

DAIR-V2X: A Large-Scale Dataset for Vehicle-Infrastructure Cooperative 3D Object Detection

Haibao Yu¹, Yizhen Luo^{1,3}, Mao Shu², Yiyi Huo^{1,4}, Zebang Yang^{1,3}, Yifeng Shi², Zhenglong Guo²,

Hanyu Li², Xing Hu², Jirui Yuan¹, Zaiqing Nie^{1*}

¹Institute for AI Industry Research(AIR), Tsinghua University

²Baidu Inc. ³Department of Computer Science and Technology, Tsinghua University

⁴University of Chinese Academy of Science

{yuhaibao@air., luoyz18@mails., yzb19@mails., yuanjirui@air., zaiqing@air.}@tsinghua.edu.cn,
 {shumao, shiyifeng, guozhenglong, lihanyu02, huxing}@baidu.com, huoyiyi18@mails.ucas.ac.cn

Abstract

Autonomous driving faces great safety challenges for a lack of global perspective and the limitation of long-range perception capabilities. It has been widely agreed that vehicle-infrastructure cooperation is required to achieve Level 5 autonomy. However, there is still NO dataset from real scenarios available for computer vision researchers to work on vehicle-infrastructure cooperation-related problems. To accelerate computer vision research and innovation for Vehicle-Infrastructure Cooperative Autonomous Driving (VICAD), we release DAIR-V2X Dataset, which is the first large-scale, multi-modality, multi-view dataset from real scenarios for VICAD. DAIR-V2X comprises 71254 LiDAR frames and 71254 Camera frames, and all frames are captured from real scenes with 3D annotations. The Vehicle-Infrastructure Cooperative 3D Object Detection problem (VIC3D) is introduced, formulating the problem of collaboratively locating and identifying 3D objects using sensory inputs from both vehicle and infrastructure. In addition to solving traditional 3D object detection problems, the solution of VIC3D needs to consider the temporal asynchrony problem between vehicle and infrastructure sensors and the data transmission cost between them. Furthermore, we propose Time Compensation Late Fusion (TCLF), a late fusion framework for the VIC3D task as a benchmark based on DAIR-V2X. Find data, code, and more up-to-date information at <https://thudair.baai.ac.cn/index> and <https://github.com/AIR-THU/DAIR-V2X>.

1. Introduction

Autonomous driving (AD) is arguably one of the hottest topics currently occupying public attention and imagina-

tion. The success of deep neural networks brings the promise of solving AD's core requirement to perceive the surrounding environment from point cloud [15, 20, 28], images [7, 18] or multi-modality data [21, 24]. Despite its

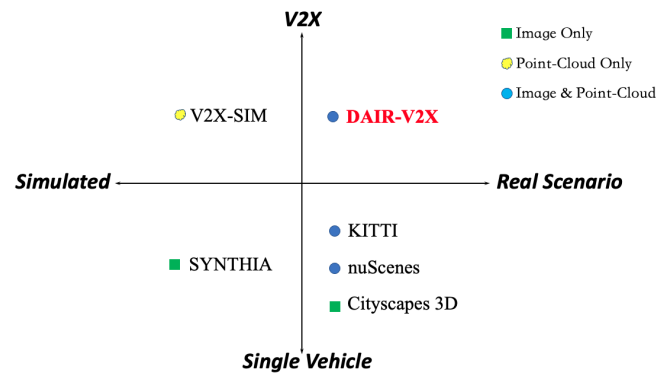


Figure 1. Datasets available for 3D Object Detection in autonomous driving. DAIR-V2X is the first real-world V2X dataset for VICAD.

great progress recently, autonomous driving still faces great safety challenges for a lack of global perspective and the limitation of long-range perception capability. It has been widely agreed that vehicle-infrastructure cooperation is required to achieve Level 5 autonomy. Utilizing both vehicle and infrastructure sensors brings a number of significant advantages, including providing a global perspective far beyond the current horizon and covering blind spots. Advances in communications like V2X (vehicle to everything) have made it possible to utilize data from infrastructure sensors [3, 22]. However, there is still NO dataset from real scenarios available for researchers to work on vehicle-infrastructure cooperation-related problems.

To accelerate computer vision research and innovation

*Corresponding author. ^{3,4} Work done while at AIR.

Table 1. A detailed comparison between autonomous driving-related datasets. - indicates that specific information is not provided. In particular, DAIR-V2X is composed of DAIR-V2X-C, DAIR-V2X-V and DAIR-V2X-I, where DAIR-V2X-C is captured by both vehicle and infrastructure sensors, DAIR-V2X-V is captured by vehicle sensors, and DAIR-V2X-I is captured by infrastructure sensors.

Dataset	Year	Real/Simulated	View	Image	Pointcloud	3D boxes	Classes
KITTI [10]	2012	real	single vehicle	15k	15k	200k	8
nuScenes [2]	2019	real	single vehicle	1.4M	400k	1.4M	23
Waymo Open [23]	2019	real	single vehicle	1M	200k	12M	4
ApolloScape [12]	2018	real	single vehicle	144k	0	70k	8-35
BBD100K [30]	2020	real	single vehicle	100M	0	0	10
ONCE [17]	2021	real	single vehicle	7M	1M	417k	5
SYNTHIA [19]	2016	simulated	single vehicle	213k	0	-	13
V2X-Sim [16]	2021	simulated	multi-vehicle	0	10k	26.6k	2
highD [13]	2018	real	infrastructure (UAV)	1.53M	0	0	1
DAIR-V2X (Our)	2021	real	vehicle-infrastructure cooperative	71k	71k	1.2M	10
- DAIR-V2X-C	2021	real	vehicle-infrastructure cooperative	39k	39k	464k	10
- DAIR-V2X-V	2021	real	single vehicle	22k	22k	239k	10
- DAIR-V2X-I	2021	real	infrastructure	10k	10k	493k	10

for Vehicle-Infrastructure Cooperative Autonomous Driving (VICAD), we release DAIR-V2X Dataset, which is the first large-scale, multi-modality, multi-view dataset for VICAD. It contains 71254 LiDAR frames and 71254 Camera frames captured in intersection scenes where a well-equipped vehicle passes through intersections with infrastructure sensors deployed. 40% of the frames are captured from infrastructure sensors and 60% of the frames are captured from vehicle sensors. All of them are precisely labeled by expert annotators. The dataset covers 10 km of city roads, 10 km of highway, 28 intersections, and 38 km² of driving regions with diverse weather and lighting variations. More details could be found in Tab. 1.

In this paper, the Vehicle-Infrastructure Cooperative 3D Object Detection (VIC3D) task is introduced, formulating the problem of cooperatively locating and identifying 3D objects using sensory inputs from both vehicle and infrastructure. In addition to solving traditional 3D object detection problems, the solution of VIC3D needs to consider the temporal asynchrony problem and data transmission cost between vehicle and infrastructure sensors.

To resolve the VIC3D object detection task and facilitate future research, we also introduce our VIC3D object detection benchmark in this paper. For data with less temporal asynchrony problems, we implement both early fusion and late fusion approaches. Results show that the average precision of fusion methods is 10 to 20 points higher than detectors that only use information from a single view. Results also show that early fusion can achieve better performance than late fusion but requires more data transmission. With the DAIR-V2X dataset, we expect more future research to achieve a performance-bandwidth trade-off. For data with severe temporal asynchrony, we propose a Time Compensation Late Fusion framework, which can effectively alleviate the temporal asynchrony problem.

The key contributions of our work are as follows:

- We release the DAIR-V2X dataset, which is the first large-scale dataset for vehicle-infrastructure cooperative autonomous driving. All frames are captured from real scenarios with 3D annotations.
- We formulate the problem of cooperatively locating and identifying 3D objects using sensory inputs from both vehicle and infrastructure as VIC3D.
- We introduce benchmarks for VIC3D object detection and single-view 3D object detection tasks. The results show the effectiveness of vehicle-infrastructure cooperation in VIC3D object detection. Especially, we propose the Time Compensation Late Fusion framework to alleviate the temporal asynchrony problem.

2. Relative Work

2.1. Autonomous Driving Datasets

In recent years, an increasing number of autonomous driving datasets have been released and greatly promoted the development of autonomous driving research. Datasets like SYNTHIA [19] and Cityscapes [5] mainly focus on 2D annotations for images. KITTI [10] and nuScenes [2] are multi-modality datasets providing camera images as well as LiDAR point clouds. Nevertheless, all datasets mentioned above only provide data from a single-vehicle view. V2X-SIM [16] is an attempt to generate a multi-vehicle view dataset, but the dataset was generated by a simulator rather than captured from real scenarios. Compared with those datasets, our DAIR-V2X dataset is the first large-scale, multi-modality, multi-view dataset captured from real scenarios for VICAD, and contains data captured from the Vehicle-Infrastructure Cooperative view. Tab. 1 shows the comparison of our dataset with the others. In our DAIR-

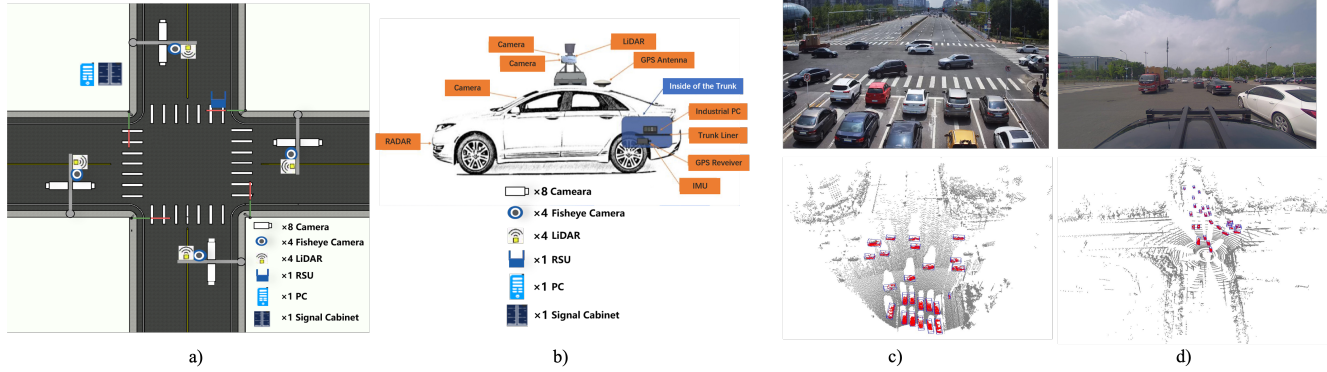


Figure 2. a) Acquisition system with infrastructure sensors. b) Acquisition system with vehicle sensors. c) Infrastructure-view image and point cloud with 3d annotation. Paired vehicle-view and infrastructure-view information complement each other in the perspective of view. d) Vehicle-view image and point cloud with 3d annotation.

V2X, we also provide a Repo3D [29] dataset composed of multi-source infrastructure images and 3D annotations, for those who are interested in Mono3D object detection and domain adaptation.

2.2. 3D Detection

3D object detection serves as the prerequisite for the success of autonomous driving. Many techniques have been introduced and can be roughly classified into three categories. a) Image-based 3D Detection refers to methods that detect 3D objects directly from 2D images. ImVoxelNet [7] is a good example to make predictions from images. b) Pointcloud-based 3D Detection stands for manners that make 3D object detection merely from point clouds. PointPillars [15], SECOND [27], and 3DSSD [28] are such approaches that achieve convincing detection results from point clouds. c) Multimodality-based 3D Detection uses both images and point clouds to make predictions. Pointpainting [24] and MVXNet [21] are practices of fusing image and LiDAR features to predict 3D bounding boxes. While 3D object detection has made great progress recently, there are still some tough problems that remain to solve such as blind spots and weak long-distance perception. To explore how to utilize the infrastructure information to solve the problems mentioned above, we conduct VIC3D object detection based on our dataset proposed in this paper.

2.3. Multi-Sensor Fusion

Multi-sensor fusion [26] is the integration of heterogeneous information collected by different sensors to alleviate the uncertainty and vulnerability of systems that rely on a single sensor. Based on the fusion stage, multi-sensor fusion can be categorized into early fusion, intermediate fusion, and late fusion. a) In early fusion, raw data from different sensors are directly transferred and fused [9]. b) In intermediate fusion, intermediate representations like features extracted from the models are fused [4,21]. c) In late fusion,

the prediction outputs like 3D information of the objects are fused [11]. VIC3D can be considered as a variant of the multi-sensor problem, so previous fusion methods can be taken into consideration to integrate the infrastructure information. However, in addition to the multi-sensor fusion challenges, VIC3D faces difficulties caused by the temporal asynchrony problem and the data transmission constraint.

2.4. V2X Cooperative Perception

V2X aims to build a communication system between vehicles and other devices in a complex traffic environment. Current V2X research mainly focuses on V2V (Vehicle-to-Vehicle) and V2I (Vehicle-to-Infrastructure) area. V2VNet [25] is a pioneering work in V2V that broadcasts compressed intermediate features and propagates message received from nearby vehicles to generate motion forecasts. Works of V2I [6,31] leverage infrastructure LiDAR data to generate and broadcast detection results. However, none of these approaches have been verified on a dataset captured from real scenarios. This may cause a huge gap between theory and practice. Therefore, we release the DAIR-V2X dataset to boost further study in this field.

3. The DAIR-V2X Dataset

In order to facilitate research on VICAD, we release DAIR-V2X, a large-scale, multi-modality, multi-view dataset from real scenarios with 3D annotations for vehicle infrastructure cooperation. Here we describe how we set up infrastructure and vehicle sensors, select interesting scenes, annotate the dataset and protect the privacy of third parties.

3.1. Setup

Equipment. Equipment for data collection are composed of infrastructure sensors and vehicle sensors. a) Infrastructure sensors. Each of the 28 intersections selected from Beijing High-level Autonomous Driving Demonstration

Table 2. Key Sensor Specifications in DAIR-V2X. Veh. stands for vehicle view, and Inf. stands for infrastructure view.

Sensor	Details
Inf. LiDAR	300 beams, 10Hz capture frequency, 100° horizontal FOV, -30° to 10° vertical FOV, $\leq 280m$ range, $\pm 3cm$ accuracy
Inf. Camera	RGB, 25Hz capture frequency, 1920x1080 resolution, JPEG compressed
Veh. LiDAR	40 beams, 10Hz capture frequency, 360° horizontal FOV, -30° to 10° vertical FOV, $\leq 200m$ range, $\pm 0.33^\circ$ vertical resolution
Veh. Camera	RGB, 20Hz capture frequency, 1920x1080 resolution, JPEG compressed
Veh. GPS & IMU	1000HZ update rate

Area are deployed with four pairs of 300-beam LiDAR and high-resolution camera. The DAIR-V2X dataset picks only one pair of them. b) Vehicle sensors. One 40-beam LiDAR and one high-quality camera looking forward are mounted on top of the autonomous vehicles. Specific layout is posted in Figure 2, and precise details are displayed in Table 2.

Coordinate. There are 5 types of coordinate systems on DAIR-V2X, *i.e.*, the LiDAR coordinate, the camera coordinate, the image coordinate, the world coordinate, and the positioning coordinate. The origin of the LiDAR coordinate system is located at the center of the LiDAR sensor, the x-axis is positive forwards, the y-axis is positive to the left, and the z-axis is positive upwards. The infrastructure LiDAR coordinate system is converted from its original system which has an inclination angle with the ground. The real-time relative pose of the equipped vehicle is obtained from GPS/IMU combined with SLAM and a local map. There is also manual secondary labeling confirmation to ensure calibration accuracy. The Lidar-to-Camera transformation is obtained by multiplying Lidar-to-World and World-to-Camera transformations.

3.2. Data Acquisition

Collection. We drive a well-equipped vehicle in the collection area and save the corresponding vehicle frames and infrastructure frames respectively. After the collection of raw data, we manually select 100 representative scenes of 20s duration. Such scenes include vehicle data and infrastructure data, where vehicles drive through intersections deployed with equipment. We sample key frames at 10Hz from both sides to form DAIR-V2X-C. In DAIR-V2X-C, it is important to note that the timestamp difference between a vehicle frame and its closest infrastructure frame could be slightly varied, due to the asynchronous triggering between vehicle sensors and infrastructure sensors. We sample 22K frames from additionally about 350 vehicle-only segments of 60s duration to form DAIR-V2X-V, and sam-

ple 10K frames from additionally about 150 infrastructure-only segments duration to form DAIR-V2X-I, to enlarge the dataset. Compared to the single-view data in DAIR-V2X-C, DAIR-V2X-V and DAIR-V2X-I contains more diverse scenes and will be more challenging to only improve the single-view performance.

Annotation. With multiple validation steps and refinement processes, expert annotators make high-quality annotations for infrastructure frames and vehicle frames respectively. Specifically, annotators exhaustively label each of the 10 object classes in every image and point cloud frame with its category attribute, occlusion state, truncated state, and a 7-dimensional cuboid modeled as $x, y, z, width, length, height$, and $yaw\ angle$. 10 categories include different vehicles, pedestrian, different cyclists. Moreover, experts also meticulously annotate objects in camera images with a rectangle bounding box modeled as $x, y, width$, and $length$.

To be mentioned, we also conduct semi-automatic labeling for the cooperative annotations with vehicle and infrastructure frame pairs. We first select vehicle and infrastructure frame pairs from DAIR-V2X-C. The timestamp differences between the two frames of the selected pairs are less than $10ms$ (We call it the Synchronous Case which is defined in Section 4.1. To obtain more cooperative annotations, we extend the threshold from $10ms$ to $30ms$). Next, we convert infrastructure 3D boxes into vehicle LiDAR coordinate system and fuse the vehicle annotations and infrastructure annotations. For each 3D box in the infrastructure annotation, if we can not find any 3D box in the vehicle annotation that has the same location and category, we add the infrastructure 3D box into the vehicle annotations; in this way, we get the vehicle-infrastructure cooperative annotations. We manually supervise and adjust the cooperative annotations to generate more accurate annotations. Here we take 9331 infrastructure frames and vehicle frames as well as the cooperative annotations to form the VIC-Sync dataset for our VIC3D object detection benchmark.

Protection. The whole dataset is desensitized before public release. Complied with local laws and regulations, we erase all localization information, including road name, map data, and positioning information, to make sure our dataset meets requirements. In addition, we utilize professional labeling tools to blur all the information suspected of privacy violation, including road signs, license plates and faces, to protect privacy and avoid violating personal rights.

4. Task & Metrics

Autonomous driving faces great safety challenges for a lack of global perspective and the limitation of long-range perception capabilities. Since 3D object detection is one of the key perception tasks in autonomous driving, in this paper, we focus on the vehicle-infrastructure cooperative (VIC) 3D object detection task, the vehicle receives and

integrates information from infrastructure to localize and recognize objects surrounding itself. Compared with traditional multi-sensor 3D object detection tasks, VIC3D object detection has the following alternative characteristics:

- **Transmission Cost.** Limited by physical communication conditions, fewer data should be transmitted from infrastructure to reduce bandwidth consumption, alleviate time delay, and satisfy real-time requirements. Thus, the solution to VIC3D object detection needs to balance the trade-off between the performance and the transmission cost.
- **Temporal Asynchrony.** Timestamps of data from the vehicle sensors and the infrastructure sensors are different due to the asynchronous triggering and time delay caused by transmission cost, to generate the temporal-spatial error. Therefore, temporal synchronization should be considered in solving VIC3D.

To better formulate the VIC3D object detection task, we will give a detailed definition to the VIC3D object detection and then provide two metrics to measure detection performance and transmission cost in this section.

4.1. VIC3D Object Detection

VIC3D object detection can be formulated as the optimization problem of effectively integrating infrastructure and vehicle information to localize and recognize 3D objects considering transmission cost. Here we discuss what the input and output of VIC3D should be.

Input. The input of VIC3D is composed of data from the vehicle and the infrastructure.

- **Vehicle Frame $I_v(t_v)$:** captured at time t_v as well as its relative pose $M_v(t_v)$, where $I_v(\cdot)$ denotes the capturing function of vehicle sensors.
- **Infrastructure Frame $I_i(t_i)$:** captured at time t_i as well as its relative pose $M_v(t_i)$, where $I_i(\cdot)$ denotes the capturing function of infrastructure sensors.

Note that t_i should be earlier than t_v because there is a time delay caused by data transmission from the infrastructure to the vehicle. Considering that the objects would move so slightly in the tiny time interval that the spatial offset can be ignored, we take the case that $|t_v - t_i| \leq 10ms$ as Synchronous Case (i.e. $t_v \approx t_i$). Similarly, we take the case that $|t_v - t_i| > 10ms$ as Asynchronous Case. In addition, we allow using more infrastructure frames previous to $I_i(t_i)$ in solving VIC3D to make full use of the infrastructure computing resources.

Ground Truth. The outputs of VIC3D object detection contain 3D information like the location, category, and orientation of objects surrounding the vehicle. The corresponding ground truth of VIC3D is the fusion result of infrastructure and vehicle ground truth, which could be for-

mulated as:

$$GT = GT_v \cup GT_i, \quad (1)$$

where GT_v is the ground truth for vehicle sensor perception and GT_i is the ground truth for infrastructure sensor perception.

VIC3D is mainly used to improve the perception performance of the self-driving vehicle. We are more concerned about a certain range of egocentric surroundings and the 3D information of objects at time t_v than at t_i . Therefore, GT_v and GT_i should both be based on time t_v . However, the timestamp of the input frames captured from the infrastructure and captured from the vehicle could be different that $t_v \neq t_i$. This not only brings challenges to fusing the infrastructure information in model prediction but also creates huge problems to generate the ground truth. That's because objects annotated with infrastructure frame at time t_i may move to different locations at time t_v , and we cannot directly get the infrastructure frame at time t_v to annotate.

In response to these difficulties, we discuss how we generate the ground truth for VIC3D based on DAIR-V2X.

- **Synchronous Case (i.e. $t_v \approx t_i$).** Under this condition, an object that appears in vehicle frame $I_v(t_v)$ should have the same spatial location as it appears in infrastructure frame $I_i(t_i)$. Therefore, we can directly take the vehicle-infrastructure cooperative 3D annotations obtained by semi-automatic labeling illustrated in Section 3.2 as ground truth.
- **Asynchronous Case (i.e. $t_v \neq t_i$).** If we can find such infrastructure frame $I_i(t'_i)$ satisfying $|t_v - t'_i| \leq 10ms$, we can generate ground truth with $I_i(t'_i)$. If not, we have to estimate the 3D states of objects at t_v to generate ground truth. This work can be carried out based on the tracking ID and kinematic equation after we provide the tracking ID in future work.

4.2. Evaluation Metrics.

VIC3D object detection has two major goals: better detection performance and less transmission cost. We describe the metrics for such two goals below.

Average Precision. *AP* (Average precision) is a popular metric for measuring the object detectors performance [8]. We also use *AP* to evaluate the 3d detection performance with cooperative annotations as ground truth. Since we are more concerned about egocentric surroundings, we remove objects outside the designed area. Here we set the designed area as a rectangular area as $[0, -39.12, 100, 39.12]$.

Transmission Cost. We use *AB* (Average Byte) to measure the transmission cost. Here Byte is a unit of digital information that consists of eight *bits*. To simplify the problem, we ignore the time consumption of data encoders and decoders during transmission. That means the less transmission cost, the less time delay. Data to be transmitted from the infrastructure can be one or a combination of the following forms.

Table 3. VIC3D object detection Benchmark on DAIR-V2X-C.

Modality	Fusion	Model	Dataset	$AP_{3D}^{IoU=0.5}$				$AP_{BEV}^{IoU=0.5}$			AB (Byte)	
				Overall	0-30m	30-50m	50-100m	Overall	0-30m	30-50m		50-100m
Image	Veh.-Only	ImvoxelNet [7]	VIC-Sync	12.03	16.25	7.25	2.28	13.62	17.66	8.58	2.82	0
	Inf.-Only	ImvoxelNet [7]	VIC-Sync	19.93	27.34	17.61	14.43	25.31	32.02	23.28	20.38	102.32
	Late Fusion	ImvoxelNet [7]	VIC-Sync	26.56	34.20	17.20	9.81	31.40	37.75	21.21	12.99	102.32
Pointcloud	Veh.-Only	PointPillars [15]	VIC-Sync	31.33	27.48	25.58	12.63	35.06	30.55	28.65	14.16	0
	Inf.-Only	PointPillars [15]	VIC-Sync	17.62	16.54	10.98	9.17	24.40	21.47	16.00	13.07	336.16
	Late Fusion	PointPillars [15]	VIC-Sync	41.90	37.65	32.72	18.84	47.96	42.40	37.65	22.08	336.16
	Early Fusion	PointPillars [15]	VIC-Sync	50.03	53.07	60.38	33.05	53.73	55.80	64.08	36.17	1382275.75
Pointcloud	Late Fusion	PointPillars [15]	VIC-Async-1	40.21	34.17	29.40	15.50	46.41	38.05	34.10	19.20	341.08
	Late Fusion	PointPillars [15]	VIC-Async-2	35.29	32.16	28.07	13.44	40.65	35.62	32.35	15.88	306.79
	Early Fusion	PointPillars [15]	VIC-Async-1	47.47	48.88	58.86	30.89	51.67	52.70	63.09	34.72	1362216.0
Pointcloud	TCLF	PointPillars [15]	VIC-Async-1	40.79	34.67	29.69	15.76	46.80	38.24	34.27	19.40	539.60
	TCLF	PointPillars [15]	VIC-Async-2	36.72	33.91	29.41	14.52	41.67	36.78	33.36	17.18	506.70

- Raw data such as images or point clouds contains complete information but requires much transmission cost.
- Intermediate representation requires less transmission cost while retaining valuable information, which may achieve a better performance-transmission trade-off. Surely, this requires a more sophisticated design to extract suitable intermediate representation.
- Object-level outputs directly provide 3D object information. Although it is transmission-efficient, it may lose valuable information.
- Other auxiliary information like scene flows help to alleviate temporal asynchrony problems.

5. Benchmark

In this section, we provide a VIC3D object detection benchmark and a Single-View (SV) 3D object detection benchmark on our DAIR-V2X dataset, analyze their characteristics and suggest avenues for future research.

5.1. Benchmark for VIC3D object detection

We provide a benchmark for VIC3D object detection on the VIC-Sync dataset extracted from DAIR-V2X-C, which is illustrated in Section 3.2. The dataset is composed of 9311 pairs of infrastructure and vehicle frames as well as their cooperative annotations as ground truth. Besides, we take the temporal asynchrony between the infrastructure frame and the vehicle frame into consideration in the benchmark, which is mainly caused by the difference in the sampling rate and transmission delay. To simulate the temporal asynchrony phenomenon, we replace each infrastructure frame in the VIC-Sync dataset with the infrastructure frame which is k -th frame previous to the original infrastructure frame to construct the VIC-Async- k dataset for the benchmark. In our experiments, we set $k = 1, 2$. We split VIC-Sync and VIC-Async- k datasets to train/valid/test part as 5:2:3 respectively. We use cooperative annotations to evaluate the detection results under the vehicle-egocentric view. The experiment results are presented in Table 3.

5.1.1 Baselines

Here we present several baselines with different modalities and fusion methods for VIC3D object detection.

LiDAR detection baseline with Late Fusion. To demonstrate the performance improvement by utilizing both infrastructure and vehicle data, we implement a late fusion framework with an infrastructure detector and a vehicle detector. Firstly, we choose PointPillars [15] as the 3D detector and train the two detectors with infrastructure-view and vehicle-view data in VIC-Sync separately. Then, we convert the infrastructure predictions into the vehicle LiDAR coordinate system and merge the prediction results with a matcher based on the Euclidian distance measurement and the Hungarian method [14] to generate fusion results.

To illustrate the temporal asynchrony problem, we also implement the LiDAR detection late fusion baseline on the VIC-Async- k dataset. In addition, based on tracking and state estimation we propose the Time Compensation Late Fusion (TCLF) framework. The TCLF is mainly composed of the following three parts: 1) Estimating the velocity of the objects with two adjacent infrastructure frames. 2) Estimating the state of the infrastructure objects at t_v . 3) Fusing the estimated infrastructure predictions and vehicle predictions following the way of LiDAR late fusion baseline. The details of the TCLF framework could be seen in Fig. 3.

Note that we also report the evaluation results only with the infrastructure data and only with the vehicle data, which are named as Veh.-Only and Inf.-Only respectively. The evaluation results are presented in Tab. 3.

Image detection baseline with Late Fusion. To examine image-only VIC3D object detection, we also implement the late fusion framework only with infrastructure images and vehicle images. We choose ImvoxelNet [7] as the 3D detector and train infrastructure detector and vehicle detector with the corresponding part of VIC-Sync training data separately. We implement the image detection late fusion following the LiDAR detection late fusion.

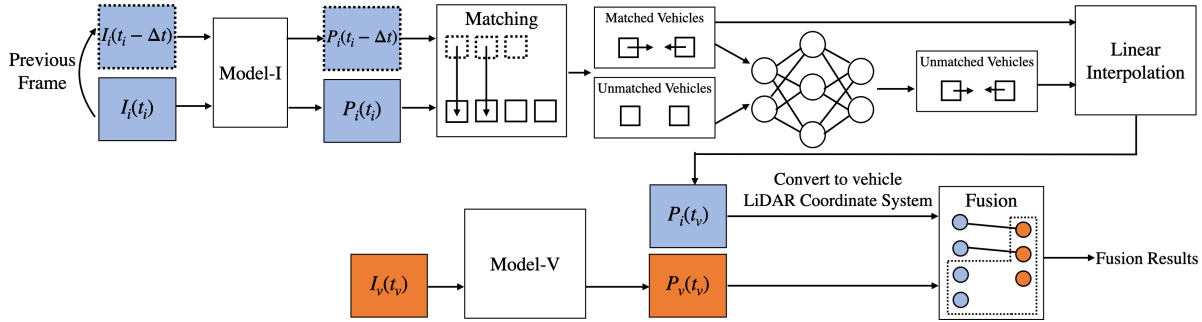


Figure 3. Time Compensation Late Fusion (TCLF) Framework. Δt denotes the sampling interval of infrastructure sensors. We predict and match the boxes between two infrastructure frames. For matched vehicles, we compute their velocities directly. For unmatched vehicles, we feed the position and motion information of the current scene into an MLP to predict their velocities. Finally, we can approximate the positions of vehicles at t_v by linear interpolation, and fuse the results of the vehicle frame.

LiDAR detection baseline with Early Fusion. To explore the fusion effect at the raw data level, we implement the early fusion with PointPillars [15] as the 3D detector on the VIC-Sync dataset. We first convert the infrastructure point cloud in the VIC-Sync dataset into the vehicle LiDAR coordinate system, then fuse the infrastructure point cloud and vehicle point cloud. We directly train and evaluate the detector with the fused point cloud. Further to illustrate the temporal asynchrony problem, we also implement the early fusion with PointPillars [15] on the VIC-Async- k dataset.

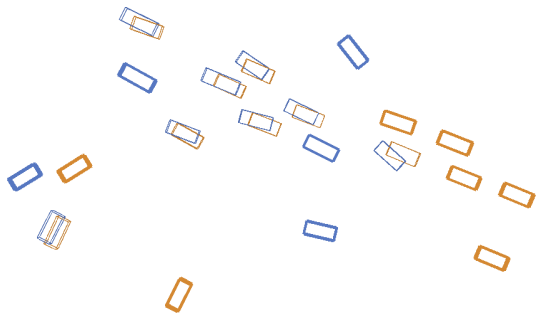


Figure 4. Prediction results of the vehicle frame (orange) and infrastructure frame (blue). We observe that infrastructure data (thick blue boxes) supplements the blind spot and extends the perception field for the vehicle.

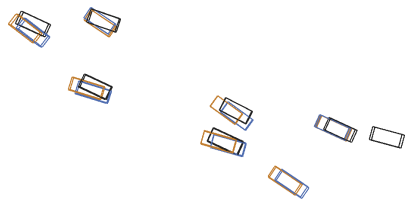


Figure 5. Prediction results with and without time compensation. The results of TCLF (blue) have a larger overlap with ground truth (black) than the results without time compensation (orange).

5.1.2 Analysis

Here we analyze the properties of the methods for the VIC3D object detection benchmark in Section 5.1.1.

Cooperative-view vs. Single-view. We compare the performance of the methods whether using both infrastructure data and vehicle data. In Tab. 3, the detection performance of late fusion is much better than the performance of Veh.-Only or Inf.-Only, whether it is Image-based or LiDAR-based or it is based on VIC-Sync dataset or VIC-Async- k dataset. For example, the LiDAR detection with Late Fusion achieves overall 41.90 AP points for 3D detection and overall 47.96 AP points for BEV detection on the VIC-Sync dataset. However, the LiDAR detection only with vehicle data just achieves overall 31.33% AP for 3D detection and overall 35.06% AP for BEV detection, and the LiDAR detection only with infrastructure data just achieves overall 17.62% AP for 3D detection and overall 24.40% AP for BEV detection. The experiment results demonstrate that fusing the infrastructure information can effectively improve the perception performance of the vehicle. This is mainly because infrastructure data provides supplementary information that makes up for the vehicle's perception field. A visualization example is shown in Fig. 4.

Temporal Asynchrony vs Time Compensation. Temporal asynchrony brings challenges to fusing the infrastructure data. Compared with the results on the VIC-Sync dataset, the performance of LiDAR detection with fusion drops significantly on VIC-Async- k (2 points on VIC-Async-1 and 6 points on VIC-Async-2). The decline is mainly due to the state changes of moving objects, resulting in matching difficulties and fusion errors. However, our TCLF can effectively improve the performance of late fusion up to 0.5% AP and 1.5% AP on VIC-Async-1 and VIC-Async-2 respectively, which demonstrates that time compensation can effectively alleviate the temporal asynchrony problems es-

Table 4. SV3D Detection Benchmark on DAIR-V2X-V

Modality	Model	Vehicle _{3D(IoU=0.5)}			Pedestrian _{3D(IoU=0.25)}			Cyclist _{3D(IoU=0.25)}		
		Easy	Middle	Hard	Easy	Middle	Hard	Easy	Middle	Hard
Image	ImvoxelNet [7]	38.37	24.28	21.54	4.54	4.54	4.54	10.38	9.09	9.09
PointCloud	PointPillars [15]	61.76	49.02	43.45	33.40	24.68	22.39	38.24	33.80	32.35
PointCloud	SECOND [27]	69.44	59.63	57.63	43.45	39.06	38.78	44.21	39.49	37.74
Image+PointCloud	MVXNet [21]	69.86	60.74	59.31	47.73	43.37	42.49	45.68	41.84	40.55

Table 5. SV3D Detection Benchmark on DAIR-V2X-I

Modality	Model	Vehicle _{3D(IoU=0.5)}			Pedestrian _{3D(IoU=0.25)}			Cyclist _{3D(IoU=0.25)}		
		Easy	Middle	Hard	Easy	Middle	Hard	Easy	Middle	Hard
Image	ImvoxelNet [7]	44.78	37.58	37.55	6.81	6.746	6.73	21.06	13.57	13.17
PointCloud	PointPillars [15]	63.07	54.00	54.01	38.53	37.20	37.28	38.46	22.60	22.49
PointCloud	SECOND [27]	71.47	53.99	54.00	55.16	52.49	52.52	54.68	31.05	31.19
Image+PointCloud	MVXNet [21]	71.04	53.71	53.76	55.83	54.45	54.40	54.05	30.79	31.06

pecially when the time delay is larger. A visualization example is provided in Fig. 5.

Early Fusion vs. Late Fusion. Compared with late fusion, early fusion achieves up to 8% AP higher under both BEV and 3D benchmarks, whether it is based on the VIC-Sync dataset or the VIC-Async-1 dataset. However, early fusion should transmit the whole point cloud and suffers an extremely high transmission cost, which is about 4000 times more than late fusion. For more practical applications, we encourage future research on achieving better performance while consuming less transmission bandwidth. We will also release the feature fusion for the benchmark in the future.

5.2. Benchmark for SV3D Detection

We present an extensive 3D detection benchmark for those who are interested in Single-View (SV) 3D detection tasks based on DAIR-V2X-V and DAIR-V2X-I datasets. Compared with the single-side data in DAIR-V2X-C, the two datasets are more diverse and could be more challenging to implement 3D object detection. Hence, we encourage researchers who just aim at improving the performance of vehicle 3D object detection or infrastructure 3D object on DAIR-V2X-V and DAIR-V2X-I.

We split DAIR-V2X-V and DAIR-V2X-I datasets to train/valid/test part as 5:2:3 respectively. We present a number of baselines with methods based on different modalities on the two datasets respectively: ImvoxelNet [7], PointPillars [15], SECOND [27] and MVXNet [21]. We evaluate 3D object detection performance using the PASCAL criteria as KITTI [10], that distant objects are filtered out based on their bounding box height in the image plane. Three types of modes are used for evaluation, including Easy, Moderate, and Hard modes. We implement these baselines with MMDetection3D Framework [1]. Evaluation results are

shown in Tab. 4 and Tab. 5.

6. Conclusion

In this paper, we introduce DAIR-V2X, the first large-scale, multi-modality, multi-view dataset for vehicle-infrastructure cooperative autonomous driving, and all frames are captured from real scenes with 3D annotations. We also define VIC3D object detection to formulate the problem of collaboratively locating and identifying 3D objects using sensory input from both vehicle and infrastructure. In addition to solving traditional 3D object detection problems, the solution of VIC3D needs to consider the temporal asynchrony problem between vehicle and infrastructure sensors and the data transmission cost between them. To facilitate future research, we provide a VIC3D benchmark for detection models with our proposed Time Compensation Late Fusion framework, as well as extensive benchmarks for 3D detection on vehicle-view and infrastructure-view datasets. Results show that integrating data from infrastructure sensors achieves an average of 15% AP higher than single-vehicle 3D detection, and TCLF can alleviate temporal asynchrony problems.

Acknowledgements

We thank Fan Yang, Ruiwen Zhang, Wenyue Wu, and Xiao Wang from Baidu Inc. for the support in data processing. We thank Jilei Mao, Taohua Zhou, Yingjuan Tang, Zan Mao, and Zhiwen Yang for their support in the benchmark construction. Thanks to Beijing High-level Autonomous Driving Demonstration Area, Beijing Connected and Autonomous Vehicles Technology Co., Ltd, Baidu Apollo, and Beijing Academy of Artificial Intelligence for their support throughout the dataset construction and release process.

References

- [1] MMDetection3D: OpenMMLab next-generation platform for general 3D object detection, 2020. **8**
- [2] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nusenes: A multi-modal dataset for autonomous driving. In *2020 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11618–11628, 2020. **2**
- [3] Shanzhi Chen, Jinling Hu, Yan Shi, Ying Peng, Jiayi Fang, Rui Zhao, and Li Zhao. Vehicle-to-everything (v2x) services supported by lte-based systems and 5g. *IEEE Communications Standards Magazine*, 1(2):70–76, 2017. **1**
- [4] Xiaozhi Chen, Huimin Ma, Ji Wan, Bo Li, and Tian Xia. Multi-view 3d object detection network for autonomous driving. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6526–6534, 2017. **3**
- [5] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. **2**
- [6] Yuepeng Cui, Hao Xu, Jianqing Wu, Yuan Sun, and Junxuan Zhao. Automatic vehicle tracking with roadside lidar data for the connected-vehicles system. *IEEE Intelligent Systems*, 34(3):44–51, 2019. **3**
- [7] Anton Konushin Danila Rukhovich, Anna Vorontsova. Imvoxelnet: Image to voxels projection for monocular and multi-view general-purpose 3d object detection. *arXiv preprint arXiv:2106.01178*, 2021. **1, 3, 6, 8**
- [8] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010. **5**
- [9] Hongbo Gao, Bo Cheng, Jianqiang Wang, Keqiang Li, Jianhui Zhao, and Deyi Li. Object classification using cnn-based fusion of vision and lidar in autonomous vehicle environment. *IEEE Transactions on Industrial Informatics*, 14(9):4224–4231, 2018. **3**
- [10] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3354–3361, 2012. **2, 8**
- [11] Mohammad-Hashem Haghbayan, Fahimeh Farahnakian, Jonne Poikonen, Markus Laurinen, Paavo Nevalainen, Juha Plosila, and Jukka Heikkonen. An efficient multi-sensor fusion approach for object detection in maritime environments. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pages 2163–2170, 2018. **3**
- [12] Xinyu Huang, Peng Wang, Xinjing Cheng, Dingfu Zhou, Qichuan Geng, and Ruigang Yang. The apolloscape open dataset for autonomous driving and its application. *IEEE transactions on pattern analysis and machine intelligence*, 42(10):2702–2719, 2019. **2**
- [13] Robert Krajewski, Julian Bock, Laurent Kloeker, and Lutz Eckstein. The highd dataset: A drone dataset of naturalistic vehicle trajectories on german highways for validation of highly automated driving systems. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pages 2118–2125. IEEE, 2018. **2**
- [14] Harold W. Kuhn. The hungarian method for the assignment problem. In *50 Years of Integer Programming*, 2010. **6**
- [15] Alex H. Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. Pointpillars: Fast encoders for object detection from point clouds. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019. **1, 3, 6, 7, 8**
- [16] Yiming Li, Shunli Ren, Pengxiang Wu, Siheng Chen, Chen Feng, and Wenjun Zhang. Learning distilled collaboration graph for multi-agent perception. *arXiv preprint arXiv:2111.00643*, 2021. **2**
- [17] Jiageng Mao, Minzhe Niu, Chenhan Jiang, Hanxue Liang, Jingheng Chen, Xiaodan Liang, Yamin Li, Chaoqiang Ye, Wei Zhang, Zhenguo Li, et al. One million scenes for autonomous driving: Once dataset. *arXiv preprint arXiv:2106.11037*, 2021. **2**
- [18] Cody Reading, Ali Harakeh, Julia Chae, and Steven L Waslander. Categorical depth distribution network for monocular 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8555–8564, 2021. **1**
- [19] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M Lopez. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3234–3243, 2016. **2**
- [20] Shaoshuai Shi, Xiaogang Wang, and Hongsheng Li. Pointcnn: 3d object proposal generation and detection from point cloud. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 770–779, 2019. **1**
- [21] Vishwanath A. Sindagi, Yin Zhou, and Oncel Tuzel. Mvxnet: Multimodal voxelnet for 3d object detection. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 7276–7282, 2019. **1, 3, 8**
- [22] Carlos Renato Storck and Fátima Duarte-Figueiredo. A 5g v2x ecosystem providing internet of vehicles. *Sensors*, 19(3):550, 2019. **1**
- [23] 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**
- [24] Sourabh Vora, Alex H. Lang, Bassam Helou, and Oscar Beijbom. Pointpainting: Sequential fusion for 3d object detection. In *2020 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4603–4611, 2020. **1, 3**
- [25] Tsun-Hsuan Wang, Sivabalan Manivasagam, Ming Liang, Binh Yang, Wenyuan Zeng, James Tu, and Raquel Urtasun. V2vnet: Vehicle-to-vehicle communication for joint perception and prediction. In *ECCV*, 2020. **3**
- [26] Zhangjing Wang, Yu Wu, and Qingqing Niu. Multi-sensor fusion in automated driving: A survey. *IEEE Access*, 8:2847–2868, 2020. **3**

- [27] Yan Yan, Yuxing Mao, and Bo Li. Second: Sparsely embedded convolutional detection. *Sensors*, 2018. 3, 8
- [28] Zetong Yang, Yanan Sun, Shu Liu, and Jiaya Jia. 3dssd: Point-based 3d single stage object detector. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 1, 3
- [29] 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. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022. 3
- [30] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2636–2645, 2020. 2
- [31] Junxuan Zhao, Hao Xu, Hongchao Liu, Jianqing Wu, Yichen Zheng, and Dayong Wu. Detection and tracking of pedestrians and vehicles using roadside lidar sensors. *Transportation Research Part C: Emerging Technologies*, 100:68–87, 2019. 3