameters of each sensor are shown in Tab. 2.

Sensor Synchronization. Time synchronization of multimodal datasets is particularly critical. We use the UTC clock as the global time reference and configure the edge computer as the master clock of Precision Time Protocol (PTP) [22]. Each LiDAR is connected to the network through an industrial-level switch and receives synchronization signals as a slave node. Each camera uses Network Time Protocol (NTP) to synchronize with the time server on the edge computer. Based on time synchronization, we further coordinate and control the acquisition trigger of the data stream. The LiDAR uses a phase-locked loop mechanism to align the rotation cycles of each device, ensuring that the scanning cycle starts and ends within the same time window. Since Axis network cameras do not support externally triggered hardware-level synchronization, we perform cross-modal time registration in post-processing, matching each LiDAR frame with the closest camera frame in time, ensuring that the maximum time deviation does not exceed half of the camera frame interval, and discarding matching pairs that exceed the threshold to guarantee stable synchronization quality.



(a) Calibration results between LiDAR and LiDAR



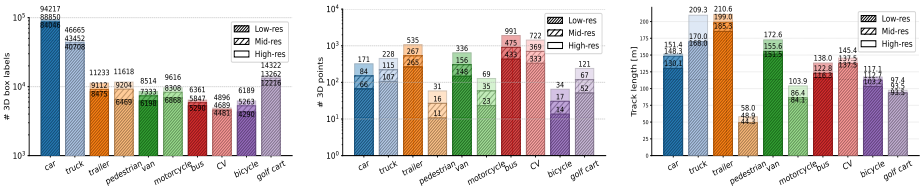
(b) Calibration results between camera and LiDAR

Fig. 2: The illustrations of temporal synchronization and spatial calibration across modalities. (a) shows calibration among LiDARs at different resolutions, including a zoomed-in view of target alignment, with camera frustums overlaid to verify camera–LiDAR extrinsics and coverage. (b) shows point clouds from different resolution LiDARs projected onto multiple camera views to assess cross-modal calibration.

Sensor Calibration. After time synchronization, we perform spatial calibration for all sensors in Fig. 2. We estimate the intrinsics and distortion coefficients of each camera using a checkerboard [66] under a standard pinhole model. For camera-LiDAR extrinsic calibration, we select multiple sets of 2D-3D corresponding points in the point cloud and image, and calculate the transformation matrix minimizing the reprojection error [37]. Furthermore, we select several 3D corresponding points in different LiDAR point clouds, use the least squares criterion to minimize the point-to-point distance to obtain an initial coarse transformation, and then use the Iterative Closest Point (ICP) algorithm [5] for fine registration, thus completing the spatial calibration from LiDAR to LiDAR.

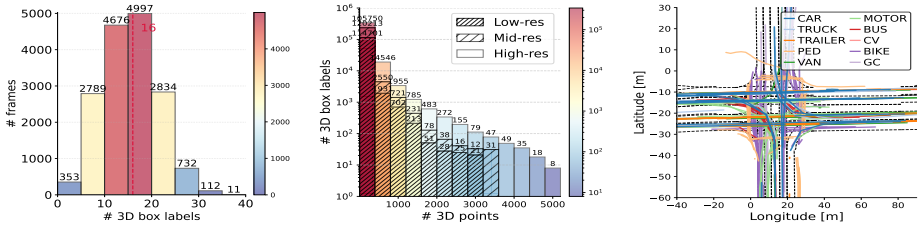
3.2 Data Collection and Annotation

RESOLVE dataset is designed and built for complex urban intersection scenarios, covering a variety of weather and lighting conditions, including sunny, cloudy, rainy, nighttime and strong backlight, to enhance scene diversity and evaluation challenges. In addition to common vehicle types found at urban intersections, the dataset also includes four categories of vulnerable road users, including golf carts, which are relatively rare in public benchmarks. All identifiable facial and license plate information is anonymized using Gaussian blur [12]. The camera and LiDAR are sampled at frequencies of 30Hz and 10Hz respectively and parsed based on ROS2 [36]. We use 10Hz as the annotation frequency to provide accurate 3D bounding boxes for each target, and assign a unique tracking ID to



(a) Valid 3D boxes per class **(b)** Avg. num. of LiDAR points **(c)** Max. track length per class

Fig. 3: Class-wise statistics under three LiDAR resolutions. (a) Lower resolution reduces valid 3D boxes, especially for small targets. (b) As resolution increases, the average number of points increases, reflecting richer geometric observations. (c) Higher resolution LiDAR provides a slight gain for small targets track, but can extend the effective tracking length of large vehicles.



(a) Valid 3D boxes per frame **(b)** LiDAR points per 3D box **(c)** BEV Trajectories per class

Fig. 4: Statistics on valid boxes per frame and in-box points per box, along with visualization of motion trajectories. (a) On average, each frame contains 16 valid bounding boxes. Resolution variations have minimal impact on this metric. (b) Distribution of LiDAR points per valid 3D box under different resolutions. (c) BEV trajectory visualization on the intersection HD map, with trajectories color-coded by class.

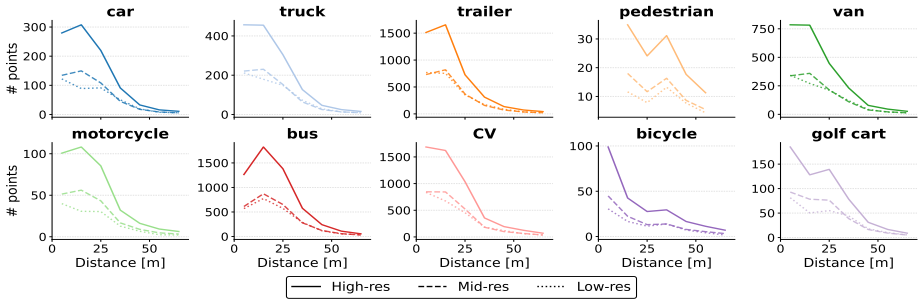


Fig. 5: Statistics on the variation of average number of points with distance from object to the center of intersection at different resolutions. The resolution-caused difference is significant at close range but diminishes with increasing distance.

the target in each sequence to support temporal tasks. We also provide a format conversion tool that can convert this dataset into nuScenes [6] format, making it

easy to be directly compatible and reused with mainstream detection framework OpenPCDet [43] and MMDetection3D [9].

3.3 Dataset Analysis

In roadside sensing, low-resolution point clouds are often too sparse for small, distant, or occluded targets, weakening detection and annotation utility. We therefore deem a 3D box valid only if it contains at least five LiDAR points. However, this definition does not alter the ground truth set used for training or evaluation. Since we employ a consistent point cloud range and identical high-resolution annotations across three resolutions, 3D boxes containing fewer than five points are retained for training and evaluation even in low- or medium-resolution settings. Figure 3a shows that the number of valid 3D boxes decreases as the resolution decreases, indicating that the lower point density directly reduces the amount of usable supervision and the coverage of the evaluation targets. Figure 3b shows that 128-beam LiDAR yields much denser in-box points than 64-beam, while the gain from 16 to 64 is smaller, indicating non-linear density scaling in this intersection setup. The tracking results in Fig. 3c further show a limited track length for small targets due to the decay of long-range point density, with marginal benefit from higher resolution. In contrast, for large vehicles, higher resolution provides richer geometry and typically extends the track length by about 20 m.

As shown in Fig. 4a, a higher LiDAR resolution yields more valid boxes per frame, improving annotation usability and coverage. Figure 4b further shows that in-box point counts shift upwards at high resolution with a heavier upper tail, confirming that higher resolution can provide richer geometric information for some targets. bird's-eye-view (BEV) visualization in Fig. 4c presents the spatial layout of traffic flow, the spatial distribution of categories, and typical motion behaviors in the scene. As shown in Fig. 5, the overall trend of the number of in-box points changing with the distance from the target to the intersection center is consistent across three resolutions. Different classes show slight differences due to variations in spatial distribution and motion patterns.

4 Tasks

4.1 3D Object Detection

We group representative 3D object detection architectures by backbone operator, covering unimodal LiDAR detectors based on sparse convolution [59], transformer [44], and mamba [14], as well as multimodal detectors that fuse LiDAR with the camera branch. We benchmark SOTA methods on our RESOLVE dataset using the nuScenes protocol [6] over 10 classes, reporting mean Average Precision (mAP) and nuScenes Detection Score (NDS). Since attributes are unavailable, we exclude mAAE and recompute NDS accordingly as in Eq. (1).

$$\mathrm{NDS}_{\mathrm{ours}} = \frac{1}{9}\left[5\,\mathrm{mAP} + \sum_{m\in\{\mathrm{mATE,mASE,mAOE,mAVE}\}} \max\bigl(1-m,\,0\bigr)\right] \tag{1}$$

4.2 3D Multi-object Tracking

Multi-object tracking is commonly categorized into tracking-by-detection (TBD) and joint detection and tracking (JDT). TBD detects objects per frame and links them across time via data association (e.g., the Hungarian algorithm [25]) with track management and is therefore largely limited by detector quality. JDT jointly learns detection and tracking within a single network, offering tighter temporal modeling but higher complexity. In this work, we focus on TBD and evaluate several classic TBD trackers under different LiDAR resolutions using the nuScenes tracking protocol, reporting Average Multiple Object Tracking Accuracy (AMOTA) and Average Multiple Object Tracking Precision (AMOTP).

4.3 Cooperative Perception

Cooperative perception improves 3D detection by enabling multiple agents to share complementary observations. In this work, we treat two diagonally deployed infrastructure sensor suites as distinct agents and categorize cooperative perception methods by their fusion stage, including no fusion, early fusion, late fusion, and intermediate fusion [8, 55, 57, 58], and report mAP over 10 categories following the nuScenes protocol [6].

5 Experiments

5.1 Implementation Details

3D Object Detection. We split the dataset into training, validation, and test sets in a 7:2:1 ratio. The point cloud is restricted to the range of $x, y \in [-60m, 88.8m]$ and $z \in [-8m, 0m]$, with a voxel size of $[0.3m, 0.3m]$. The image size of the fusion models is set to 256×704. All models are trained for 20 epochs with a batch size of 4 per GPU, using the adam optimizer and a onecycle scheduler with a learning rate of 0.001. The same training protocol is applied to three LiDAR resolution settings to ensure fair comparisons. The experiments are conducted on four NVIDIA RTX PRO 6000 BLACKWELL GPUs and evaluated on a single GPU. All experiments are implemented in OpenPCDet [43]. We have integrated and open-sourced all model implementations in a unified code repository and provided training weights to facilitate further research.

3D Multi-object Tracking. Since different TBD methods are tuned for specific benchmarks and even have special settings for different classes, we do not perform exhaustive optimization for each method. Therefore, the performance of some trackers may be lower than the results reported in their original benchmark tests. However, we adopt a uniform evaluation protocol to ensure fair comparisons. Specifically, we fix the distance threshold at 0.4, the threshold scan count at 40, the minimum recall rate at 0.1, and the ID switching time window as unrestricted. We have integrated and open-sourced all tracking models into a unified detection framework so that it can be used seamlessly with various detectors.

Table 3: Evaluation results of 3D detection methods on our RESOLVE dataset under three LiDAR resolutions. L: LiDAR, C: Camera, SparseConv: Sparse Convolution.

Model	Modality	LiDAR Backbone	mAP↑			NDS↑		
			Low	Mid	High	Low	Mid	High
PointPillars [26]	L	SparseConv	75.1	79.7	80.6	71.2	72.6	73.6
SECOND [59]	L	SparseConv	68.2	76.6	78.0	66.5	71.2	73.1
CenterPoint [62]	L	SparseConv	79.9	86.4	87.4	73.9	79.9	80.9
TransFusion-L [2]	L	SparseConv	82.5	86.6	89.1	76.3	80.1	82.4
VoxSeT [17]	L	Transformer	87.4	89.1	90.1	82.9	81.6	83.1
DSVT [46]	L	Transformer	85.9	94.1	94.6	81.5	87.0	87.5
Voxel Mamba [65]	L	Mamba	85.7	94.9	95.4	81.7	88.5	89.1
LION [32]	L	Mamba	88.1	95.1	95.9	84.2	89.8	91.1
BEVFusion [33]	L+C	SparseConv	86.3	92.6	93.1	79.8	84.8	86.4
UniTR [47]	L+C	Transformer	89.7	94.4	94.9	83.7	87.3	87.4

Cooperative Perception. All experiments are implemented in OpenCOOD [58]. To ensure a fair comparison across different cooperative perception paradigms, we employ a unified PointPillars-based backbone, where all collaborative detectors share the same anchor box definitions, focal classification loss, Smooth-L1 regression loss, and NMS threshold as the single-agent PointPillars baseline.

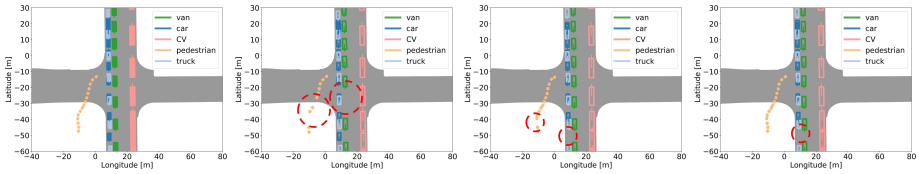
5.2 Benchmark Results

3D Object Detection. As shown in Tab. 3, increasing LiDAR resolution from low to medium boosts the mean mAP across all models by about 7.4%. The gains for SparseConv, Transformer, Mamba, and multimodal methods are about 8.0%, 5.7%, 9.3%, and 6.3%, respectively. Sparse convolution benefits more from higher intra-voxel point density [19]. Transformer-based models are relatively robust to sparsity due to global interactions, while Mamba models, due to serialization modeling, have higher requirements for token quality. Therefore, when medium resolution brings more continuous spatial structure and more reliable local statistics, their performance improvement is more significant. Multimodal methods benefit from improved cross-modal alignment and complementarity as LiDAR geometry becomes more reliable. When resolution increases further from medium to high, the overall mean mAP increases only by about 1.0%. Mid-resolution already captures most discriminative geometry, and additional densification yields limited new information. Moreover, encoding stages such as voxelization and sampling can introduce information bottlenecks, preventing extra points from translating into proportional gains.

3D Multi-object Tracking. According to Tab. 4, a higher resolution generally improves detection quality, thus benefiting downstream association. This is

Table 4: Evaluation results of multi-object tracking methods on our RESOLVE dataset under three LiDAR resolutions. We use BEVFusion [33] as detector. The best performance is marked in red. CV, Tra, Motor, Bic, Ped, GC: Construction Vehicle, Trailer, Motorcycle, Bicycle, Pedestrian, Golf Cart. Res.: Resolution.

Model	Res.	AMOTA ↑										AMOTP ↓									
		Car	Truck	CV	Bus	Tra	Van	Motor	Bic	Ped	GC	Car	Truck	CV	Bus	Tra	Van	Motor	Bic	Ped	GC
AB3DMOT [52]	Low	47.4	48.7	62.6	56.2	40.5	52.4	7.6	37.9	54.6	47.6	44.5	51.5	65.3	56.0	113.2	64.5	164.3	120.7	67.3	29.7
	Mid	39.2	42.4	64.2	41.1	44.9	35.8	4.8	44.1	55.3	35.4	34.0	45.0	44.5	49.0	106.7	56.4	155.1	105.3	72.7	29.7
	High	35.0	39.4	69.1	57.1	46.8	33.0	8.0	45.4	58.7	37.7	30.9	36.5	31.3	34.4	100.6	42.3	140.4	97.1	70.6	18.7
CenterPoint [62]	Low	88.9	82.8	87.9	92.8	74.6	69.1	76.9	93.0	78.1	92.9	32.9	41.3	50.5	49.1	60.6	29.6	69.1	32.2	51.7	26.9
	Mid	85.1	83.8	93.1	93.5	83.3	70.5	82.1	93.3	71.9	94.8	31.0	37.6	39.1	36.4	44.8	33.0	46.2	23.3	70.7	22.1
	High	91.9	87.1	94.0	95.8	90.0	66.8	92.0	99.3	92.7	96.7	22.0	32.6	25.2	25.7	32.8	28.9	33.9	15.0	27.6	15.3
SimpleTrack [38]	Low	88.5	85.6	92.6	94.1	84.2	71.9	52.8	95.3	73.4	97.8	34.8	42.2	42.4	48.2	55.3	34.9	70.3	29.1	58.1	20.8
	Mid	87.2	85.2	97.0	97.1	89.5	74.6	53.1	98.4	77.8	96.0	34.2	40.4	31.2	35.7	39.3	38.9	50.5	18.0	68.5	21.8
	High	90.0	88.4	97.1	97.0	89.8	74.9	56.7	99.7	82.9	96.8	27.0	31.8	20.6	27.6	31.7	30.5	39.2	13.4	44.5	16.7
Poly-MOT [29]	Low	79.6	72.6	65.9	77.6	70.0	74.2	28.2	70.8	90.8	80.4	33.1	43.6	72.9	68.2	48.4	48.3	81.1	36.9	31.9	23.3
	Mid	69.6	63.5	61.3	64.6	68.8	65.3	39.3	81.2	74.7	75.7	45.2	53.8	89.1	79.3	49.1	62.3	66.6	36.7	77.8	26.5
	High	69.6	65.0	67.2	80.1	70.1	72.2	44.6	84.8	74.7	79.0	38.4	41.6	74.1	63.4	38.6	45.6	47.4	31.0	72.0	15.8
MCTrack [49]	Low	70.3	63.1	60.9	62.8	60.5	62.8	26.0	45.9	72.0	78.7	41.9	63.3	93.6	91.2	74.6	99.9	147.7	115.9	70.3	51.7
	Mid	66.9	63.9	74.9	61.5	64.2	41.3	49.3	59.0	73.1	86.8	35.7	47.2	63.4	80.8	58.9	91.4	85.7	84.4	70.4	33.6
	High	68.1	64.4	85.0	85.9	72.2	42.9	56.8	79.7	82.7	89.9	26.9	32.6	32.7	30.5	39.0	68.5	49.5	39.6	41.7	16.7



(a) Ground Truth **(b)** Low-res tracking **(c)** Mid-res tracking **(d)** High-res tracking

Fig. 6: Qualitative multi-object tracking results of SimpleTrack [38] on our RESOLVE across three resolution settings. Ground truths and predictions with a score threshold of 0.3 are overlaid and visualized at 15 Hz. Failure cases are marked in red circles.

achieved by improving positioning accuracy. The average AMOTP decreases by 5.1% from low to medium resolution and by 14.1% from medium to high resolution. In contrast, AMOTA is not necessarily monotonically positively correlated with resolution due to factors such as association strategies and category characteristics. CenterPoint [62] provides more stable and reliable association and localization for small objects, such as fast-moving motorcycles and slow-moving pedestrians. SimpleTrack [38] achieves the highest AMOTA across most classes, demonstrating stronger matching robustness. As shown in Fig. 6, we visualize multi-frame tracking results. To avoid clutter, we display five representative trajectories of different categories in a single scene. It can be observed that higher resolution improves localization accuracy, especially for small targets.

Cooperative Perception. As shown in Fig. 7, cooperative fusion consistently improves over the no-fusion baseline, but different fusion paradigms show different sensitivity to resolution. Early fusion performs well at higher resolutions due to raw-level aggregation, but becomes less robust with sparse LiDAR inputs. Late fusion brings limited gains, as decision-level aggregation cannot recover

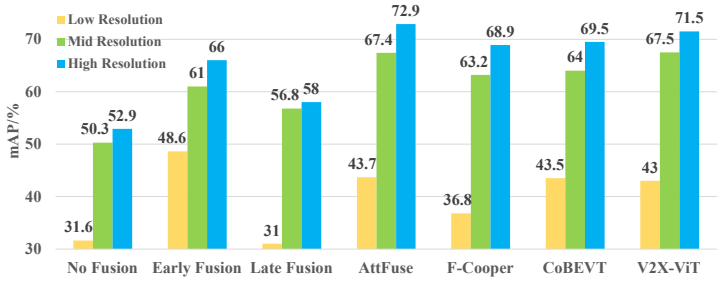


Fig. 7: Agent-level cooperative perception performance under three LiDAR resolutions.

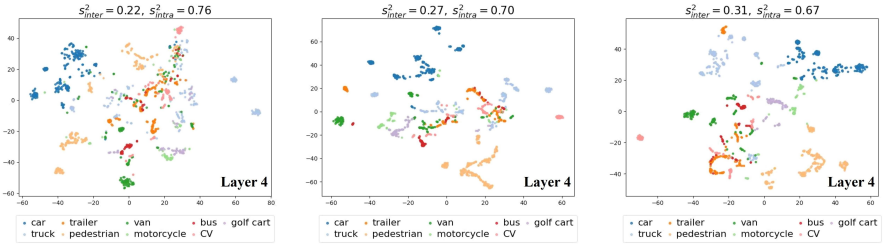
geometric details. By exchanging learned BEV features, intermediate fusion methods (e.g., AttFuse [58], F-Cooper [8], CoBEVT [55], and V2X-ViT [57]), which account for agent heterogeneity and inter-agent synchronization, outperform late and early fusion in most settings. The advantage of feature-level fusion is most evident at medium resolution, where feature-level collaboration compensates for reduced point density using complementary viewpoints. However, the clear drop at low resolution suggests that cooperation alone cannot fully recover severely missing geometry, highlighting the need to explicitly account for resolution-induced point cloud sparsity.

5.3 Analysis

Multimodal Perception Gain under point sparsity. Higher-resolution LiDAR improves unimodal detection performance primarily through enhanced feature discriminability. To examine this effect, we visualize the features of the last backbone layer using t-SNE [35] and quantify class separability using inter-class and intra-class variance [54], as shown in Fig. 8. The results show that higher resolution leads to larger inter-class variance and smaller intra-class variance, indicating improved separability across classes and stronger clustering within the same class. We further observe that multimodal fusion models using lower-resolution data match or even outperform unimodal models using higher-resolution data, as shown in Tab. 5, suggesting that camera features may partially compensate for LiDAR sensing sparsity.

Table 5: Comparison of unimodal and multi-modal models with same LiDAR modules. Results in the same color form a paired comparison.

Model	Modality	LiDAR Backbone	mAP↑		
			Low	Mid	High
TransFusion-L [2]	L	SparseConv	82.5	86.6	89.1
BEVFusion [33]	L+C	SparseConv	86.3(-0.3)	92.6(+3.5)	93.1



(a) Low-res, mAP=82.5 **(b)** Mid-res, mAP=86.6 **(c)** High-res, mAP=89.1

Fig. 8: t-SNE plots of the impact of LiDAR resolutions on feature representation capability of unimodal models under Transfusion-L [2] architecture.

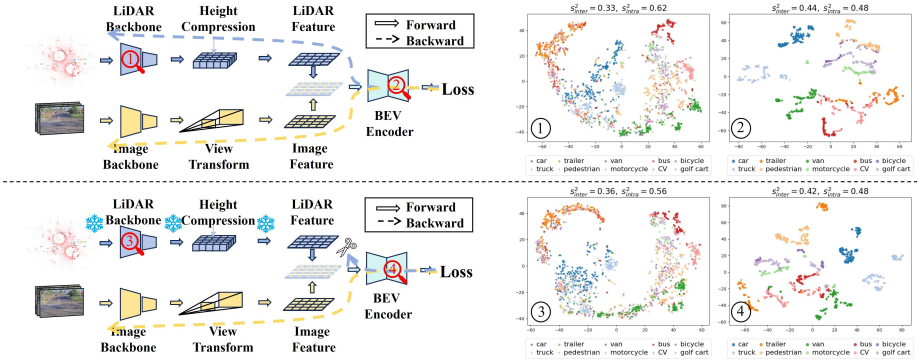
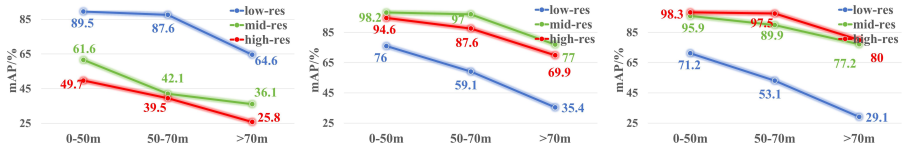


Fig. 9: Controlled experimental design under a general multimodal architecture. Top: standard backpropagation, mAP=93.1. Bottom: stop gradients in the LiDAR branch and freeze LiDAR-module parameters, mAP=89.4. ① and ③ visualize the features of the first layer of LiDAR backbone, while ② and ④ visualize the fused features.

To further investigate how multimodal learning influences LiDAR feature representations, we follow the protocol in [50] and conduct two training settings in the BEVFusion [33] architecture, as shown on the left of Fig. 9: (i) allow standard training with gradient backpropagation to both LiDAR and camera branches, and (ii) truncate the gradient from the multimodal loss propagated to the LiDAR backbone, and substitute it the unimodal LiDAR loss. We analyze feature distributions from early LiDAR backbone layers and late fusion layers using t-SNE visualizations and variance metrics (right of Fig. 9), allowing both qualitative and quantitative investigation of how camera information influences LiDAR feature encoding.

Compared to the truncated-gradient setting, standard multimodal training (marked number 1 in Fig. 9) leads to reduced class separability in early LiDAR backbone layers, indicating that multimodal loss alters feature learning both in late fusion stages and in the LiDAR branch during early representation stages. Despite this reduced early-layer discriminability due to gradient from multi-



(a) Under low-res training **(b)** Under mid-res training **(c)** Under high-res training

Fig. 10: Unimodal performance degradation caused by different LiDAR resolutions used during training and inference. Voxel Mamba [65] is used as unimodal detector.

modal loss [50], the fusion model achieves higher detection performance and stronger feature separability at the fusion layers (marked number 2 in Fig. 9). This implies that multimodal learning encourages the LiDAR backbone to rely less on point-density cues and instead align more effectively between camera and LiDAR features [15]. Overall, these results suggest that multimodal fusion can mitigate LiDAR sensing sparsity by reshaping LiDAR feature learning, enabling competitive perception performance even with lower-resolution sensors and offering clues for designing more cost-effective multimodal perception systems in roadside cooperative perception scenarios.

Impact of LiDAR Resolution Mismatch. As shown in Fig. 10, benchmark results also imply that LiDAR resolution shifts between training and inference introduce a clear domain gap that degrades unimodal detection performance. While previous work [20, 28, 51] on domain adaptation uses downsampling or cross-dataset settings to investigate this issue, RESOLVE provides a controlled protocol that isolates resolution shifts. The model trained at low resolution drops by 40% when transferred to higher resolution, while a high-resolution model drops only 4.6% on mid resolution but still drops 40% on low resolution with severe long-range degradation. This gap is even more detrimental for multimodal detection (see supplementary materials), indicating that resolution mismatch disrupts both unimodal feature learning and cross-modal alignment.

6 Conclusion

We propose RESOLVE, the first real-world multi-resolution, multi-modal dataset for roadside cooperative perception. It benchmarks how resolution-induced point-cloud distribution shifts affect model performance on downstream roadside perception tasks, while isolating other sensing factors. We also conduct comprehensive experiments on unimodal and multimodal models. The results show how multimodal fusion can compensate for degraded point-cloud branch feature learning under reduced LiDAR resolution, providing insights into the design of more robust and cost-effective roadside perception models that explicitly account for resolution mismatch, target distance, and object diversity. One limitation of RESOLVE is that it is restricted to single-intersection scenarios. We will explore multi-intersection and corridor-level scenarios in future work.

Acknowledgements

The author would like to thank the Maricopa County Department of Transportation, Arizona, USA, for its technical support. This work was partially supported by an NVIDIA Academic Grant Program Award.

References

1. Ahmad, F., Shin, C.S., Pang, W., Leong, B., Ghosh, P., Govindan, R.: Cooperative infrastructure perception. In: Ninth IEEE/ACM International Conference on Internet-of-Things Design and Implementation, IoTDI 2024, Hong Kong, May 13-16, 2024. pp. 61–72 (2024). https://doi.org/10.1109/IOTDI61053.2024.00010
2. Bai, X., Hu, Z., Zhu, X., Huang, Q., Chen, Y., Fu, H., Tai, C.: Transfusion: Robust lidar-camera fusion for 3d object detection with transformers. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022. pp. 1080–1089 (2022). https://doi.org/10.1109/CVPR52688.2022.00116
3. Bai, Z., Wu, G., Barth, M.J., Liu, Y., Sisbot, E.A., Oguchi, K.: Pillargrid: Deep learning-based cooperative perception for 3d object detection from onboard-roadside lidar. In: 2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC). p. 1743–1749. IEEE (Oct 2022). https://doi.org/10.1109/itsc55140.2022.9921947
4. Bai, Z., Wu, G., Barth, M.J., Liu, Y., Sisbot, E.A., Oguchi, K.: Vinet: Lightweight, scalable, and heterogeneous cooperative perception for 3d object detection. Mechanical Systems and Signal Processing **204**, 110723 (2023). https://doi.org/10.1016/j.ymssp.2023.110723
5. Besl, P.J., McKay, N.D.: Method for registration of 3-D shapes. In: Schenker, P.S. (ed.) Sensor Fusion IV: Control Paradigms and Data Structures. vol. 1611, pp. 586 – 606. International Society for Optics and Photonics, SPIE (1992). https://doi.org/10.1117/12.57955
6. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuscenes: A multimodal dataset for autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020)
7. Cai, X., Jiang, W., Xu, R., Zhao, W., Ma, J., Liu, S., Li, Y.: Analyzing infrastructure lidar placement with realistic lidar simulation library. In: 2023 IEEE International Conference on Robotics and Automation (ICRA). pp. 5581–5587 (2023). https://doi.org/10.1109/ICRA48891.2023.10161027
8. 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), https://doi.org/10.1145/3318216.3363300
9. Contributors, M.: MMDetection3D: OpenMMLab next-generation platform for general 3D object detection. https://github.com/open-mmlab/mmdetection3d (2020), accessed: 25 Jun 2026
10. Corral-Soto, E.R., Grandhi, A., He, Y.Y., Rochan, M., Liu, B.: Improving lidar 3d object detection via range-based point cloud density optimization (2023), https://arxiv.org/abs/2306.05663, accessed: 25 Jun 2026

11. Creß, C., Bing, Z., Knoll, A.C.: Intelligent transportation systems using roadside infrastructure: A literature survey. IEEE Transactions on Intelligent Transportation Systems **25**(7), 6309–6327 (2024). https://doi.org/10.1109/TITS.2023.3343434
12. European Parliament and Council of the European Union: Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 (general data protection regulation). Official Journal of the European Union, L 119, 4 May 2016, pp. 1–88 (2016), https://eur-lex.europa.eu/eli/reg/2016/679/oj/eng, cELEX: 32016R0679; accessed 25 Jun 2026
13. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: The KITTI dataset. Int. J. Robotics Res. **32**(11), 1231–1237 (2013). https://doi.org/10.1177/0278364913491297
14. Gu, A., Dao, T.: Mamba: Linear-time sequence modeling with selective state spaces (2024), https://arxiv.org/abs/2312.00752, accessed: 25 Jun 2026
15. Hadgi, S., Li, L., Ovsjanikov, M.: To supervise or not to supervise: Understanding and addressing the key challenges of point cloud transfer learning. In: Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LXXVIII. pp. 146–163 (2024). https://doi.org/10.1007/978-3-031-73229-4_9
16. Hao, R., Fan, S., Dai, Y., Zhang, Z., Li, C., Wang, Y., Yu, H., Yang, W., Yuan, J., Nie, Z.: Rcooper: A real-world large-scale dataset for roadside cooperative perception. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024. pp. 22347–22357 (2024). https://doi.org/10.1109/CVPR52733.2024.02109
17. He, C., Li, R., Li, S., Zhang, L.: Voxel set transformer: A set-to-set approach to 3d object detection from point clouds. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022. pp. 8407–8417 (2022). https://doi.org/10.1109/CVPR52688.2022.00823
18. He, Y., Cao, P., Suo, D., Liu, X.: A joint optimization of beam distribution and deployment for roadside lidar systems to maximize vehicle perception. IEEE Transactions on Intelligent Vehicles **10**(2), 1300–1314 (2025). https://doi.org/10.1109/TIV.2024.3426524
19. Hu, J.S.K., Kuai, T., Waslander, S.L.: Point density-aware voxels for lidar 3d object detection. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022. pp. 8459–8468 (2022). https://doi.org/10.1109/CVPR52688.2022.00828
20. Hu, Q., Liu, D., Hu, W.: Density-insensitive unsupervised domain adaption on 3d object detection. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023. pp. 17556–17566 (2023). https://doi.org/10.1109/CVPR52729.2023.01684
21. Hu, Y., Lu, Y., Xu, R., Xie, W., Chen, S., Wang, Y.: Collaboration helps camera overtake lidar in 3d detection. In: The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
22. IEEE 1588: Ieee standard for a precision clock synchronization protocol for networked measurement and control systems (2020). https://doi.org/10.1109/IEEESTD.2020.9120376
23. Jiang, W., Xiang, H., Cai, X., Xu, R., Ma, J., Li, Y., Lee, G.H., Liu, S.: Optimizing the placement of roadside lidars for autonomous driving. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 18335–18344 (2023). https://doi.org/10.1109/ICCV51070.2023.01685

24. Krammer, A., Schöller, C., Gulati, D., Lakshminarasimhan, V., Kurz, F., Rosenbaum, D., Lenz, C., Knoll, A.: Providentia — a large-scale sensor system for the assistance of autonomous vehicles and its evaluation. Field Robotics **2**, 1156–1176 (2022). https://doi.org/10.55417/fr.2022038
25. Kuhn, H.W.: The hungarian method for the assignment problem. In: 50 Years of Integer Programming 1958-2008 - From the Early Years to the State-of-the-Art, pp. 29–47. Springer (2010). https://doi.org/10.1007/978-3-540-68279-0_2
26. Lang, A.H., Vora, S., Caesar, H., Zhou, L., Yang, J., Beijbom, O.: Pointpillars: Fast encoders for object detection from point clouds. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019. pp. 12697–12705 (2019). https://doi.org/10.1109/CVPR.2019.01298
27. Li, H., Cao, B., Liang, Z., Li, W., Oh, J., Chen, Y., Liang, S., Zhou, H., Ma, C., Liu, J., Li, Z., Zhang, P., Long, K., Liu, M., Jiang, J., Yu, C., Liu, S., Yu, H., Li, X.: Cats-v2v: A real-world vehicle-to-vehicle cooperative perception dataset with complex adverse traffic scenarios (2025), https://arxiv.org/abs/2511.11168, accessed: 25 Jun 2026
28. Li, S., Ma, L., Li, X.: Domain generalization of 3d object detection by density-resampling. In: Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LXIV. pp. 456–473 (2024). https://doi.org/10.1007/978-3-031-73039-9_26
29. Li, X., Xie, T., Liu, D., Gao, J., Dai, K., Jiang, Z., Zhao, L., Wang, K.: Polymot: A polyhedral framework for 3d multi-object tracking. In: 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 9391–9398 (2023). https://doi.org/10.1109/IROS55552.2023.10341778
30. Liu, C., Zhu, M., Ma, C.: H-v2x: A large scale highway dataset for bev perception. In: Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part I. Lecture Notes in Computer Science, vol. 15059, pp. 139–157. Springer (2024). https://doi.org/10.1007/978-3-031-73232-4_8
31. Liu, C., Zhu, M., Zhang, Z., Song, L., Zhao, X., Luo, Q., Wang, Q., Guo, C., Su, K.: Tad-e2e: A large-scale end-to-end autonomous driving dataset. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 26600–26609 (October 2025)
32. Liu, Z., Hou, J., Wang, X., Ye, X., Wang, J., Zhao, H., Bai, X.: LION: linear group RNN for 3d object detection in point clouds. In: Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024 (2024)
33. 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: IEEE International Conference on Robotics and Automation, ICRA 2023, London, UK, May 29 - June 2, 2023. pp. 2774–2781 (2023). https://doi.org/10.1109/ICRA48891.2023.10160968
34. Ma, C., Qiao, L., Zhu, C., Liu, K., Kong, Z., Li, Q., Zhou, X., Kan, Y., Wu, W.: Holovic: Large-scale dataset and benchmark for multi-sensor holographic intersection and vehicle-infrastructure cooperative. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024. pp. 22129–22138 (2024). https://doi.org/10.1109/CVPR52733.2024.02089
35. van der Maaten, L., Hinton, G., Rachmad, Y.: Visualizing data using t-sne. Journal of Machine Learning Research **9**, 2579–2605 (11 2008)

36. Macenski, S., Foote, T., Gerkey, B., Lalancette, C., Woodall, W.: Robot operating system 2: Design, architecture, and uses in the wild. Science Robotics **7**(66), eabm6074 (2022). https://doi.org/10.1126/scirobotics.abm6074
37. Madsen, K., Nielsen, H.B., Tingleff, O.: Methods for non-linear least squares problems (2nd ed.). society for industrial & applied mathematics (2004)
38. Pang, Z., Li, Z., Wang, N.: Simpletrack: Understanding and rethinking 3d multi-object tracking. In: Computer Vision - ECCV 2022 Workshops - Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part I. pp. 680–696 (2022). https://doi.org/10.1007/978-3-031-25056-9_43
39. Qu, A., Huang, X., Suo, D.: Seip: Simulation-based design and evaluation of infrastructure-based collective perception. In: 2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC). pp. 3871–3878 (2023). https://doi.org/10.1109/ITSC57777.2023.10422006
40. Richter, J., Faion, F., Feng, D., Becker, P.B., Sielecki, P., Glaeser, C.: Understanding the domain gap in lidar object detection networks (2022), https://arxiv.org/abs/2204.10024, accessed: 25 Jun 2026
41. Sekaran, K.C., Geisler, M., Rößle, 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. In: Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diago, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025 (2025)
42. Sun, P., Kretzschmar, 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: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2443–2451 (2020). https://doi.org/10.1109/CVPR42600.2020.00252
43. Team, O.D.: Openpcdet: An open-source toolbox for 3d object detection from point clouds. https://github.com/open-mmlab/OpenPCDet (2020), accessed: 25 Jun 2026
44. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. p. 6000–6010. NIPS'17, Curran Associates Inc., Red Hook, NY, USA (2017)
45. Wang, B., Meng, S., Zhang, L., Wang, C., Huang, J., Li, Y., Ren, H., Xiao, Y., Peng, Y., Ji, J., Zhang, Y., Zhang, Y.: Corp: A multi-modal dataset for campus-oriented roadside perception tasks (2024), https://arxiv.org/abs/2404.03191, accessed: 25 Jun 2026
46. Wang, H., Shi, C., Shi, S., Lei, M., Wang, S., He, D., Schiele, B., Wang, L.: DSVT: dynamic sparse voxel transformer with rotated sets. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023. pp. 13520–13529 (2023). https://doi.org/10.1109/CVPR52729.2023.01299
47. Wang, H., Tang, H., Shi, S., Li, A., Li, Z., Schiele, B., Wang, L.: Unitr: A unified and efficient multi-modal transformer for bird's-eye-view representation. In: IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023. pp. 6769–6779 (2023). https://doi.org/10.1109/ICCV51070.2023.00625

48. Wang, H., Zhang, X., Li, Z., Li, J., Wang, K., Lei, Z., Haibing, R.: Ips300+: a challenging multi-modal data sets for intersection perception system. In: 2022 International Conference on Robotics and Automation (ICRA). p. 2539–2545 (2022). https://doi.org/10.1109/ICRA46639.2022.9811699
49. Wang, X., Qi, S., Zhao, J., Zhou, H., Zhang, S., Wang, G., Tu, K., Guo, S., Zhao, J., Li, J., Qin, H., Yang, M.: Mctrack: A unified 3d multi-object tracking framework for autonomous driving. In: 2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 4551–4558 (2025). https://doi.org/10.1109/IROS60139.2025.11245874
50. Wei, S., Luo, C., Luo, Y.: Boosting multimodal learning via disentangled gradient learning. In: IEEE/CVF International Conference on Computer Vision, ICCV 2025, Honolulu, HI, USA, October 19-25, 2025. pp. 1–10 (2025). https://doi.org/10.1109/ICCV51701.2025.02124
51. Wei, Y., Wei, Z., Rao, Y., Li, J., Zhou, J., Lu, J.: Lidar distillation: Bridging the beam-induced domain gap for 3d object detection. In: Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXXIX. pp. 179–195 (2022). https://doi.org/10.1007/978-3-031-19842-7_11
52. Weng, X., Wang, J., Held, D., Kitani, K.: 3d multi-object tracking: A baseline and new evaluation metrics. In: IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2020, Las Vegas, NV, USA, October 24, 2020 - January 24, 2021. pp. 10359–10366 (2020). https://doi.org/10.1109/IROS45743.2020.9341164
53. Xiang, H., Zheng, Z., Xia, X., Xu, R., Gao, L., Zhou, Z., Han, X., Ji, X., Li, M., Meng, Z., Jin, L., Lei, M., Ma, Z., He, Z., Ma, H., Yuan, Y., Zhao, Y., Ma, J.: V2x-real: A large-scale dataset for vehicle-to-everything cooperative perception. In: Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part LII. p. 455–470 (2024). https://doi.org/10.1007/978-3-031-72943-0_26
54. Xiao, A., Guan, D., Zhang, X., Lu, S.: Domain adaptive lidar point cloud segmentation with 3d spatial consistency. IEEE Transactions on Multimedia **26**, 5536–5547 (2024). https://doi.org/10.1109/TMM.2023.3335879
55. 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, 14-18 December 2022, Auckland, New Zealand. pp. 989–1000 (2022), https://proceedings.mlr.press/v205/xu23a.html, accessed: 25 Jun 2026
56. 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 (CVPR) (2023)
57. Xu, R., Xiang, H., Tu, Z., Xia, X., Yang, M., Ma, J.: V2x-vit: Vehicle-to-everything cooperative perception with vision transformer. In: Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXXIX. pp. 107–124 (2022). https://doi.org/10.1007/978-3-031-19842-7_7
58. 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

2022, Philadelphia, PA, USA, May 23-27, 2022. pp. 2583–2589 (2022). https://doi.org/10.1109/ICRA46639.2022.9812038
59. Yan, Y., Mao, Y., Li, B.: Second: Sparsely embedded convolutional detection. Sensors **18**(10) (2018). https://doi.org/10.3390/s18103337
60. Yang, L., Zhang, X., Li, J., Wang, C., Ma, J., Song, Z., Zhao, T., Song, Z., Wang, L., Zhou, M., Shen, Y., Lv, C.: V2x-radar: A multi-modal dataset with 4d radar for cooperative perception. Advances in Neural Information Processing Systems (NeurIPS) (2025)
61. Ye, X., Shu, M., Li, H., Shi, Y., Li, Y., Wang, G., Tan, X., Ding, E.: Rope3D: The Roadside Perception Dataset for Autonomous Driving and Monocular 3D Object Detection Task. In: Conference on Computer Vision and Pattern Recognition. pp. 21341–21350 (2022). https://doi.org/10.1109/CVPR52688.2022.02065
62. Yin, T., Zhou, X., Krähenbühl, P.: Center-based 3d object detection and tracking. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021. pp. 11784–11793 (2021). https://doi.org/10.1109/CVPR46437.2021.01161
63. Yongqiang, D., Dengjiang, W., Gang, C., Bing, M., Xijia, G., Yajun, W., Jianchao, L., Yanming, F., Juanjuan, L.: Baai-vanjee roadside dataset: Towards the connected automated vehicle highway technologies in challenging environments of china (2021), https://arxiv.org/abs/2105.14370, accessed: 25 Jun 2026
64. 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)
65. Zhang, G., Fan, L., He, C., Lei, Z., Zhang, Z., Zhang, L.: Voxel mamba: Group-free state space models for point cloud based 3d object detection. In: Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024 (2024)
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, L.Z.Y.L.A.L.L.L.R.M.B.B.M.: Omnihd-scenes: A next-generation multimodal dataset for autonomous driving. IEEE transactions on pattern analysis and machine intelligence **48 6**, 7032–7049 (2026). https://doi.org/10.1109/TPAMI.2026.3663672
68. Zhu, X., Sheng, H., Cai, S., Deng, B., Yang, S., Liang, Q., Chen, K., Gao, L., Song, J., Ye, J.: Roscenes: A large-scale multi-view 3d dataset for roadside perception. In: European Conference on Computer Vision. pp. 331–347. Springer (2024)
69. Zimmer, W., Creß, C., Nguyen, H.T., Knoll, A.C.: Tumtraf intersection dataset: All you need for urban 3d camera-lidar roadside perception. In: 2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC). pp. 1030–1037 (2023). https://doi.org/10.1109/ITSC57777.2023.10422289
70. Zimmer, W., Wardana, G.A., Sritharan, S., Zhou, X., Song, R., Knoll, A.C.: Tumtraf v2x cooperative perception dataset. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 22668–22677 (2024)