TUMTraf V2X Cooperative Perception Dataset

Walter Zimmer[1] Gerhard Arya Wardana[1] Suren Sritharan[1]
Xingcheng Zhou[1] Rui Song[1,2] Alois C. Knoll[1]

[1]Technical University of Munich [2]Fraunhofer IVI

https://tum-traffic-dataset.github.io/tumtraf-v2x

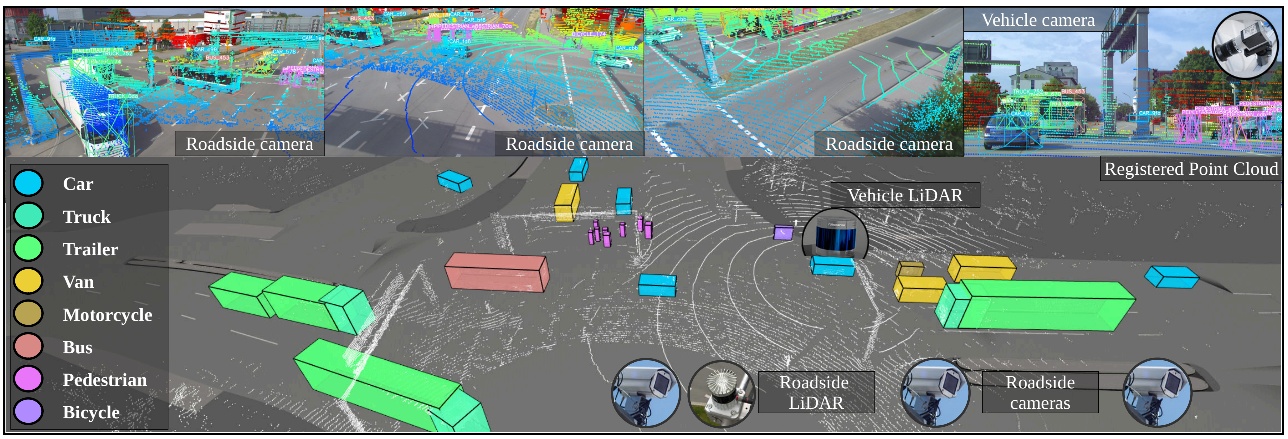


Figure 1. **Visualization** of 3D box labels and tracks in our **TUMTraf V2X Cooperative Perception Dataset**. The top part shows the labels projected into the four camera images. The part below shows a point cloud from two LiDARs with 3D box labels of the same scene.

Abstract

Cooperative perception offers several benefits for enhancing the capabilities of autonomous vehicles and improving road safety. Using roadside sensors in addition to onboard sensors increases reliability and extends the sensor range. External sensors offer higher situational awareness for automated vehicles and prevent occlusions. We propose CoopDet3D, a cooperative multi-modal fusion model, and TUMTraf-V2X, a perception dataset, for the cooperative 3D object detection and tracking task. Our dataset contains 2,000 labeled point clouds and 5,000 labeled images from five roadside and four onboard sensors. It includes 30k 3D boxes with track IDs and precise GPS and IMU data. We labeled nine categories and covered occlusion scenarios with challenging driving maneuvers, like traffic violations, near-miss events, overtaking, and U-turns. Through multiple experiments, we show that our CoopDet3D camera-LiDAR fusion model achieves an increase of +14.36 3D mAP compared to a vehicle camera-LiDAR fusion model. Finally, we make our dataset, model, labeling tool, and devkit publicly available on our website.

1 . Introduction

Cooperative perception involves the fusion of onboard sensor data and roadside sensor data, and it offers several advantages for enhancing the capabilities of autonomous vehicles and improving road safety. Using data from multiple sources makes the perception more robust to sensor failures or adverse environmental conditions. Roadside sensors provide an elevated view that helps to detect obstacles early. Moreover, they are also beneficial for precise vehicle localization and reduce the computational load of automated vehicles by offloading some perception tasks to the roadside sensors. Roadside sensors provide a global perspective of the traffic and offer a comprehensive situational awareness when fused with onboard sensor data. There are also fewer false positives or negatives because cooperative perception cross-validates the information from different sensors.

Infrastructure sensors can share perception-related information with vehicles through V2X. Due to minimal delay, and real-time capabilities, the infrastructure-based perception systems can further enhance the situational awareness and decision-making processes of vehicles.

Intelligent Transportation Systems (ITS) like the Testbed

Table 1. Comparison of 3D cooperative V2X perception datasets with our proposed TUMTraf-V2X Cooperative Perception dataset (I=Infrastructure, V=Vehicle).

Dataset	OPV2V [36]	V2XSet [35]	V2X-Sim [20]	V2V4Real [37]	DAIR-V2X-C [41]	V2X-Seq (SPD) [43]	**TUMTraf-V2X (Ours)**
Year	2022	2022	2022	2022	2022	2023	2024
V2X	V2V	V2V&I	V2V&I	V2V	V2I	V2I	V2I
Real data	-	-	-	✓	✓	✓	✓
Annotation range	120 m	120 m	70 m	200 m	280 m	280 m	200 m
Day & night scenes	-	-	-	-	✓	✓	✓
# object classes	1	1	1	5	10	9	8
Track IDs	-	-	✓	✓	-	✓	✓
HD Maps	✓	✓	✓	✓	-	✓	✓
# of sensors (I | V)	- | 6*	- | 6*	5 | 7	- | 8‡	2 | 3	2 | 3	5 | 4
Available worldwide	✓	✓	✓	✓	-	-	✓
Traffic violations	-	-	-	-	-	-	✓
Labeled attributes§	-	-	-	-	-	-	✓
OpenLABEL format	-	-	-	-	-	-	✓
# Point Clouds	11k	11k	10k	20k	39k	15k	2.0k
# Images	44k	44k	60k	40k†	39k	15k	5.0k
# 3D Boxes	233k	233k	26k	240k	464k	10.45k	29.38k
Location	CARLA	CARLA	CARLA	USA	China	China	Germany

† Image dataset has not been released yet.
* Value per vehicle. Multiple Conn. and Autom. Vehicles (CAVs) are used.
‡ Total sensors from 2 CAVs.
§ Weather, time of day, orientation, #3D points, occlusion, color, sub type

for Autonomous Driving [17] aim to improve safety by providing real-time traffic information. According to [9], testbeds extensively start using LiDAR sensors in their setups to create an accurate live digital twin of the traffic. Connected vehicles get a far-reaching view which enables them to react to breakdowns or accidents early. ITS systems also provide lane and speed recommendations to improve the traffic flow.

The key challenge with ego-centric vehicle datasets is that there are many occlusions from a vehicle perspective, e.g., if a large truck in front of the ego vehicle obscures the view. Roadside sensors located at a smart intersection provide a broad overview of the intersection and a full-surround view. Given the immense potential of ITS, there is a specific need for V2X datasets. Despite the high costs associated with collecting and labeling such datasets, this work addresses this challenge as a crucial step toward realizing large-scale ITS implementations.

Our contributions are as follows:

- We provide a high-quality V2X dataset for the cooperative 3D object detection and tracking task with 2,000 labeled point clouds and 5,000 labeled images. In total, 30k 3D bounding boxes with track IDs were labeled in challenging traffic scenarios like near-miss events, overtaking scenarios, U-turn maneuvers, and traffic violation events.
- We open-source our 3D bounding box annotation tool (*3D BAT* v24.3.2) to label multi-modal V2X datasets.
- We propose *CoopDet3D*, a cooperative 3D object detection model, and show in extensive experiments and ablation studies that it outperforms single view models on our V2X dataset by +14.3 3D mAP.
- Finally, we provide a development kit to load the annotations in the widely recognized and standard format OpenLABEL [13], to facilitate a seamless integration and utilization of the dataset. Furthermore, it can preprocess, visualize, and convert labels to and from different dataset formats, and evaluate perception and tracking methods.

2 . Related work

3D autonomous driving datasets are mainly categorized based on the viewpoint. Table 1 highlights the main differences between our proposed dataset and other V2X datasets.

2.1. Single viewpoint datasets

Single viewpoint datasets are obtained from a single point of reference, either an ego-vehicle or roadside infrastructure. Onboard sensor-based datasets like KITTI [12], nuScenes [6], and Waymo [30] contain a diverse set of sensor data collected from a moving vehicle equipped with multiple sensors, including high-resolution cameras, LiDARs, radars, and GPS/INS systems. These datasets are abundant and provide many annotated data, including bounding boxes, track IDs, segmentation masks, and depth maps under different urban driving scenarios.

On the other hand, roadside sensor-based datasets are in the infancy stage. High-quality multi-modal (camera and LiDAR) datasets are presented in [5, 10, 48], which

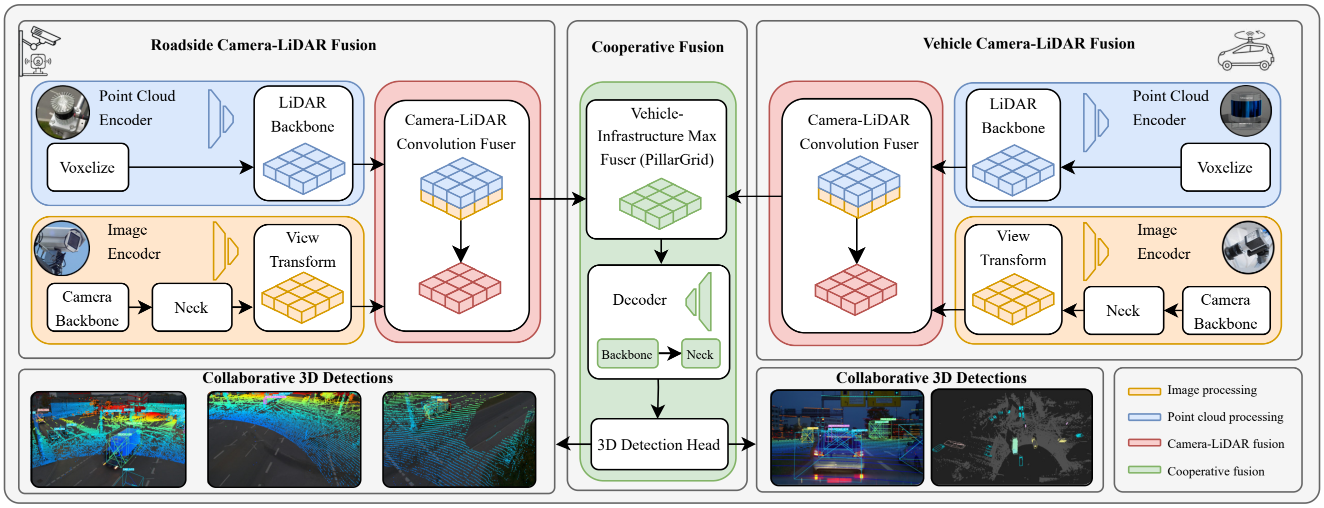


Figure 2. Our CoopDet3D framework is a multi-modal cooperative fusion system, comprising three distinct fusion pipelines. 1) The roadside camera-LiDAR fusion pipeline fuses three camera images and one LiDAR point cloud by extracting features and transforming them into a BEV representation. 2) The vehicle camera-LiDAR fusion pipeline fuses the vehicle camera feature map with the vehicle point cloud feature map using a convolutional fuser. 3) The vehicle and infrastructure feature maps are then fused by applying an element-wise max-pooling operation (Max Fuser). In the end, we use the TransFusion [1] 3D detection head to obtain 3D bounding box predictions.

are obtained from Infrastructure Perception Systems (IPS). Similarly, in [40], the authors provide a dataset consisting of only images taken from different viewpoints and under varying traffic conditions. These datasets provide a top-down view of a crowded intersection under different conditions and, as such, can overcome issues such as occlusions created by other vehicles and thereby have a higher number of object labels than onboard sensor-based datasets.

2.2. V2X datasets

V2X datasets exploit the information from multiple viewpoints to gain additional knowledge regarding the environments. In this way, they overcome the limitations of single viewpoint datasets such as occlusion, limited field of view (FOV), and low point cloud density.

DAIR-V2X dataset family [41] is one of the foremost cooperative multi-modal datasets introduced. It contains three subsets: an intersection, a vehicle, and a cooperative dataset. The cooperative dataset contains 464k 3D box labels belonging to 10 classes, making it one of the largest cooperative datasets. The V2X-Seq dataset [43] extends selected sequences of the DAIR-V2X dataset with track IDs and is partitioned into a sequential perception dataset (SPD) and a trajectory forecasting dataset. Despite these, the lack of specific information, such as the labeling methodology used, the exact models of the sensors deployed, the distribution of the classes, and the scenarios within the dataset, leads to uncertainty in the extendability and application of this dataset in varying conditions.

In V2V4Real [37], the authors propose a multi-modal cooperative dataset focusing only on V2V perception. Two vehicles equipped with cameras, LiDAR, and GPS/IMU integration systems are used to collect multi-modal sensor data for diverse scenarios. As opposed to all other V2X datasets, this focuses on V2V perception, and though it is of similar size to other cooperative datasets, it contains fewer classes and 3D bounding box information.

Simulated multi-agent perception datasets have been proposed in [20, 35, 36]. These datasets contain multi-modal sensor data (camera and LiDAR) obtained from roadside units (RSUs) and multiple ego vehicles, which enable collaborative perception. They use a combination of simulators such as SUMO [24], CARLA [11], and OpenCDA [36] for flow simulation, data retrieval, and V2X communication. However, the utility of the dataset is still limited due to the simulated nature of the data, and its extendability to real-life applications has not been studied in detail.

2.3. V2X perception models for object detection

The datasets presented above have been used to develop various models for a wide variety of tasks, with the majority focusing on 3D object detection. Different approaches have been taken depending on the availability and challenges, and these methods are grouped based on the number of nodes employed and the modalities used for detection.

Most 3D object detection models use multi-modal sensor data obtained from a single point of view, which is often an ego-vehicle. Due to the popularity and abundant availability of onboard datasets [6, 12, 30], most models use images, point clouds, or both modalities. Image-based models were the pioneers in 3D object detection due to their low cost and simplicity, and both vehicular camera-based models [18, 34] and infrastructure camera-based models [39] have been proposed. LiDAR-based 3D object detection models [19, 46, 49] became popular since LiDAR point clouds provide 3D depth information and are robust, especially in adverse weather conditions and nighttime scenarios. Fusion models combine the information obtained from both images and point clouds and have been shown to outperform the prior methods [47]. Single viewpoint fusion models use either vehicular camera and LiDAR [23, 32, 38] or infrastructure camera and LiDAR [47] for 3D object detection.

Cooperative perception models, which use data from multiple viewpoints, have been shown to overcome issues related to occlusion, which were often present in vehicular sensor-based models. V2I cooperative perception models [2, 3, 14, 26, 33, 35, 42] use the sensor data from both vehicles and infrastructure and V2V models [15, 29] communicate the sensor data between multiple vehicles. In this work, our cooperative multi-modal dataset is one contribution among others. Thus, while most of the prior works focus on unimodal cooperative perception using either LiDAR point clouds [2, 7] or camera images [15], we benchmark our dataset with CoopDet3D, a deep fusion based cooperative multi-modal 3D object detection model based on BEVFusion [15] and PillarGrid [2].

3 . TUMTraf-V2X Dataset

Our TUMTraf V2X Cooperative Perception Dataset focuses on challenging traffic scenarios and various day and nighttime scenes. The data is further annotated, emphasizing high-quality labels through careful labeling and high-quality review processes. It also contains dense traffic and fast-moving vehicles, which reveals the specific challenges in cooperative perception, such as pose estimation errors, latency, and synchronization. Furthermore, we provide sensor data from nine different sensors covering the same traffic scenes under diverse weather conditions and lighting variations. The infrastructure sensors are oriented in all four directions of the intersection to get a 360° view, which leads to better perception results. Finally, it contains rare events like traffic violations where pedestrians cross the road at a busy four-way intersection while the crossing light is lit red.

3.1. Sensor setup

Our TUMTraf V2X Cooperative Perception Dataset was recorded on an ITS system with nine sensors.

The infrastructure sensor setup is the following:

- 1x Ouster LiDAR OS1-64 (gen. 2), 64 vert. layers, 360° FOV, below horizon config., 10 cm acc. @120 m range
- 4x Basler ace acA1920-50gc, 1920×1200, Sony IMX174 with 8 mm lenses

On the vehicle, the following sensors were used:

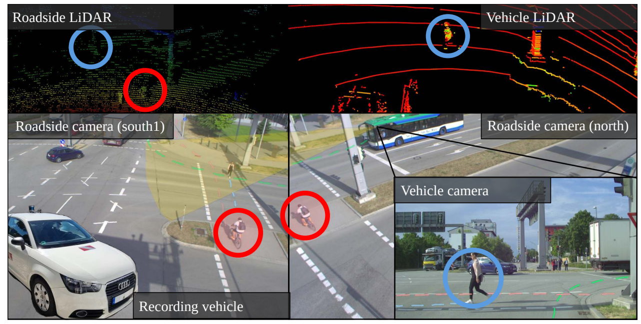


Figure 3. Demonstration of a possible V2X occlusion scenario. A pedestrian (blue) is crossing the road in front of the ego vehicle. An occluded bicycle is marked in red. The recording vehicle with the sensor setup is shown in the bottom left corner.

- 1x Robosense RS-LiDAR-32, 32 vert. layers, 360° FOV, 3 cm accuracy @200 m range
- 1x Basler ace acA1920-50gc, 1920×1200, Sony IMX174 with 16 mm lens
- 1x Emlid Reach RS2+ multi-band RTK GNSS receiver
- 1x XSENS MTi-30-2A8G4 IMU

3.2. Sensor calibration and registration

We synchronize the cameras and LiDARs in the spatial and temporal domain. First, we determine the intrinsic camera parameters and the radial and tangential image distortions by using a checkerboard target. We then calibrate the roadside LiDAR with the roadside cameras by picking 100-point pairs in the point cloud and camera image. Extrinsic parameters (rotation and translation) are calculated by minimizing the reprojection error of 2D-3D point correspondences [25]. We follow the same procedure for onboard camera-LiDAR calibration. Finally, we calibrate the onboard LiDAR to the roadside LiDAR. This spatial registration is done by first estimating a coarse transformation. We pick ten 3D point pairs in each point cloud and minimize their distance using the least squares method. Then, we apply the point-to-point Iterative Closest Point (ICP) algorithm [4] to get the fine transformation between the point clouds.

We label the vehicle and infrastructure point clouds after registering them. The coarse registration was done by measuring the GPS position of the onboard LiDAR and the roadside LiDAR. Then, we transform every 10th onboard point cloud to the coordinate system of the infrastructure point cloud. The fine registration was done by applying the point-to-point ICP to get an accurate V2I transformation matrix. All rotations of the point cloud frames in between are interpolated based on the spherical linear interpolation (SLERP) [28] method:

$$SLERP(q_0, q_1, t) = q_0(q_0^{-1}q_1)^t, \quad (1)$$

where q_0 and q_1 are the quaternions representing the ro-

Table 2. Evaluation results (mAP_{BEV} and mAP_{3D}) of CoopDet3D on our TUMTraf-V2X test set in south2 FOV.

Config.		$\mathbf{mAP_{BEV}}$ ↑	$\mathbf{mAP_{3D}}$ ↑			
Domain	**Modality**		**Easy↑**	**Mod.↑**	**Hard↑**	**Avg.↑**
Vehicle	Camera	46.83	31.47	37.82	30.77	30.36
Vehicle	LiDAR	85.33	85.22	76.86	69.04	80.11
Vehicle	Cam+LiDAR	84.90	77.60	72.08	73.12	76.40
Infra.	Camera	61.98	31.19	46.73	40.42	35.04
Infra.	LiDAR	92.86	86.17	88.07	75.73	84.88
Infra.	Cam+LiDAR	92.92	87.99	**89.09**	**81.69**	87.01
Coop.	Camera	68.94	45.41	42.76	57.83	45.74
Coop.	LiDAR	93.93	92.63	78.06	73.95	85.86
Coop.	Cam+LiDAR	**94.22**	**93.42**	88.17	79.94	**90.76**

Table 3. Evaluation results of infrastructure-only CoopDet3D vs. InfraDet3D [47] on TUMTraf Intersection test set [48]. South 1 and South 2 refer to sensors covering different FOVs.

			$\mathbf{mAP_{3D}}$ ↑			
Model	**FOV**	**Modality**	**Easy↑**	**Mod.↑**	**Hard↑**	**Avg.↑**
InfraDet3D	south 1	LiDAR	75.81	47.66	**42.16**	55.21
CoopDet3D	south 1	LiDAR	**76.24**	**48.23**	35.19	**69.47**
InfraDet3D	south 2	LiDAR	38.92	46.60	**43.86**	43.13
CoopDet3D	south 2	LiDAR	**74.97**	**55.55**	39.96	**69.94**
InfraDet3D	south 1	Cam+LiDAR	67.08	31.38	35.17	44.55
CoopDet3D	south 1	Cam+LiDAR	**75.68**	**45.63**	**45.63**	**66.75**
InfraDet3D	south 2	Cam+LiDAR	58.38	19.73	33.08	37.06
CoopDet3D	south 2	Cam+LiDAR	**74.73**	**53.46**	**41.96**	**66.89**

tations of the start and end frames and $t \in [0, 1]$. Translation vectors $\mathbf{T}_0$ and $\mathbf{T}_1$ were obtained using linear interpolation:

$$\mathbf{T}(t) = \mathbf{T}_0 + t(\mathbf{T}_1 - \mathbf{T}_0). \quad (2)$$

This dual interpolation strategy ensures that the estimated transformations between the frames are smooth and geometrically accurate, thus adhering closely to the actual movements of the vehicle over time.

3.3. Data selection and labeling

We selected the data based on challenging traffic scenarios, like U-turns, tailgate events, and traffic violation maneuvers. Besides the high traffic density of 31 objects per frame, we selected frames with high-class coverage. We selected 700 frames during sunny daytime and 100 frames during cloudy nighttime for labeling. The camera and LiDAR data were recorded into rosbag files at 15 Hz and 10 Hz, respectively. We extracted and synchronized the data based on ROS [27] timestamps and labeled it with our *3D BAT* (v24.3.2) annotation tool[1]. We improved the 3D BAT [45] baseline labeling tool to label 3D objects faster and more precisely with a one-click annotation feature. The annotators were instructed to label traffic participants while examining the images. Objects are still labeled, even if they have no 3D points inside, but are visible in the images. Extremities (e.g., pedestrian limbs) are included in the bounding box, but side mirrors of vehicles aren't. If a pedestrian carries an object, that object is included in the bounding box. If two or more pedestrians are carrying an object, only the box of one will include the object. After labeling, each annotator checked the work of other annotators manually frame-by-frame. When errors were found, the original annotator was notified, and they fixed it. This helps ensure that the labels in our dataset are high quality.

3.4. Data structure and format

We record eight different scenes, each 10 sec. long, from vehicle and infrastructure perspectives using nine sensors and split the data into a train (80%), val. (10%), and test (10%) set. We use stratified sampling to distribute all sets' object classes equally (see Fig. 6a). Labels are provided in the ASAM OpenLABEL [13] standard.

3.5. Dataset development kit

We provide a devkit to work with our dataset. In addition to generating the data statistics, it provides modules for multi-class stratified splitting (train/val/test), point cloud registration, loading annotations in OpenLABEL format, evaluation of detection and tracking results, pre-processing steps such as point cloud filtering, and post-processing such as bounding box filtering. The statistics (Fig. 4, 5, and 6) were created using our devkit. It also contains modules to convert the labels from OpenLABEL to KITTI or our custom nuScenes format with timestamps instead of tokens and vice versa. This devkit enables users to migrate their models and make them compatible with our dataset format. We release our devkit[2] under the MIT license and the dataset under the Creative Commons (CC) BY-NC-ND 4.0 license.

4 . Benchmark

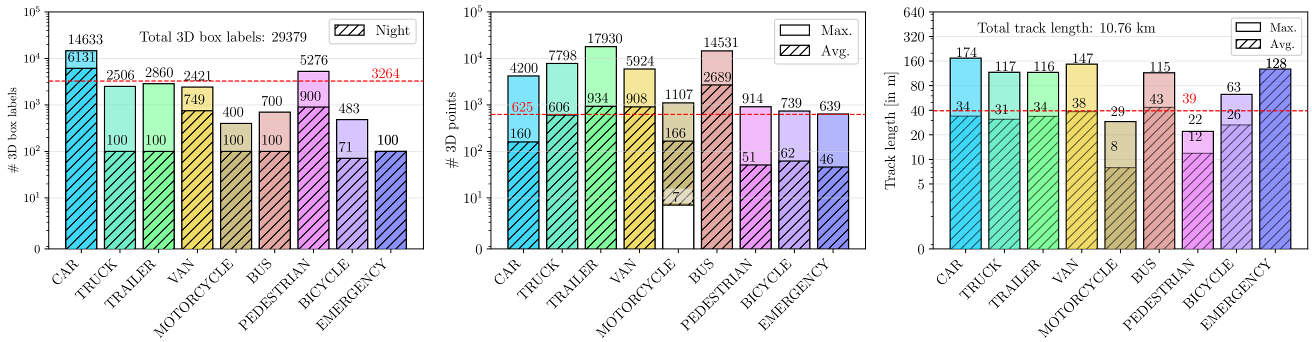
We propose CoopDet3D, an extension of BEVFusion [23] and PillarGrid [2] for deep cooperative multi-modal 3D object detection and benchmark it on *TUMTraf-V2X mini*.

4.1. Evaluation metrics

The accuracy is measured in terms of the mean average precision (mAP). Two types of mAP measures are used: BEV mAP considers the BEV center distance, and the results are obtained using the same evaluation methodology used in BEVFusion, which in turn uses the evaluation protocol of nuScenes [6]. Similarly, the 3D mAP measure considers the intersection in 3D, and the results are obtained using the evaluation script of our TUM Traffic dataset devkit. The runtime is evaluated using frames per second (FPS) as the metric and the results were obtained by measuring the time

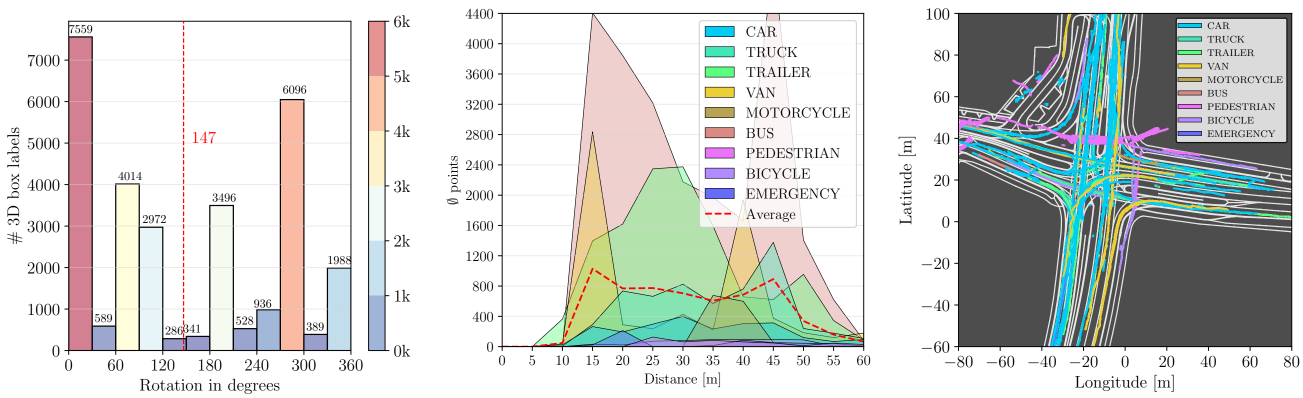
[1] https://github.com/walzimmer/3d-bat

[2] https://github.com/tum-traffic-dataset/tum-traffic-dataset-dev-kit



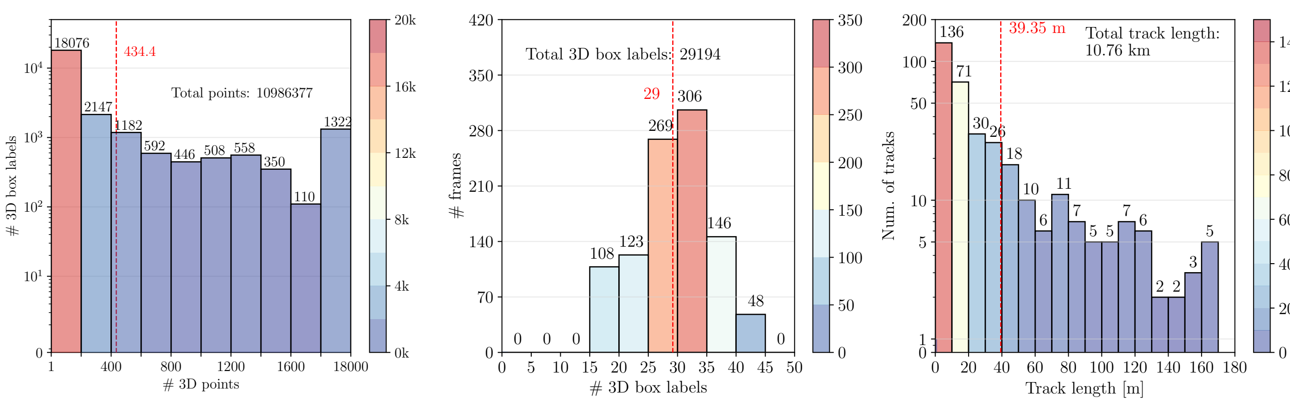
(a) Distribution of objects between day and night. (b) Avg. and max. num. of 3D points for each class. (c) Avg. and max. track length for all classes.

Figure 4. Our TUMTraf-V2X dataset contains 30k 3D box labels in total and is balanced among nine different object classes. (a) Cars (14,633) and pedestrians (5,276) are highly represented in the dataset. (b) 3D box labels contain, on average 625 points inside which shows the density of the labeled objects. The BUS class has the highest point density. (c) All traffic participants are tracked for 39 m on average. Emergency vehicles have the highest average track length of 128 m, whereas the CAR class contains a max. track length of 174 m.



(a) Visualization of rotations (yaw). (b) 3D points grouped by distance. (c) BEV visualization of tracks.

Figure 5. Our dataset was recorded at a crowded intersection with many left and right turns. (a) Most of the vehicles (7,559) are driving in the east direction (0 degrees). (b) 3D boxes were labeled up to 200 m range and are very dense between 10 and 60 m. (c) The visualization of BEV tracks shows where pedestrians and bicycles are crossing the road.



(a) Histogram of 3D points. (b) Histogram of 3D box labels. (c) Histogram of track lengths.

Figure 6. (a) Most 3D box labels contain between 1 and 200 3D points inside, with an average of 434 3D points, excluding empty boxes. Objects that were close to both LiDAR sensors even contained up to 18k 3D points. (b) Frames contain between 15 and 45 traffic participants, with an average of 29. (c) Objects were tracked up to 170 m, and the average track length is 39.35 m.

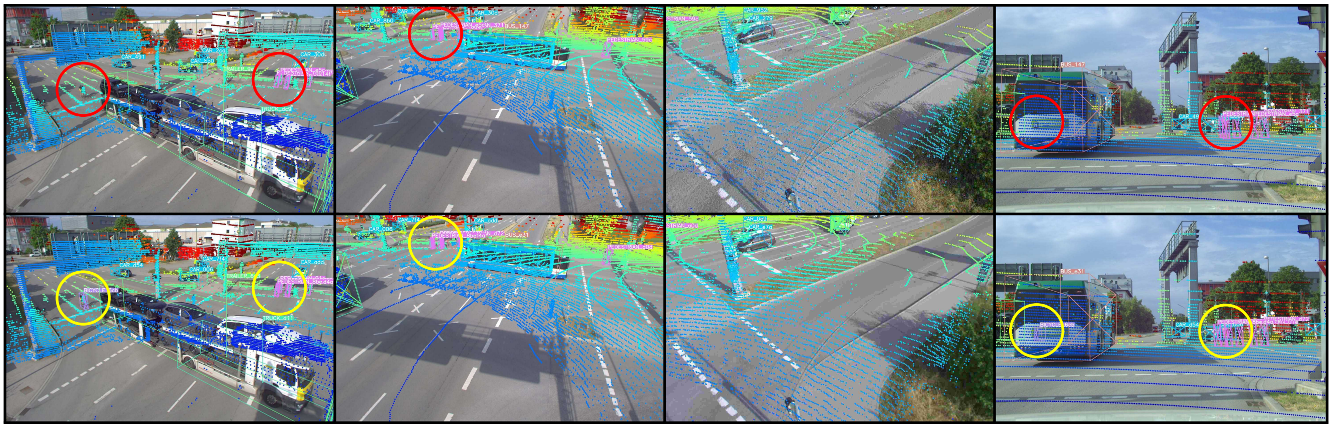


Figure 7. Qualitative results on the TUMTraf-V2X Cooperative Perception test set. The first row shows the inference results of the onboard (vehicle-only) camera-LiDAR fusion with 23 detected objects. In the second row, the results of the cooperative vehicle-infrastructure camera-LiDAR fusion are visualized. Here, all 25 traffic participants could be detected with the support of roadside sensors.

Table 4. Ablation study on cooperative 3D object detection with 11 combinations of camera and LiDAR backbones. The best trade-off between speed and accuracy is highlighted in gray. ND = non-deterministic, TS = TorchSparse, ↻ = retrained. VRAM is in GiB.

Backbone Configuration	$\mathrm{mAP_{BEV}}\uparrow$	**FPS↑**	**VRAM↓**
VoxelNet ND + SwinT	93.47	6.30	6.69
VoxelNet ND + YOLOv8 s	92.94	7.24	6.39
VoxelNet TS + SwinT	93.51	8.84	4.61
VoxelNet TS + YOLOv8 s	92.94	10.66	**4.28**
VoxelNet TS + YOLOv8 s ↻	94.31	10.66	**4.28**
PointPillars 512 + Swin T	**94.43**	9.00	4.94
PointPillars 512 + YOLOv8 s	94.27	11.14	4.63
PointPillars 512 + YOLOv8 s ↻	94.25	11.14	4.63
PointPillars 512_2x + Swin T	92.79	9.06	4.94
PointPillars 512_2x + YOLOv8 s	94.16	**11.20**	4.63
PointPillars 512_2x + YOLOv8 s ↻	94.22	**11.20**	4.63

needed by the model to run one full inference, including data preprocessing and voxelization. The first five iterations are skipped as a warmup since they are usually considerably slower than the average. Finally, the complexity of the model is measured in terms of the maximal VRAM usage across all GPUs during training and testing.

4.2. CoopDet3D model

Our *CoopDet3D* uses a BEVFusion-based backbone for camera-LiDAR fusion on the vehicle and the infrastructure sides separately to obtain the vehicle and infrastructure features. The best backbone for image and point cloud feature extraction was chosen through multiple ablation studies. Then, inspired by the method proposed by PillarGrid [2], an element-wise max-pooling operation is proposed to fuse the resulting fused camera-LiDAR features of vehicle and infrastructure together. Finally, the detection head from BEVFusion is used for 3D detection from the fused feature.

The architecture of *CoopDet3D* is shown in Fig. 2.

First, we disable the camera feature extraction nodes and train the LiDAR-only model for 20 epochs. Then, we use pre-trained weights for the cooperative model and fine-tune the entire model for eight further epochs. Hyperparameter tuning revealed that the default hyperparameters of BEVFusion [23] gave the best results, and such were not modified. The preprocessing steps are also the same as for BEVFusion, but we change the point cloud range to $[-75, 75]$ in the x- and y-scale and $[-8, 0]$ in the z-scale since the dataset used in this case is different. Furthermore, we use 3x NVIDIA RTX 3090 GPUs with 24 GB VRAM for training and a single GPU for evaluation. We open-source our model and provide pre-trained weights[3].

4.3. Experiments and ablation studies

The objective of these experiments is to highlight the importance of our V2X multi-viewpoint dataset as opposed to single-viewpoint datasets. As such, we conduct multiple experiments and ablation studies with data obtained from each viewpoint and compare the results on the proposed model.

Ablation on viewpoint and modality. We conduct multiple experiments with all possible combinations of a) viewpoints: vehicle-only, infrastructure-only, cooperative, and b) modalities: camera-only, LiDAR-only, and camera-LiDAR fusion. Table 2 shows the mAP achieved by CoopDet3D for each of these combinations.

We observe that the results follow a general pattern of cooperative performance being better than infrastructure-only, which is, in turn, better than vehicle-only. Furthermore, fusion models perform better than LiDAR-only models, which in turn are better than camera-only models. Figure 7 shows qualitative results between our vehicle-only

[3] https://github.com/tum-traffic-dataset/coopdet3d

camera-LiDAR fusion model and our cooperative vehicle-infrastructure camera-LiDAR fusion model. Again, we observe from these samples that the cooperative perception model is able to detect 25 traffic participants, whereas the vehicle-only model is only able to detect 23 objects due to occlusions and a limited field of view.

Ablation on fusion level. We compare our CoopDet3D model to the current SOTA camera-LiDAR fusion method InfraDet3D [47] on the TUMTraf Intersection test set [48]. The proposed method uses deep fusion, whereas the InfraDet3D method is a late fusion method. Table 3 shows the performance of our model against InfraDet3D. The results show that the proposed deep fusion method outperforms the SOTA late fusion model in all metrics, except in the hard difficulty in LiDAR-only mode. Furthermore, Fig. 8 shows two sample images taken during day and nighttime, wherein deep fusion again outperforms late fusion.

We note that these experiments were conducted in an offline setting, disregarding other considerations for simplicity. However, when deploying it in real life, factors such as the transmission bandwidth should also be considered. Since we observed that deep feature fusion generally leads to higher efficacy, the V2I transmissions should contain these features instead of infrastructure bounding boxes.

Ablation on backbones. As an ablation study, we present the results of the experiments to find the best backbone and model configuration for the cooperative camera-LiDAR fusion model. For the camera backbone, SwinT [22] and MMYOLO's [8] implementation of YOLOv8 [16] were considered. For the LiDAR backbone, VoxelNet [44] and PointPillars [19] were considered. In addition, VoxelNet was implemented with two different backends, namely SPConv v2 and Torchsparse [31]. For PointPillars, two grid sizes are considered 512×512 for both train and test grids (PointPillars 512) and 512×512 train grid with 1024×1024 test grid (PointPillars 512_2x). The results of these experiments are shown in Table 4.

The results show that only models that use any combination of VoxelNet Torchsparse, both PointPillars variants, and YOLOv8 are able to run above 10 FPS. From these configurations, we choose PointPillars 512_2x with YOLOv8 as the best configuration for all the above experiments as it achieves the best results across all the ablation studies. This is a promising result since we also know that this backbone configuration is able to run in real-time (11.2 FPS) on an RTX 3090 without using TensorRT acceleration.

An interesting observation is that utilizing pre-trained weights for transfer learning of YOLOv8 is not always beneficial, as the results from PointPillars 512 + YOLOv8 s show. This is likely because the pre-trained weights were from MS COCO [21], and they have a very different data domain compared to our dataset. Since MS COCO is also much larger than our dataset in terms of camera images, retraining harms the performance of the model slightly.

Figure 8. Qualitative results of our CoopDet3D (left) and the InfraDet3D (right) model on the TUMTraf Intersection test set during day and nighttime. Detected objects marked with a red circle were classified correctly by CoopDet3D.

In terms of efficiency, the goal of these experiments was to verify that the proposed CoopDet3D model with the best configuration provides the highest accuracy while also being able to run in real-time (minimum of 10 Hz). Furthermore, it should also be feasible to train the model on a high-performance GPU and perform inference on a mid-range consumer GPU deployable on an edge device. The results concerning the VRAM usage during inference show that the complexity of the model makes this feasible.

5 . Conclusion and future work

This work proposes the TUMTraf-V2X dataset, a multi-modal multi-view V2X dataset for cooperative 3D object detection and tracking. Our dataset focuses on challenging traffic scenarios at an intersection and provides views from the infrastructure and the ego vehicle. To benchmark the dataset, we propose CoopDet3D – a baseline model for cooperative perception. Experiments show that cooperative fusion leads to higher efficacy than its unimodal and single-view camera-LiDAR fusion counterparts. Furthermore, cooperative fusion leads to an improvement of +14.3 3D mAP compared to vehicle-only perception, highlighting the need for V2X datasets. Finally, we provide our *3D BAT* v24.3.2 labeling tool and devkit to load, parse, and visualize the dataset. It also includes modules for pre- and postprocessing and evaluation. Future efforts will integrate this platform into online environments, enabling a broader range of infrastructure-based, real-time perception applications.

Acknowledgment

This research was supported by the Federal Ministry of Education and Research in Germany within the *AUTOtech.agil* project, Grant Number: 01IS22088U.

References

[1] Xuyang Bai, Zeyu Hu, Xinge Zhu, Qingqiu Huang, Yilun Chen, Hongbo Fu, and Chiew-Lan Tai. Transfusion: Robust lidar-camera fusion for 3d object detection with transformers, 2022. 3

[2] Zhengwei Bai, Guoyuan Wu, Matthew J Barth, Yongkang Liu, Emrah Akin Sisbot, and Kentaro Oguchi. Pillargrid: Deep learning-based cooperative perception for 3d object detection from onboard-roadside lidar. In *IEEE 25th International Conference on Intelligent Transportation Systems (ITSC)*, pages 1743–1749. IEEE, 2022. 4, 5, 7

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

[4] Paul J Besl and Neil D McKay. Method for registration of 3-d shapes. In *Sensor fusion IV: control paradigms and data structures*, pages 586–606. Spie, 1992. 4

[5] Steffen Busch, Christian Koetsier, Jeldrik Axmann, and Claus Brenner. Lumpi: The leibniz university multi-perspective intersection dataset. In *2022 IEEE Intelligent Vehicles Symposium (IV)*, pages 1127–1134. IEEE, 2022. 2

[6] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11621–11631, 2020. 2, 3, 5

[7] Ziming Chen, Yifeng Shi, and Jinrang Jia. Transiff: An instance-level feature fusion framework for vehicle-infrastructure cooperative 3d detection with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18205–18214, 2023. 4

[8] MMYOLO Contributors. MMYOLO: OpenMMLab YOLO series toolbox and benchmark. `https://github.com/open-mmlab/mmyolo`, 2022. 8

[9] Christian Creß, Zhenshan Bing, and Alois C. Knoll. Intelligent transportation systems using roadside infrastructure: A literature survey. *IEEE Transactions on Intelligent Transportation Systems*, pages 1–0, 2023. 2

[10] Christian Creß, Walter Zimmer, Leah Strand, Maximilian Fortkord, Siyi Dai, Venkatnarayanan Lakshminarasimhan, and Alois Knoll. A9-dataset: Multi-sensor infrastructure-based dataset for mobility research. In *2022 IEEE Intelligent Vehicles Symposium (IV)*, pages 965–970, 2022. 2

[11] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. CARLA: An open urban driving simulator. In *Proceedings of the 1st Annual Conference on Robot Learning*, pages 1–16, 2017. 3

[12] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013. 2, 3

[13] Nicco Hagedorn. OpenLABEL Concept Paper. 2, 5

[14] Yuze He, Li Ma, Zhehao Jiang, Yi Tang, and Guoliang Xing. Vi-eye: Semantic-based 3d point cloud registration for infrastructure-assisted autonomous driving. In *Proceedings of the 27th Annual International Conference on Mobile Computing and Networking*, pages 573–586, 2021. 4

[15] Yue Hu, Yifan Lu, Runsheng Xu, Weidi Xie, Siheng Chen, and Yanfeng Wang. Collaboration helps camera overtake lidar in 3d detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9243–9252, 2023. 4

[16] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. Ultralytics yolo, 2023. 8

[17] Annkathrin Krämmer, Christoph Schöller, Dhiraj Gulati, Venkatnarayanan Lakshminarasimhan, Franz Kurz, Dominik Rosenbaum, Claus Lenz, and Alois Knoll. Providentia-a large-scale sensor system for the assistance of autonomous vehicles and its evaluation. *Journal of Field Robotics*, pages 1156–1176, 2022. 2

[18] Abhinav Kumar, Garrick Brazil, Enrique Corona, Armin Parchami, and Xiaoming Liu. Deviant: Depth equivariant network for monocular 3d object detection. In *European Conference on Computer Vision*, pages 664–683, 2022. 4

[19] Alex H Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. Pointpillars: Fast encoders for object detection from point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12697–12705, 2019. 4, 8

[20] Yiming Li, Dekun Ma, Ziyan An, Zixun Wang, Yiqi Zhong, Siheng Chen, and Chen Feng. V2x-sim: Multi-agent collaborative perception dataset and benchmark for autonomous driving. *IEEE Robotics and Automation Letters*, 7(4):10914–10921, 2022. 2, 3

[21] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Proceedings of European Conference on Computer Vision, ECCV 2014, Part V 13*, pages 740–755. Springer, 2014. 8

[22] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 8

[23] Zhijian Liu, Haotian Tang, Alexander Amini, Xinyu Yang, Huizi Mao, Daniela L Rus, and Song Han. Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. In *2023 IEEE Int. Conference on Robotics and Automation (ICRA)*, pages 2774–2781, 2023. 4, 5, 7

[24] Pablo Alvarez Lopez, Michael Behrisch, Laura Bieker-Walz, Jakob Erdmann, Yun-Pang Flötteröd, Robert Hilbrich, Leonhard Lücken, Johannes Rummel, Peter Wagner, and Evamarie Wießner. Microscopic traffic simulation using sumo. In *The 21st IEEE International Conference on Intelligent Transportation Systems*. IEEE, 2018. 3

[25] Kaj Madsen, Hans Bruun Nielsen, and Ole Tingleff. Methods for non-linear least squares problems. 2004. 4

[26] Donghao Qiao and Farhana Zulkernine. Cobevfusion: Cooperative perception with lidar-camera bird's-eye view fusion. *arXiv preprint arXiv:2310.06008*, 2023. 4

[27] Morgan Quigley, Ken Conley, Brian Gerkey, Josh Faust, Tully Foote, Jeremy Leibs, Rob Wheeler, Andrew Y Ng,

et al. Ros: an open-source robot operating system. In *ICRA workshop on open source software*, page 5. Kobe, Japan, 2009. 5

[28] Ken Shoemake. Animating rotation with quaternion curves. In *Proceedings of the 12th annual conference on Computer graphics and interactive techniques*, pages 245–254, 1985. 4

[29] Rui Song, Chenwei Liang, Hu Cao, Zhiran Yan, Walter Zimmer, Markus Gross, Andreas Festag, and Alois Knoll. Collaborative semantic occupancy prediction with hybrid feature fusion in connected automated vehicles. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 4

[30] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2446–2454, 2020. 2, 3

[31] Haotian Tang, Shang Yang, Zhijian Liu, Ke Hong, Zhongming Yu, Xiuyu Li, Guohao Dai, Yu Wang, and Song Han. TorchSparse++: Efficient Point Cloud Engine. In *Computer Vision and Pattern Recognition Workshops CVPRW*, 2023. 8

[32] Sourabh Vora, Alex H Lang, Bassam Helou, and Oscar Beijbom. Pointpainting: Sequential fusion for 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4604–4612, 2020. 4

[33] Sizhe Wei, Yuxi Wei, Yue Hu, Yifan Lu, Yiqi Zhong, Siheng Chen, and Ya Zhang. Asynchrony-robust collaborative perception via bird's eye view flow. In *Advances in Neural Information Processing Systems*, pages 28462–28477. Curran Associates, Inc. 4

[34] Zizhang Wu, Yuanzhu Gan, Lei Wang, Guilian Chen, and Jian Pu. Monopgc: Monocular 3d object detection with pixel geometry contexts. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4842–4849, 2023. 4

[35] Runsheng Xu, Hao Xiang, Zhengzhong Tu, Xin Xia, Ming-Hsuan Yang, and Jiaqi Ma. V2x-vit: Vehicle-to-everything cooperative perception with vision transformer. In *European conference on computer vision*, pages 107–124. Springer, 2022. 2, 3, 4

[36] Runsheng Xu, Hao Xiang, Xin Xia, Xu Han, Jinlong Li, and Jiaqi Ma. Opv2v: An open benchmark dataset and fusion pipeline for perception with vehicle-to-vehicle communication. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 2583–2589. IEEE, 2022. 2, 3

[37] Runsheng Xu, Xin Xia, Jinlong Li, Hanzhao Li, Shuo Zhang, Zhengzhong Tu, Zonglin Meng, Hao Xiang, Xiaoyu Dong, Rui Song, et al. V2v4real: A real-world large-scale dataset for vehicle-to-vehicle cooperative perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13712–13722, 2023. 2, 3

[38] Junjie Yan, Yingfei Liu, Jianjian Sun, Fan Jia, Shuailin Li, Tiancai Wang, and Xiangyu Zhang. Cross modal transformer: Towards fast and robust 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18268–18278, 2023. 4

[39] Lei Yang, Kaicheng Yu, Tao Tang, Jun Li, Kun Yuan, Li Wang, Xinyu Zhang, and Peng Chen. Bevheight: A robust framework for vision-based roadside 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21611–21620, 2023. 4

[40] Xiaoqing Ye, Mao Shu, Hanyu Li, Yifeng Shi, Yingying Li, Guangjie Wang, Xiao Tan, and Errui Ding. Rope3d: The roadside perception dataset for autonomous driving and monocular 3d object detection task. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21341–21350, 2022. 3

[41] Haibao Yu, Yizhen Luo, Mao Shu, Yiyi Huo, Zebang Yang, Yifeng Shi, Zhenglong Guo, Hanyu Li, Xing Hu, Jirui Yuan, et al. Dair-v2x: A large-scale dataset for vehicle-infrastructure cooperative 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21361–21370, 2022. 2, 3

[42] Haibao Yu, Yingjuan Tang, Enze Xie, Jilei Mao, Jirui Yuan, Ping Luo, and Zaiqing Nie. Vehicle-infrastructure cooperative 3d object detection via feature flow prediction. *arXiv preprint arXiv:2303.10552*, 2023. 4

[43] Haibao Yu, Wenxian Yang, Hongzhi Ruan, Zhenwei Yang, Yingjuan Tang, Xu Gao, Xin Hao, Yifeng Shi, Yifeng Pan, Ning Sun, et al. V2x-seq: A large-scale sequential dataset for vehicle-infrastructure cooperative perception and forecasting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5486–5495, 2023. 2, 3

[44] Yin Zhou and Oncel Tuzel. Voxelnet: End-to-end learning for point cloud based 3d object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4490–4499, 2018. 8

[45] Walter Zimmer, Akshay Rangesh, and Mohan Trivedi. 3d bat: A semi-automatic, web-based 3d annotation toolbox for full-surround, multi-modal data streams. In *2019 IEEE Intelligent Vehicles Symposium (IV)*, pages 1816–1821. IEEE, 2019. 5

[46] Walter Zimmer, Emec Ercelik, Xingcheng Zhou, Xavier Jair Diaz Ortiz, and Alois Knoll. A survey of robust 3d object detection methods in point clouds. *arXiv preprint arXiv:2204.00106*, 2022. 4

[47] Walter Zimmer, Joseph Birkner, Marcel Brucker, Huu Tung Nguyen, Stefan Petrovski, Bohan Wang, and Alois C. Knoll. Infradet3d: Multi-modal 3d object detection based on roadside infrastructure camera and lidar sensors. In *2023 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2023. 4, 5, 8

[48] Walter Zimmer, Christian Creß, Huu Tung Nguyen, and Alois C Knoll. Tumtraf intersection dataset: All you need for urban 3d camera-lidar roadside perception [best student paper award]. In *2023 IEEE Intelligent Transportation Systems ITSC*. IEEE, 2023. 2, 5, 8

[49] Walter Zimmer, Jialong Wu, Xingcheng Zhou, and Alois C Knoll. Real-time and robust 3d object detection with roadside lidars. In *Proceedings of the 12th International Scientific Conference on Mobility and Transport: Mobility Innovations for Growing Megacities*, pages 199–219. Springer, 2023. 4