同行专家业内评价意见书编号：20250854322

附件1

浙江工程师学院（浙江大学工程师学院）
同行专家业内评价意见书

姓名：祝贺

学号：22260013

申报工程师职称专业类别（领域）：电子信息

浙江工程师学院（浙江大学工程师学院）制

2025年03月10日

填表说明

一、本报告中相关的技术或数据如涉及知识产权保护、军工项目保密等内容，请作脱密处理。

二、请用宋体小四字号撰写本报告，可另行附页或增加页数，A4纸双面打印。

三、表中所涉及的签名都必须用蓝、黑色墨水笔，亲笔签名或签字章，不可以打印代替。

四、同行专家业内评价意见书编号由工程师学院填写，编号规则为：年份4位＋申报工程师职称专业类别(领域)4位+流水号3位，共11位。

一、个人申报

（一）基本情况【围绕《浙江工程师学院（浙江大学工程师学院）工程类专业学位研究生工程师职称评审参考指标》，结合该专业类别(领域)工程师职称评审相关标准，举例说明】

1.对本专业基础理论知识和专业技术知识掌握情况(不少于200字)

在本专业的基础理论知识和专业技术知识方面，我具备扎实的理论基础和丰富的实践经验。在基础理论方面，我系统学习了与工程相关的数学、自然科学、人文与社会科学基础知识，深入理解了人工智能算法与系统、计算机视觉、模式识别等核心课程内容，并能够灵活运用这些理论知识解决实际问题。在专业技术知识方面，我掌握了行业前沿技术及应用方法，包括硬件设备、软件算法及其在现实中的具体应用。我熟悉本行业的技术标准、工作流程以及相关的政策制度和法律法规，能够在实际工作中严格遵循技术规范，确保工程项目的合规性和安全性。同时，我具备应用现代研究工具和专业软件进行数据分析和仿真模拟的能力，能够掌握使用传感器和设备在现场采集数据的方法，并开展相关研究与工程实践，为工程建设和技术研究提供有力支持。

2.工程实践的经历(不少于200字)

我在参与企业项目过程中，我阅读了大量书籍及文献，不断积累工程与学术知识，提高了撰写中英文文献的能力，提升了解决复杂工程问题的能力，并积极关注行业技术前沿动态及发展趋势，以保持自身知识体系的先进性。我不仅学习到本专业的知识，还学习到许多交叉的知识，具备了较强的技术应用创新和工程实践能力，能够结合企业需求，提出切实可行的技术改进方案，并推动技术成果的转化应用，为企业的技术进步和项目建设贡献力量。此外，通过此次实践，我培养了职业能力，及时与合作者沟通以发现并解决问题，提高了自己的表达与组织能力。在遇到困难时不断磨练，勇于担当，学会了保持平和的心态，抱着精益求精和追求卓越的工匠精神对待工作和学习。

3.在实际工作中综合运用所学知识解决复杂工程问题的案例（不少于1000字)

（1）针对单模态与多模态感知模型部署与集成，我们搭建了一个自动驾驶实车平台，进行数据的采集与标注，并实现了基于激光雷达、多相机和多传感器融合的3D目标检测模型的训练、测试与部署。具体来说，模型首先基于nuScenes数据集进行预训练，然后在采集构建的单车自动驾驶3D目标检测数据集上进行微调，最后对模型通过TensorRT推理引擎进行量化与优化实现加速推理并设计CUDA核函数实现并行计算，使用ROS通信框架集成从而部署到实车平台上实时运行，满足自动驾驶场景中对高精度和低时延的实际应用需求。其中，激光雷达感知采用Centerpoint模型，对其进行FP16精度量化，在NVIDIA AGX Orin的计算平台上运行帧率达到25Hz；多相机融合感知采用基于BEVDet改进的模型，对其进行FP16精度量化，在NVIDIA AGX Orin的计算平台上运行帧率达到55FPS；多模态融合采用基于BEVFusion改进的模型，对其进行INT8和FP16混合精度量化，在NVIDIA AGX Orin的计算平台上运行帧率为23FPS。

（2）针对单车感知的不足，我们搭建了一个车路协同感知系统，利用路端提供的上帝视角，辅助车端感知，提供一个更加安全、稳定和智能的自动驾驶环境感知。在实际应用场景中，考虑到车辆的高速度、实时性要求以及通信带宽的限制，路侧设备传输的是感知结果而不是原始传感器数据或特征图，并且受各种环境因素的影响，传输延迟是动态的。在中国成都的一段高速公路搭建车路感知系统并进行数据采集与标注，构建了第一个在真实场景中从路端到车端在线传输的、多模态和多视图的车路协同3D目标检测数据集OTVIC。此外，我们还提出了一种基于Transformer的新型端到端多模态后融合框架LfFormer，模型对图像和点云进行特征提取，在BEV视角下实现图像特征和点云特征的融合，再与历史时刻的 BEV

特征进行时序融合，利用过去多帧的路侧感知数据进行预测并编码与 BEV 特征进行交互，从而实现车路协同感知。实验证明了所提出的融合框架有效弥补了时延和坐标系转换误差所带来的的时间异步和空间错位，具有良好的精度和鲁棒性，同时通信带宽与速度满足实际应用需求。

（3）在3D占用网格预测任务中，由于计算复杂度高和存储开销大的问题，现有的一些方法利用稀疏特征表示进行建模实现了较为高效的预测，但检测精度较低。此外，时序融合已经通过验证可以帮助提高检测精度，增加感知的鲁棒性，但伴随而来的是引入额外的计算用于多帧信息的融合上，因此需要设计一个高效且稀疏的时序特征融合机制。在此，我们提出了一个新型方法SparseOcc4D，将环视图像从2D-to-3D提取到的3D特征转化为稀疏张量，并与历史帧的稀疏特征进行融合后，利用基于3D稀疏卷积设计的骨干网络和特征金字塔网络对3D稀疏特征进行学习，通过粗略到细化的分割头得到3D占用网格的预测结果在多个公开数据集上的实验结果表明，SparseOcc4D在效率和准确性方面均优于稠密特征建模的现有方法。相比于稀疏特征建模方法，由于引入时序融合，虽然计算开销与显存占用有所增加，但实现了更高的精度，特别是对动态场景的精确建模，展现了其在自动驾驶感知任务中的潜力。

（二）取得的业绩（代表作）【限填3项，须提交证明原件（包括发表的论文、出版的著作、专利证书、获奖证书、科技项目立项文件或合同、企业证明等）供核实，并提供复印件一份】					
1. 公开成果代表作**【论文发表、专利成果、软件著作权、标准规范与行业工法制定、著作编写、科技成果获奖、学位论文等】**					
成果名称	成果类别 [含论文、授权专利（含发明专利申请）、软件著作权、标准、工法、著作、获奖、学位论文等]	发表时间/授权或申请时间等	刊物名称/专利授权或申请号等	本人排名/总人数	备注
特征-结果级融合的车路协同感知方法、介质及电子设备	授权发明专利	2024年06月18日	专利号：ZL 202310490780.1	2/3	导师为第一作者
OTVIC: A Dataset with Online Transmission for Vehicle-to-Infrastructure Cooperative 3D Object Detection	会议论文	2024年12月25日	2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)	1/8	EI会议收录

2.其他代表作【主持或参与的课题研究项目、科技成果应用转化推广、企业技术难题解决方案、自主研发设计的产品或样机、技术报告、设计图纸、软课题研究报告、可行性研究报告、规划设计方案、施工或调试报告、工程实验、技术培训教材、推动行业发展中发挥的作用及取得的经济社会效益等】

（三）在校期间课程、专业实践训练及学位论文相关情况	
课程成绩情况	按课程学分核算的平均成绩： 90 分
专业实践训练时间及考核情况(具有三年及以上工作经历的不作要求)	累计时间： 2 年（要求1年及以上） 考核成绩： 83 分
本人承诺	
个人声明：本人上述所填资料均为真实有效，如有虚假，愿承担一切责任，特此声明！ **申报人签名：**祝贺	

22260013

二、日常表现考核评价及申报材料审核公示结果

日常表现考核评价	非定向生由德育导师考核评价、定向生由所在工作单位考核评价： ☑优秀　　□良好　　□合格　　□不合格 德育导师/定向生所在工作单位分管领导签字（公章）：2025年3月11日 中共浙江大学工程师学院委员会
申报材料审核公示	根据评审条件，工程师学院已对申报人员进行材料审核（学位课程成绩、专业实践训练时间及考核、学位论文、代表作等情况），并将符合要求的申报材料在学院网站公示不少于5个工作日，具体公示结果如下： □通过　　□不通过（具体原因：　　　　　　　　） 工程师学院教学管理办公室审核签字（公章）：　　　　　　年　月　日

浙江大学研究生院

攻读硕士学位研究生成绩表

学号：22260013	姓名：祝贺	性别：男	学院：工程师学院	专业：电子信息	学制：2.5年
毕业时最低应获：26.0学分		已获得：28.0学分		入学年月：2022-09	毕业年月：
学位证书号：			毕业证书号：		授予学位：

学习时间	课程名称	备注	学分	成绩	课程性质	学习时间	课程名称	备注	学分	成绩	课程性质
2022-2023学年秋季学期	工程技术创新前沿		1.5	84	专业学位课	2022-2023学年冬季学期	新时代中国特色社会主义理论与实践		2.0	89	专业学位课
2022-2023学年秋季学期	数值计算方法		2.0	100	专业选修课	2022-2023学年春季学期	自然辩证法概论		1.0	88	专业学位课
2022-2023学年秋季学期	人工智能算法与系统		2.0	100	专业选修课	2022-2023学年夏季学期	研究生英语基础技能		1.0	免修	公共学位课
2022-2023学年秋冬学期	工程伦理		2.0	89	专业学位课	2022-2023学年春夏学期	人工智能制造技术		3.0	92	专业学位课
2022-2023学年冬季学期	模式识别与人工智能		2.0	90	专业选修课	2022-2023学年春夏学期	高阶工程认知实践		3.0	82	专业学位课
2022-2023学年冬季学期	产业技术发展前沿		1.5	92	专业学位课	2022-2023学年夏季学期	研究生英语		2.0	免修	专业学位课
2022-2023学年秋冬学期	研究生论文写作指导		1.0	93	专业选修课		硕士生读书报告		2.0	通过	
2022-2023学年冬季学期	计算机视觉		2.0	96	专业选修课						

说明：1.研究生课程按三种方法计分：百分制，两级制（通过、不通过），五级制（优、良、中、及格、不及格）。

2.备注中“*”表示重修课程。

学院成绩校核章：

浙江大学研究生院（60）成绩校核章

成绩校核人：张梦依

打印日期：2025-03-20

证书号第7107695号

专利公告信息

发明专利证书

发 明 名 称：特征-结果级融合的车路协同感知方法、介质及电子设备

专 利 权 人：浙江大学

地　　　址：310058 浙江省杭州市西湖区余杭塘路866号

发 明 人：王越;祝贺;熊蓉

专 利 号：ZL 2023 1 0490780.1　　授权公告号：CN 116958763 B

专利申请日：2023年05月04日　　授权公告日：2024年06月18日

申请日时申请人：浙江大学

申请日时发明人：王越;祝贺;熊蓉

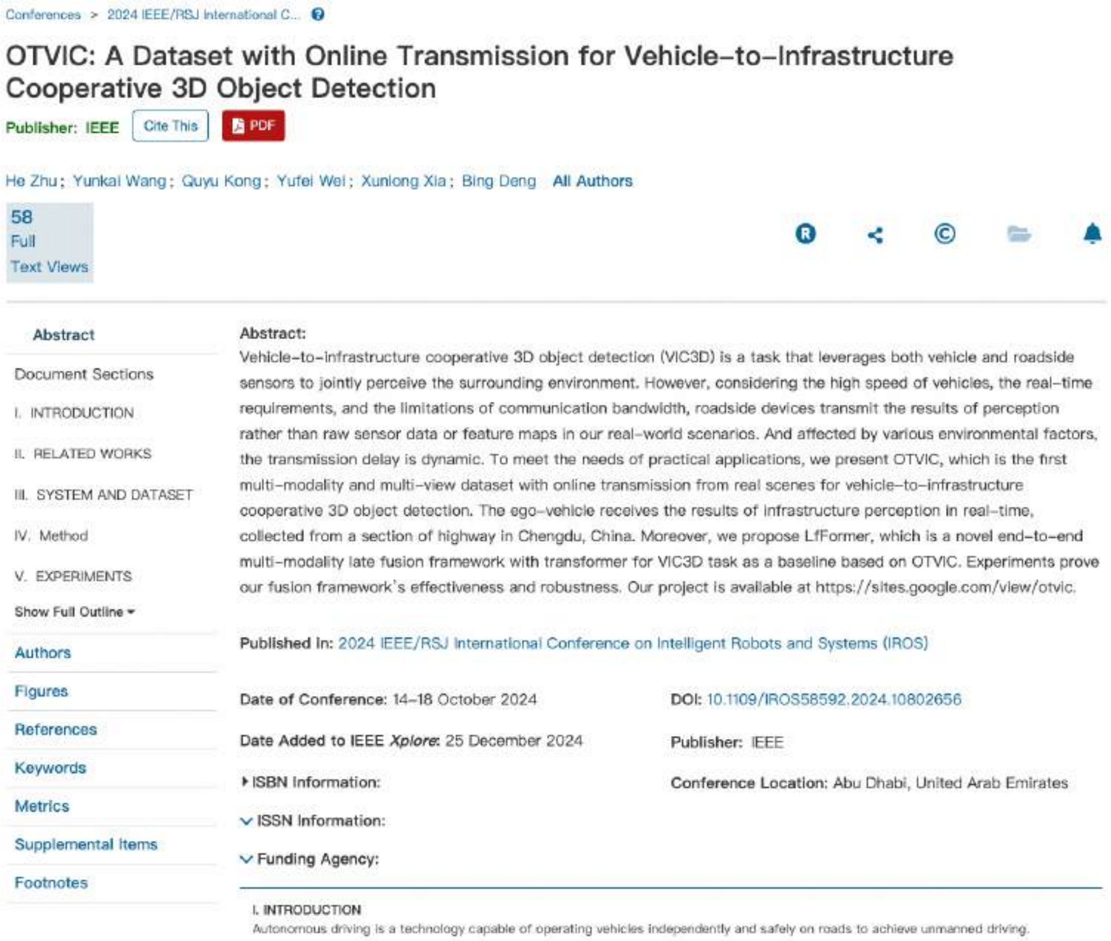
国家知识产权局依照中华人民共和国专利法进行审查，决定授予专利权，并予以公告。

专利权自授权公告之日起生效。专利权有效性及专利权人变更等法律信息以专利登记簿记载为准。

局长
申长雨

申长雨

国家知识产权局
2024年06月18日



OTVIC: A Dataset with Online Transmission for Vehicle-to-Infrastructure Cooperative 3D Object Detection

Publisher: IEEE

He Zhu; Yunkai Wang; Quyu Kong; Yufei Wei; Xunlong Xia; Bing Deng All Authors

Abstract:
Vehicle-to-infrastructure cooperative 3D object detection (VIC3D) is a task that leverages both vehicle and roadside sensors to jointly perceive the surrounding environment. However, considering the high speed of vehicles, the real-time requirements, and the limitations of communication bandwidth, roadside devices transmit the results of perception rather than raw sensor data or feature maps in our real-world scenarios. And affected by various environmental factors, the transmission delay is dynamic. To meet the needs of practical applications, we present OTVIC, which is the first multi-modality and multi-view dataset with online transmission from real scenes for vehicle-to-infrastructure cooperative 3D object detection. The ego-vehicle receives the results of infrastructure perception in real-time, collected from a section of highway in Chengdu, China. Moreover, we propose LfFormer, which is a novel end-to-end multi-modality late fusion framework with transformer for VIC3D task as a baseline based on OTVIC. Experiments prove our fusion framework's effectiveness and robustness. Our project is available at https://sites.google.com/view/otvic.

Published in: 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)

Date of Conference: 14-18 October 2024

Date Added to IEEE *Xplore*: 25 December 2024

▸ISBN Information:

⌄ISSN Information:

⌄Funding Agency:

DOI: 10.1109/IROS58592.2024.10802656

Publisher: IEEE

Conference Location: Abu Dhabi, United Arab Emirates

I. INTRODUCTION

Autonomous driving is a technology capable of operating vehicles independently and safely on roads to achieve unmanned driving.

链接：https://ieeexplore.ieee.org/abstract/document/10802656

《Ei Compendex》收录证明

经检索“Engineering Village”，下述论文被《Ei Compendex》收录。（检索时间：2025年3月5日）。

<RECORD 1>
Accession number:20250517794072
Title:OTVIC: A Dataset with Online Transmission for Vehicle-to-Infrastructure Cooperative 3D Object Detection
Authors:Zhu, He (1, 2); Wang, Yunkai (1, 2); Kong, Quyu (2); Wei, Yufei (1); Xia, Xunlong (2); Deng, Bing (2); Xiong, Rong (1); Wang, Yue (1)
Author affiliation:(1) Zhejiang University, Hangzhou, China; (2) Alibaba Cloud, Hangzhou, China
Corresponding author:Wang, Yue(wangyue@iipc.zju.edu.cn)
Source title:IEEE International Conference on Intelligent Robots and Systems
Abbreviated source title:IEEE Int Conf Intell Rob Syst
Part number:1 of 1
Issue title:2024 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2024
Issue date:2024
Publication year:2024
Pages:10732-10739
Language:English
ISSN:21530858
E-ISSN:21530866
CODEN:85RBAH
ISBN-13:9798350377705
Document type:Conference article (CA)
Conference name:2024 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2024
Conference date:October 14, 2024 - October 18, 2024
Conference location:Abu Dhabi, United arab emirates
Conference code:205527
Publisher:Institute of Electrical and Electronics Engineers Inc.
Number of references:31
Main heading:Vehicle detection
Uncontrolled terms:3D object - Communication bandwidth - High Speed - Multi-modality - Objects detection - On-line transmission - Raw sensor - Real time requirement - Surrounding environment - Vehicle-to-infrastructure
Classification code:1106.3.1 Image Processing
DOI:10.1109/IROS58592.2024.10802656
Funding text:This work was supported by the National Nature Science Foundation of China under Grant 62373322. This research was supported by Zhejiang Provincial Natural Science Foundation of China under Grant No. LD24F030001. This work was supported in part by the Alibaba Group through Alibaba Innovative Research (AIR) Program.
Database:Compendex
Compilation and indexing terms, Copyright 2025 Elsevier Inc.

注：

1. 以上检索结果来自 CALIS 查收查引系统。

2. 以上检索结果均得到委托人及被检索作者的确认。

教育部科技查新工作站（Z09）
检索人（签章）：李佳
2025年3月5日

OTVIC: A Dataset with Online Transmission for Vehicle-to-Infrastructure Cooperative 3D Object Detection

He Zhu[1,2], Yunkai Wang[1,2], Quyu Kong[2], Yufei Wei[1], Xunlong Xia[2], Bing Deng[2], Rong Xiong[1], Yue Wang[1†]

Abstract— **Vehicle-to-infrastructure cooperative 3D object detection (VIC3D) is a task that leverages both vehicle and road-side sensors to jointly perceive the surrounding environment. However, considering the high speed of vehicles, the real-time requirements, and the limitations of communication bandwidth, roadside devices transmit the results of perception rather than raw sensor data or feature maps in our real-world scenarios. And affected by various environmental factors, the transmission delay is dynamic. To meet the needs of practical applications, we present OTVIC, which is the first multi-modality and multi-view dataset with online transmission from real scenes for vehicle-to-infrastructure cooperative 3D object detection. The ego-vehicle receives the results of infrastructure perception in real-time, collected from a section of highway in Chengdu, China. Moreover, we propose LfFormer, which is a novel end-to-end multi-modality late fusion framework with transformer for VIC3D task as a baseline based on OTVIC. Experiments prove our fusion framework's effectiveness and robustness. Our project is available at https://sites.google.com/view/otvic.**

I. INTRODUCTION

Autonomous driving is a technology capable of operating vehicles independently and safely on roads to achieve unmanned driving. Currently, there are two main technological strategies: single-vehicle perception and vehicle-to-infrastructure cooperative perception [1]. Vehicle-to-infrastructure cooperative perception allows the ego-vehicle to communicate with infrastructure and improve the perceptive capability, which can solve the shortcomings of single-vehicle perception, such as the limited sight-of-view and sensor occlusion or failure [2].

Currently, most datasets for vehicle-to-infrastructure or vehicle-to-everything cooperative perception are collected from simulated environments, such as CoopInf [3], CARTI [4], V2X-Sim [5], V2XSet [6], and so on. However, in real-world scenarios, perception data from infrastructure needs to be transmitted to the vehicles in real-time. And there are three main issues in our scenario: 1) Transmission delays and inference time of the infrastructure perception algorithm can lead to temporal asynchrony. Due to the influence of environmental factors such as geographical location or weather conditions, the transmission delay changes dynamically in real scene. 2) In highway scenarios, because of the high vehicle speeds, even minimal delays can result in significant spatial misalignments. And it will lead to the problem of feature blurring, which could potentially decrease the performance of vehicle-to-infrastructure cooperative perception. 3) The roadside device provides perception for an area of $800 \times 80\ m^2$ based on 7 or 8 images with 4K resolution [7]. Due to the peak data transfer rate from RSU to OBU is 31.7 Mbps theoretically and only 15.6 Mbps in practice, it's hard to transmit raw sensor data or feature maps in real-time.

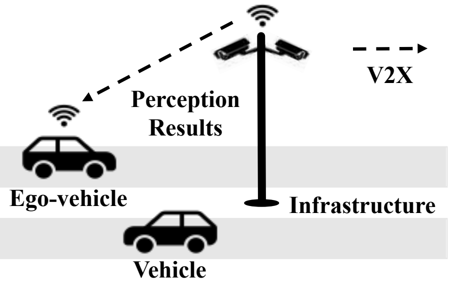


Fig. 1. A diagram illustrating vehicle-to-infrastructure scenarios. Roadside devices utilize multiple cameras for perception, transmitting structured data of perception results through Road Side Unit (RSU) online. The ego-vehicle is equipped with cameras, lidar, IMU, GPS, and other sensors. When the vehicle gets close to the roadside devices, it uses the On-Board Unit (OBU) to receive perception data from the infrastructure and fuses it with the vehicle's own sensors data to achieve vehicle-to-infrastructure cooperative perception.

In this paper, we present OTVIC, which is the first multi-modality and multi-view dataset with **O**nline **T**ransmission from real scenes for **V**ehicle-to-**I**nfrastructure **C**ooperative perception. Online transmission refers to the real-time data transfer between vehicles and infrastructure under varying communication conditions and noise levels. The purpose of the dataset is to improve the robustness and generalization performance of late fusion in challenging environments such as dynamic delay, high vehicle speeds and communication noises. A diagram illustrating vehicle-to-infrastructure scenarios is shown in Fig. 1. Each frame of the dataset contains four images captured by the ego-vehicle's cameras (including front, rear, left, and right view images), lidar point clouds, ego-vehicle localization and motion information (including the vehicle's position, velocity, acceleration, heading angle, and angular velocity), as well as the results of infrastructure perception (including the objects' type, position, heading angle, velocity, acceleration, tracking ID, and delay).

*This work was supported by the National Nature Science Foundation of China under Grant 62373322.

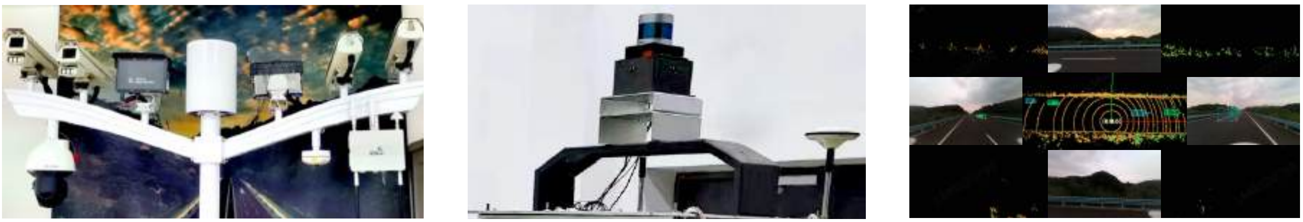
*This research was supported by Zhejiang Provincial Natural Science Foundation of China under Grant No. LD24F030001.

*This work was supported in part by the Alibaba Group through Alibaba Innovative Research (AIR) Program.

[1]Zhejiang University, Hangzhou, P.R. China.

[2]Alibaba Cloud, Hangzhou, P.R. China.

† Corresponding author, Yue Wang wangyue@iipc.zju.edu.cn



(a) Infrastructure's sensors

(b) Ego-vehicle's sensors

(c) An example from our dataset with 3D annotation

Fig. 2. Visualization of infrastructure and vehicle system as well as dataset. Subfigures (a) and (b) show respectively the actual sensors of the infrastructure and the ego-vehicle. In subfigure (c), green boxes are ground truth and blue boxes are the detection results of infrastructure perception. Due to inference time of algorithm and transmission delays from the roadside to the vehicle, the infrastructure perception data lags behind the vehicle's perception data.

To address the challenge of vehicle-to-infrastructure cooperative 3D object detection in real-world scenarios, we propose an end-to-end multi-modality late fusion framework based on transformer. The key idea is to encode the anchors predicted from infrastructure perception into infrastructure querie for fusion. Its input consists of the sensor data from vehicle and the received perception results from infrastructure, with the output being 3D object detection results in the ego-vehicle's lidar coordinate system. Its input can also be replaced with other agents' perception results, so it can be readily extended to Vehicle-to-Vehicle (V2V) and Vehicle-to-Everything (V2X) collaborative perception scenarios.

In summary, our contributions are two-fold:

- We propose OTVIC, the first multi-modality and multi-view dataset with online transmission for vehicle-to-infrastructure cooperative 3D object detection. All frames are captured from real scenes in which the vehicle receives the results of infrastructure perception in real-time.
- We introduce LfFormer, a novel end-to-end multi-modality late fusion framework with transformer as a baseline based on OTVIC. The results show the effectiveness and robustness of our fusion framework.

II. Related Works

A. Vehicle-to-Infrastructure Datasets

The vehicle-to-infrastructure (V2I) datasets can primarily be collected either from simulators or real-world. Although collecting from simulators is low-cost and easy to implement, it is challenging to simulate the variety of problems that may be encountered in real-world scenes. DAIR-V2X-C [8] is the first multi-modality and multi-view V2I dataset from real scenarios. V2X-Seq (SPD) [9] is the first temporal perception dataset for V2I cooperative 3D object detection and tracking in real-world scenarios. However, in both datasets, the vehicle is only equipped with a single forward camera, which does not allow for the research of multi-view bird's eye view (BEV) perception algorithms of camera-only. The speed of vehicles in these dataset is also slower than ours. Furthermore, these datasets do not account for the dynamics of delay, bandwidth and real-time requirements of actual V2X communication.

TABLE I

Vehicle Hardware Specifications

Sensor	Details
LiDAR	Velodyne VLP-32C, 32 beams, 10 Hz capture frequency, 360° horizontal FOV, -25° to 15° vertical FOV, 200 m capture range
Camera	OAK-FFC-4P board with four OV9782 cameras, RGB, 20 Hz capture frequency, 1280 × 800 resolution, 120° FOV
GPS/RTK	CHCNAV P3DU, 20 Hz update rate
IMU	Xsens MTi-G-700, 400 Hz update rate
CAN/LIN	Kvaser Hybrid, 100 Hz update rate
OBU	Nebula, LTE-V2X, 5905-5915 MHz frequency bands

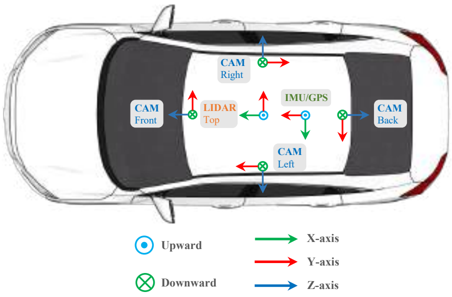


Fig. 3. Sensor setup for the ego-vehicle in OTVIC.

B. Multi-Modality Fusion Perception

Multi-modality fusion perception is the integration of heterogeneous information collected by different sensors, such as lidar, radar, and camera, which can enhance the effectiveness and robustness of perception compared to using a single sensor. Based on the different fusion stages, it can be categorized into early fusion, intermediate fusion, and late fusion. The core idea of early fusion is to extract information from images to enhance or filter the point clouds, followed by a point cloud detector to obtain the perception results. Examples include F-PointNet [10], PointPainting [11], PointAugmenting [12], and so on. The notion of intermediate fusion is to fuse the feature maps which are extracted from the different sensors' data, such as BEVFusion [13], [14] and TransFusion [15]. Late fusion is to fuse the perception results obtained from different sensors, like CLOCs [16] and Fast-CLOCs [17].

C. Vehicle-to-Infrastructure Collaborative Perception

Vehicle-to-infrastructure cooperative perception uses the sensors from both vehicle and infrastructure to jointly accomplish the perception task of the surrounding environment.

TABLE II

COMPARISON BETWEEN VEHICLE-TO-INFRASTRUCTURE COOPERATIVE PERCEPTION DATASET.

Dataset	Year	Source	Scenario	Transmission	Image	Point cloud	IMU/GPS	Frames
CoopInf [3]	2020	CARLA [23]	T-junction & Roundabout	Offline	✓	×	×	10,000
CARTI [4]	2022	CARLA [23]	Crossroads	Offline	×	✓	×	11,000
WIBAM [24]	2021	Real-World	Crossroads	Offline	✓	×	×	33,092
DAIR-V2X-C [8]	2021	Real-World	Intersections	Offline	✓	✓	✓	9,331
V2X-Seq (SPD) [9]	2023	Real-World	Intersections	Offline	✓	✓	✓	15,000
OTVIC (Ours)	2024	Real-World	Highway	Online	✓	✓	✓	15,045

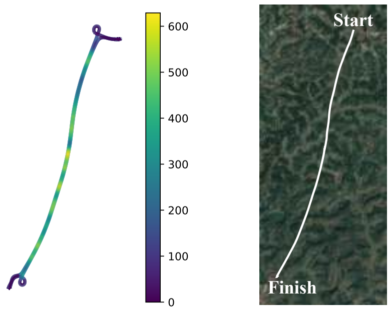
Similarly, the cooperative perception models can be categorized based on the fusion stage into early, intermediate, and late fusion. Early fusion [18], [19] directly transforms raw data and merges it to form a comprehensive perception. This method tend to require a large communication bandwidth due to the large scale of the raw data and is difficult to operate in real-time. Intermediate fusion [6], [20], [21] fuses the feature maps from both sides into a unified feature representation. This method achieves a balance between accuracy and transmission bandwidth. However, compression and decompression of feature maps may result in some loss. Additionally, due to significant temporal asynchrony and spatial misalignment in highway scenarios, this may lead to blurring and misalignment of the feature maps, which can easily lead to performance degradation. Late fusion combines the outputs of perception from infrastructure and vehicle. Existing works often use non-maximum suppression (NMS) [22]. Although this method requires minimal communication bandwidth, which can meet the requirements of practical applications, the perception accuracy of this method is relatively low. In a real-time system, we need to consider issues such as delay, bandwidth limitations, and communication noise, which are critical for the vehicle-to-infrastructure cooperative perception system.

III. SYSTEM AND DATASET

In order to research vehicle-to-infrastructure cooperative perception that can be practically applied, we establish systems for infrastructure and vehicle in real world and propose the OTVIC dataset. Here we describe how to collect data in real-time and annotate the dataset. Finally, we present a statistical analysis of the dataset. A visualization of the systems and the dataset is depicted in Fig. 2.

A. Infrastructure System

The infrastructure perception system is comprised of 7 or 8 cameras, an edge computing device, a RSU, and a cloud platform [7]. The cameras are mounted on 4 poles with different pitch angles at a height of 10 to 20 meters above the ground. These poles are installed at set intervals in the middle of the highway or on the side. Typically, the cameras on each pole are set to two different focal lengths to cover both near-range and far-range vehicles. The edge computing device supports the fusion perception of seven or eight cameras, providing computational power for BEV detection algorithms. By utilizing multi-sensor fusion, the infrastructure can provide perception within an area of 800 by 80 meters. RSU is the communication hub with a coverage range of 800 meters, transmissing message between the vehicles, other RSUs, and the cloud. It transmits data to the OBU with frequencies from 5915 to 5925 MHz. The cloud platform, supported by Alibaba Cloud, is used to collect data from all roadside devices and conduct real-time monitoring of the entire road.



(a) Spatial data coverage

(b) Route

Fig. 4. Map for the OTVIC dataset. In subfigure (a), colors indicate the number of frames with ego vehicle poses within an 80m radius across all scenes. Subfigure (b) shows the route of data collection in the real world.

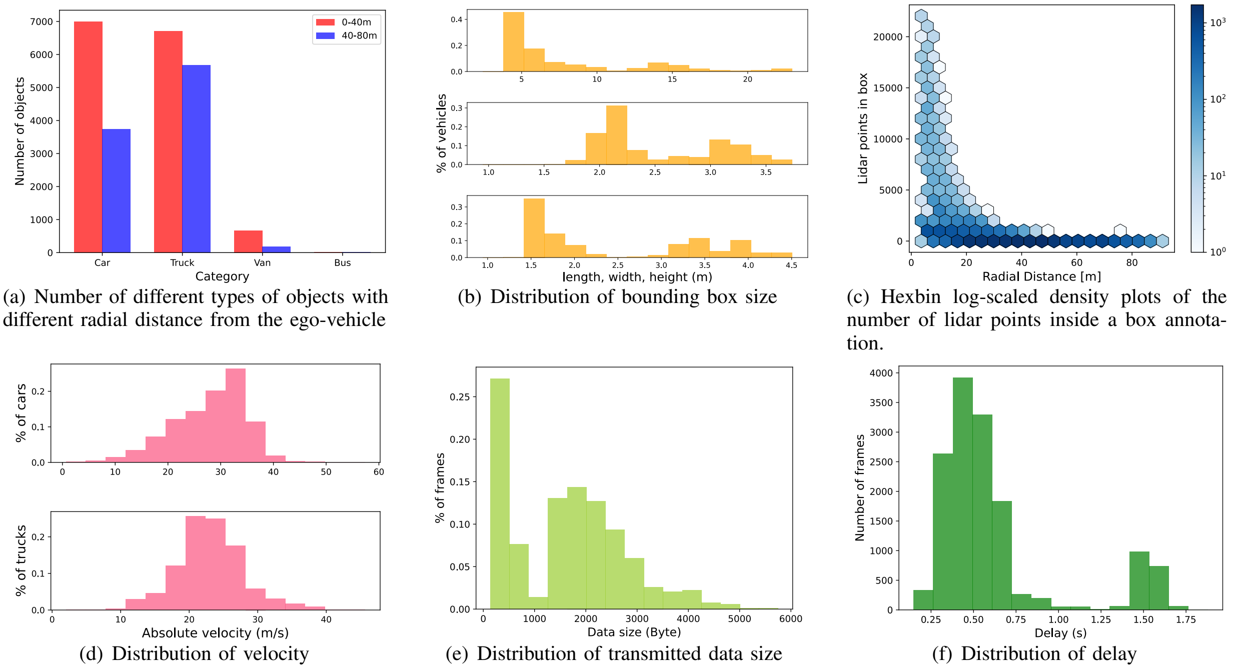
B. Vehicle System

The Vehicle System is composed of the perception module, the localization module, the communication module, and a computer. A more detailed description of the hardware configurations is depicted in the Table I. Specifically, we carefully calibrate the extrinsics and intrinsics of every sensor. The middleware framework we use is Robot Operating System (ROS 1).

The perception module consists of four cameras and one lidar as Fig. 3 shows. The four cameras are oriented towards the front, rear, left, and right directions of the vehicle, achieving timestamp alignment among multiple cameras through millisecond-level hardware synchronization. And the lidar is mounted on the top of the vehicle.

The localization module consists of an Inertial Measurement Unit (IMU) and a Global Positioning System (GPS). They are mounted at the center position of the vehicle's rear axle. We employ an algorithm based on the Extended Kalman Filter (EKF) for multi-sensor fusion to achieve accurate localization. This approach uses data from IMU to predict the vehicle's location and applies GPS data to correct the pose and motion estimates.

The communication module is the On-Board Unit (OBU) for receiving information from roadside devices in real-time. Additionally, the vehicle system achieves clock synchronization with the infrastructure system through Network Time Protocol (NTP).



(a) Number of different types of objects with different radial distance from the ego-vehicle

(b) Distribution of bounding box size

(c) Hexbin log-scaled density plots of the number of lidar points inside a box annotation.

(d) Distribution of velocity

(e) Distribution of transmitted data size

(f) Distribution of delay

Fig. 5. Analysis of OTVIC dataset. Subfigures (a), (b) and (c) reveal the statistics of the 3D bounding box annotations in our dataset. Subfigure (d) shows the distribution of speeds for cars and trucks. Subfigure (e) presents the distribution of data size transmitted from RSU to OBU. And subfigure (f) illustrates the statistics of delay from infrastructure to vehicle. This delay includes the inference time of infrastructure perception algorithm, the real-time transmission latency from RSU to OBU and the time difference between the timestamp of the received roadside data and the lidar timestamp.

C. *OTVIC Dataset*

Data Collection. Our dataset is collected on highways at speeds ranging from 70 to 110 km/h. When the ego-vehicle drives into the detectable area of the roadside devices, the perception data received from the roadside and the sensor data from the ego-vehicle are saved to the local hard drive of vehicle. Based on the lidar timestamps, the saved vehicle sensor data are sampled at a frequency of 10Hz to obtain discrete frames. Each frame of vehicle is matched with the closest frame of infrastructure which is received in real-time before the current lidar timestamp. After the data synchronization, we manually select 112 representative scenes, each encompassing several seconds in duration. Fig. 4 shows spatial coverage across all scenes and route in the real world. The dataset totals 15,045 frames. Each frame includes images, point clouds, and localization and motion information outputted by the localization module from ego-vehicle, as well as the perception results from infrastructure.

Data Annotation. In order to get vehicle-to-infrastructure cooperative annotations, we convert 3D bounding boxes of infrastructure into the ego-vehicle coordinate system and fuse the infrastructure annotations and vehicle annotations. Through multiple validation and refinement steps, expert annotators make high-quality annotations for each frame of the dataset. In particular, annotators comprehensively annotate each of the four object classes in every image and point cloud with its type, position, size, yaw angle, 3D bounding box, and ID. 4 categories include car, $truck$, van, and bus. A total of 24,452 manually annotated vehicles, including 10,823 cars, 12,750 trucks, 848 vans, and 31 buses.

Data Analysis. As Table II depicts, we compare the open vehicle-to-infrastructure cooperative perception datasets with OTVIC. Our dataset is collected in real-time from high-speed scenarios in which infrastructure sends the results of perception to vehicles for late fusion. Statistically, the average speed for moving car and truck categories are 27.99 and 23.04 m/s. The packet loss rate for infrastructure data is less than 1%. More statistics about the dataset are illustrated in Fig. 5. In our dataset, the infrastructure data is a vector with the dimensions (N, 8), including the type, position, heading angle, velocity, acceleration, ID, and delay for each object, where N is the number of objects. It has converted from the coordinate system of the infrastructure localization (GCJ02) to the ego-vehicle's lidar coordinate system. Additionally, we provide the ego-vehicle's location and motion information estimated by IMU and GPS, which is a vector with 6 dimensions, including ego-vehicle's position, velocity, acceleration, yaw angle, and angular velocity. It has also transformed to the lidar coordinate system.

Data Protection. Before the public release, we mask license plates and faces to protect privacy because of local laws and regulations. We also erase real geographic information by transforming the position into a coordinate system of a virtual world.

IV. Method

In this section, we propose an end-to-end multi-modality fusion framework based on the OTVIC dataset as a baseline method. We introduce the overall architecture of the fusion framework in Fig. 6 and then show the details of LfFormer. Finally, we demonstrate the loss functions for model training.

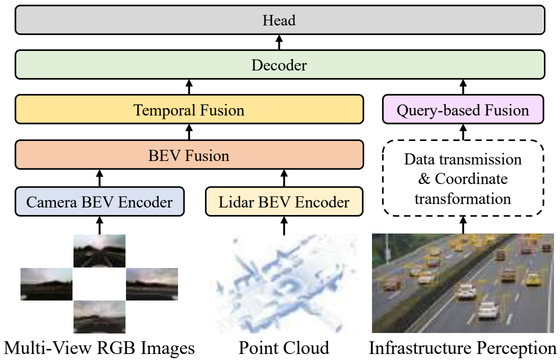


Fig. 6. Architecture of Our Fusion Framework (LfFormer). The input is the multi-view RGB images and point cloud from the ego-vehicle sensor, as well as the perception results from the infrastructure. The output is 3D object detection results in the ego-vehicle's lidar coordinate system.

A. Overall Architecture

In order to addresses various challenges about collaborative perception between vehicle and infrastructure, we propose a feature-result level fusion framework tailored for real-world application scenarios. In this method, the roadside unit provides result-level data and sends to the vehicle in real-time. Given the smaller data volume of result-level data, it meets the requirements of actual communication bandwidth. However, infrastructure perception data has asynchronous and heterogeneous characteristics, with inherent errors and delays that necessitate spatial and temporal alignment and compensation. Moreover, the vehicle fuses multi-modality data from images and point clouds to obtain feature-level data. Then, we use a novel network based on the Transformer [25] to achieve the feature-result level fusion, thereby accomplishing the task of vehicle-to-infrastructure collaborative perception. It consists of seven sub-modules: Camera BEV Encoder, Lidar BEV Encoder, BEV Fusion, Temporal Fusion, Query-based Fusion, Decoder, and Head.

B. LfFormer

Camera and Lidar BEV Encoder. We use ResNet [26] for feature extraction from images to obtain 2D features. Inspired by BEVFormer [27], it employs spatial cross-attention to learn feature representations in the bird's-eye view (BEV) space. This method extracts spatial features from regions of interest across camera views based on a predefined grid of BEV queries. The Lidar BEV Encoder, using VoxelNet [28] or PointPillars [29], converts the point clouds into Voxel or Pillars features, which are further flatten into BEV feature.

BEV Fusion. We use a 3×3 convolution layer as the BEV fusion module, which is designed to fuse BEV features from both the camera and lidar effectively. It reduces the BEV features of camera and lidar from the dimensions $[B, C_{camera} + C_{lidar}, H, W]$ to $[B, C_{fusion}, H, W]$.

Temporal Fusion. In order to fuse historical BEV features and learn rich information such as the motion characteristics of detected objects, we use a temporal fusion module based on temporal self-attention [27] to enhance the performance of perception. It uses the ego-vehicle motion information to align the BEV features of the previous frame to the current frame, and then fuses with the BEV features of the current frame.

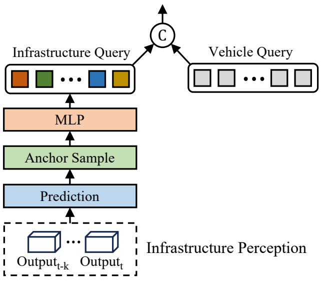


Fig. 7. The details of Query-based Fusion.

Query-based Fusion. To integrate roadside perception data, we propose a query-based fusion module. The key idea is to encode the anchors from infrastructure perception into infrastructure queries after prediction and sampling. The details of this module is illustrated in Fig. 7. First, we use a historical sequence of infrastructure perception to predict based on the equations of motion with constant acceleration, and then anchors are sampled around the predicted positions. Subsequently, these anchors are encoded into infrastructure query using a MLP network. Finally, they are concatenated with vehicle query which is predicted by heatmap [15].

Decoder. The decoder is a standard Transformer [25] decoder which is composed of a self-attention layer, a cross-attention layer, and a feed forward network. The fused BEV features output by the temporal fusion module serve as the key and value for the decoder, while the output from the query-based fusion module is the decoder's query.

Head. The head includes two branches: classification and regression, which is composed of fully connected networks. The output of classification branch is the confidence score of being an object or backgroud. And the output of regression branch is $(x, y, z, l, w, h, \theta)$, denoting the position, size, and yaw angle of the bounding box. Additionally, this fusion framework does not require non-maximum suppression (NMS) post-processing.

C. Loss Function

In this fusion framework, we use the bipartite matching between the predicted bounding boxes and ground truth through the Hungarian algorithm [30]. We adopt a l_1 loss for regression of bounding boxes and a focal loss for object classification. And the Gaussian focal loss function is used for the prediction of the heatmap. Our total loss function consists of a weighted sum of the regression loss, classification loss, and heatmap loss:

$$Loss = \omega_r L_r + \omega_c L_c + \omega_h L_h \tag{1}$$

where ω_r, ω_c and ω_h represent respectively the weight of the regression loss, classification loss and heatmap loss.

V. EXPERIMENTS

In this section, we present a V2I 3D object detection benchmark on our OTVIC dataset and analyze the ex-

TABLE III

3D OBJECT DETECTION BENCHMARK ON OTVIC.

Modality	**Fusion**	**Model**	mAP_{0-40m}	mAP_{40-80m}	$mAP_{overall}$
Image	No Fusion	Vehicle-only-C	0.518	0.228	0.421
	Late Fusion	NMS	0.527	0.249	0.432
	Late Fusion	TCLF [8]	0.530	0.261	0.439
	Late Fusion	LfFormer-C	**0.531**	**0.309**	**0.458**
Image & Pointcloud	No Fusion	Vehicle-only	0.792	0.652	0.761
	Late Fusion	NMS	0.794	0.679	0.768
	Late Fusion	TCLF [8]	0.803	0.681	0.773
	Late Fusion	LfFormer (Ours)	**0.807**	**0.702**	**0.784**

perimental results quantitatively and qualitatively. Finally, we conduct robustness and ablation studies on the OTVIC dataset for the LfFormer model.

A. Benchmark Models

To reduce sensor costs, some autonomous vehicles are equipped only with multiple cameras, omitting the use of lidar. Here we compare the performance of different methods based on two modalities.

Since the OTVIC dataset contains infrastructure perception results rather than raw sensor data, we choose the late fusion method, including Non-Maximum Suppression (NMS) and Time Compensation Late Fusion (TCLF) [8] as our baselines. In order to compare cooperative perception and single-vehicle perception, we also investigate the model performance without infrastructure perception, named vehicle-only.

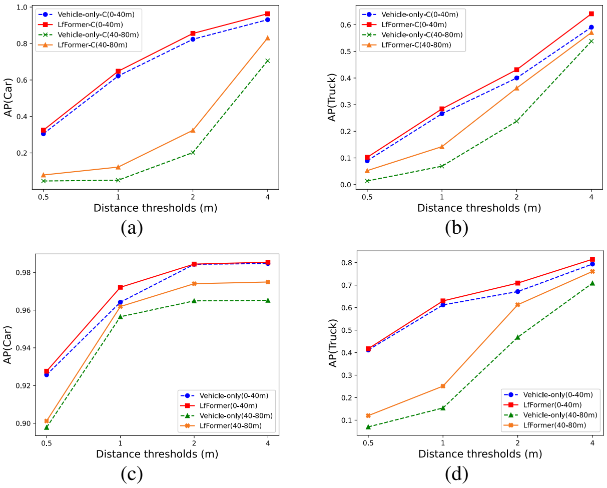
Vehicle-only. We use the aforementioned fusion framework (LfFormer) without infrastructure perception input as the vehicle-only perception model. The model without the lidar stream is called Vehicle-only-C. Here we choose BEVFormer [27] as the camera stream and VoxelNet [28] as the lidar stream.

Non-Maximum Suppression (NMS). We use vehicle-only models to estimate the bounding boxes with confidence scores. The perception results from infrastructure are predicted using a constant acceleration model. NMS is applied to these proposals from both vehicle and infrastructure to generate the final 3D object detection.

Time Compensation Late Fusion (TCLF). The TCLF predicts and matches the bounding boxes across successive infrastructure frames. For matched vehicles, it computes their velocities directly. For unmatched vehicles, a learning-based method is used to predict their velocities. And then it approximates the positions of the current frame by linear interpolation and fuses them with the perception results from the ego-vehicle.

B. Metrics

The evaluation metric uses the mean Average Precision (mAP) for all categories of objects across different distance ranges. Similar to nuscenes [31], we calculate precision-recall curve at different thresholds, defining a match by the 2D center distance d on the ground plane, rather than intersection over union (IOU) for each object category. Then, we calculate the Average Precision (AP) as the normalized area under the precision-recall curve, excluding operating points with recall or precision below 10% to reduce noise impact. If no points meet this criterion, the AP for that category is zero. Finally, we average over matching thresholds of $\mathbb{M} = \{0.5, 1, 2, 4\}$ meters and summarize the AP values across all categories to obtain the mAP. Assuming the set of classes is $\mathbb{N}$, the formula for computing mAP is as follows:



(a)
(b)
(c)
(d)

Fig. 8. Average Precisio (AP) for each class at different distance thresholds. Subfigure (a) and (b) compare the AP for car and truck using the Vehicle-only-C and LfFormer-C models. Subfigure (c) and (d) compare the AP for car and truck using the Vehicle-only and LfFormer models.

$$mAP = \frac{1}{|\mathbb{M}||\mathbb{N}|} \sum_{m \in \mathbb{M}} \sum_{n \in \mathbb{N}} AP_{m,n} \qquad (2)$$

C. Experiment Details

The dataset is divided into training, validation, and test sets in a 7:1:2 ratio. Given the scarcity of vans and buses in the dataset, the model training will only consider two categories of objects: car and truck. The LfFormer model training adopts a three-stage approach. We use the AdamW optimizer to iteratively update the network parameters for all stages. Firstly, we train the camera stream for 24 epochs where the initial learning rate is set as $5e^{-5}$ and the weight decay is set as 0.01. Secondly, we train the lidar stream for 12 epochs. We set the learning rate to $6.25e^{-6}$ and set the weight decay to 0.01. Finally, based on the camera and lidar streams from the above two steps, our vehicle-to-infrastructure collaborative perception model is trained for 6 epochs with the initial learning rate $1.25e^{-4}$ and the weight decay 0.05. Our proposed detection network is trained on four Nvidia 3090 GPU with batch size 4. We set the

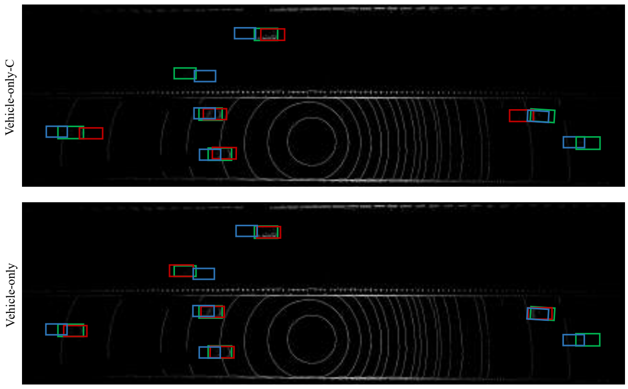


Fig. 9. Visualization of detection results for no fusion methods. Green boxes are ground truth, red boxes are the detection results of vehicle-only-C or vehicle-only, and blue boxes are the results of infrastructure perception with prediction.

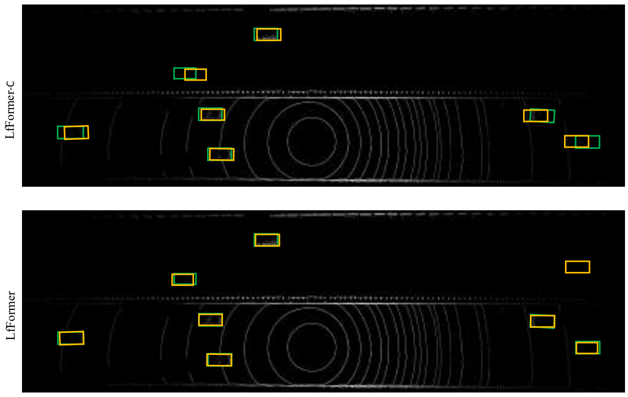


Fig. 10. Visualization of detection results for LfFormer-C and LfFormer. Green and yellow boxes are ground truth and detection results respectively.

perceptual range to $x \in [-80, 80]$ meters, $y \in [-80, 80]$ meters.

D. Benchmark Analysis

Quantitative evaluation. We compare the performance of different methods and present a 3D object detection benchmark for V2I cooperative 3D object detection across three range intervals: 0-40 m, 40-80 m, and 0-80 m (overall). In Table III, we can see that the detection performance of cooperative perception is better than single-vehicle perception and our proposed fusion framework achieves the highest mAP. Due to the short focal length of the camera, distant objects in the image are small, resulting in a lower mAP for camera-only methods. However, the infrastructure perception data can improve the performance for the detection of distant objects. The AP curves for each category at different distance thresholds are shown in Fig. 8. Compared to single-vehicle perception, our cooperative perception fusion framework has a major increase of AP at distance thresholds of 1, 2, and 4.

TABLE IV
ROBUSTNESS EXPERIMENTS UNDER DIFFERENT SCENARIOS ON THE OTVIC DATASET FOR LFFORMER.

Scenarios	Interval / Ratio	mAP
Delay	$0 - 300$ ms	0.786
	$300 - 600$ ms	0.785
	$600 - 900$ ms	0.778
	≥ 900 ms	0.777
Packet dropout	25%	0.782
	50%	0.779
	75%	0.775
	100%	0.761

Qualitative evaluation. Because it's difficult to compare differences in 3D bounding boxes within images, we project the detection results into the lidar coordinate system for all methods. Firstly, we illustrate the detection results of vehicle-only in Fig. 9. Due to the effects of temporal asynchrony and spatial misalignment, the precision of infrastructure perception is lower. However, it can detect all objects because of a larger field of view for perception. Influenced by the limited perception field of view for ego-vehicle, there are problems of detecting distant objects, whereas the detection accuracy for nearby objects is relatively high. Additionally, vehicle-only method also struggle to detect objects that are obscured, such as the oncoming vehicles obscured by the median strip. And we also illustrate the detection results of LfFormer-C and LfFormer in Fig. 10. With the help of infrastructure perception, LfFormer-C has better performance than Vehicle-only-C. Because of the input of point clouds, LfFormer exhibits better performance than LfFormer-C, achieving precise compensation and outstanding detection results.

E. Robustness Analysis

To test the robustness of the model, we select two scenarios: different online transmission delays and packet dropout ratios for perception data received from the infrastructure, as shown in Table IV.

Robustness to delay. In our dataset, the inference latency of the roadside perception algorithm and the transmission delay from infrastructure to vehicle are dynamically changing. Here we calculate the performance at various delays: 0-300 ms, 300-600 ms, 600-900 ms, and over 900 ms. Because we retrieve historical infrastructure perception and predict in order to compensate for the delay in our fusion framework, we can see that LfFormer still performs well even when the delay is high.

Robustness to packet dropout. We simulate different packet loss rates by randomly dropping the data received every second from infrastructure. Here we simulate the infrastructure perception packet dropout ratios of 25%, 50%, 75%, as well as 100% and evaluate the performance of our fusion framework under each condition. The results indicate that as the packet dropout ratio grows, there is a decline in the model's performance.

F. Ablation Study

We conduct ablation study on the OTVIC dataset for the LfFormer model. Table V assesses the effectiveness of the proposed operations, including prediction on infrastructure perception and anchor sampling. We can see that: i) the prediction module for infrastructure perception can effectively compensate for the temporal asynchrony. ii) Sampling anchors at the predicted locations can mitigate the impact of

TABLE V
ABLATION STUDY ON THE OTVIC DATASET FOR LFFORMER.

Prediction	Sample	mAP
×	×	0.766
✓	×	0.773
✓	✓	0.784

spatial misalignments caused by inaccuracies in infrastructure perception or prediction.

VI. CONCLUSION

In this paper, we propose a dataset and a late fusion framework based on the various issues and challenges of vehicle-to-infrastructure cooperative perception in real-world scenarios. This method effectively leverages the accuracy of vehicle perception and the global perspective of infrastructure perception with small communication bandwidth, providing a safer and more reliable perception for autonomous driving. It can be easily extended to Vehicle-to-Vehicle (V2V) and Vehicle-to-Everything (V2X) collaborative scenarios for further research.

REFERENCES

[1] J. Mao, S. Shi, X. Wang, and H. Li, "3d object detection for autonomous driving: A review and new outlooks," *arXiv preprint arXiv:2206.09474*, 2022.

[2] S. Liu, C. Gao, Y. Chen, X. Peng, X. Kong, K. Wang, R. Xu, W. Jiang, H. Xiang, J. Ma, *et al.*, "Towards vehicle-to-everything autonomous driving: A survey on collaborative perception," *arXiv preprint arXiv:2308.16714*, 2023.

[3] E. Arnold, M. Dianati, R. de Temple, and S. Fallah, "Cooperative perception for 3d object detection in driving scenarios using infrastructure sensors," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 3, pp. 1852–1864, 2020.

[4] Z. Bai, G. Wu, M. J. Barth, Y. Liu, E. A. Sisbot, and K. Oguchi, "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)*, pp. 1743–1749, IEEE, 2022.

[5] Y. Li, D. Ma, Z. An, Z. Wang, Y. Zhong, S. Chen, and C. Feng, "V2x-sim: Multi-agent collaborative perception dataset and benchmark for autonomous driving," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 10914–10921, 2022.

[6] R. Xu, H. Xiang, Z. Tu, X. Xia, M.-H. Yang, and J. Ma, "V2x-vit: Vehicle-to-everything cooperative perception with vision transformer," in *European conference on computer vision*, pp. 107–124, Springer, 2022.

[7] X. Zhu, H. Sheng, S. Cai, B. Deng, S. Yang, Q. Liang, K. Chen, L. Gao, J. Song, and J. Ye, "Roscenes: A large-scale multi-view 3d dataset for roadside perception," *arXiv preprint arXiv:2405.09883*, 2024.

[8] H. Yu, Y. Luo, M. Shu, Y. Huo, Z. Yang, Y. Shi, Z. Guo, H. Li, X. Hu, J. 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*, pp. 21361–21370, 2022.

[9] H. Yu, W. Yang, H. Ruan, Z. Yang, Y. Tang, X. Gao, X. Hao, Y. Shi, Y. Pan, N. 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*, pp. 5486–5495, 2023.

[10] C. R. Qi, W. Liu, C. Wu, H. Su, and L. J. Guibas, "Frustum pointnets for 3d object detection from rgb-d data," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 918–927, 2018.

[11] S. Vora, A. H. Lang, B. Helou, and O. Beijbom, "Pointpainting: Sequential fusion for 3d object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4604–4612, 2020.

[12] C. Wang, C. Ma, M. Zhu, and X. Yang, "Pointaugmenting: Cross-modal augmentation for 3d object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11794–11803, 2021.

[13] T. Liang, H. Xie, K. Yu, Z. Xia, Z. Lin, Y. Wang, T. Tang, B. Wang, and Z. Tang, "Bevfusion: A simple and robust lidar-camera fusion framework," *Advances in Neural Information Processing Systems*, vol. 35, pp. 10421–10434, 2022.

[14] Z. Liu, H. Tang, A. Amini, X. Yang, H. Mao, D. L. Rus, and S. Han, "Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation," in *2023 IEEE international conference on robotics and automation (ICRA)*, pp. 2774–2781, IEEE, 2023.

[15] X. Bai, Z. Hu, X. Zhu, Q. Huang, Y. Chen, H. Fu, and C.-L. Tai, "Transfusion: Robust lidar-camera fusion for 3d object detection with transformers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1090–1099, 2022.

[16] S. Pang, D. Morris, and H. Radha, "Clocs: Camera-lidar object candidates fusion for 3d object detection," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 10386–10393, IEEE, 2020.

[17] S. Pang, D. Morris, and H. Radha, "Fast-clocs: Fast camera-lidar object candidates fusion for 3d object detection," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 187–196, 2022.

[18] Q. Chen, S. Tang, Q. Yang, and S. Fu, "Cooper: Cooperative perception for connected autonomous vehicles based on 3d point clouds," in *2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS)*, pp. 514–524, IEEE, 2019.

[19] Q. Chen, X. Ma, S. Tang, J. Guo, Q. Yang, and S. Fu, "F-cooper: Feature based cooperative perception for autonomous vehicle edge computing system using 3d point clouds," in *Proceedings of the 4th ACM/IEEE Symposium on Edge Computing*, pp. 88–100, 2019.

[20] K. Yang, D. Yang, J. Zhang, M. Li, Y. Liu, J. Liu, H. Wang, P. Sun, and L. Song, "Spatio-temporal domain awareness for multi-agent collaborative perception," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 23383–23392, 2023.

[21] Y. Hu, Y. Lu, R. Xu, W. Xie, S. Chen, and Y. Wang, "Collaboration helps camera overtake lidar in 3d detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9243–9252, 2023.

[22] R. Xu, H. Xiang, X. Xia, X. Han, J. Li, and J. 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)*, pp. 2583–2589, IEEE, 2022.

[23] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "Carla: An open urban driving simulator," in *Conference on robot learning*, pp. 1–16, PMLR, 2017.

[24] M. Howe, I. Reid, and J. Mackenzie, "Weakly supervised training of monocular 3d object detectors using wide baseline multi-view traffic camera data," *arXiv preprint arXiv:2110.10966*, 2021.

[25] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[27] Z. Li, W. Wang, H. Li, E. Xie, C. Sima, T. Lu, Y. Qiao, and J. Dai, "Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers," in *European conference on computer vision*, pp. 1–18, Springer, 2022.

[28] Y. Zhou and O. 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*, pp. 4490–4499, 2018.

[29] A. H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang, and O. Beijbom, "Pointpillars: Fast encoders for object detection from point clouds," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12697–12705, 2019.

[30] H. W. Kuhn, "The hungarian method for the assignment problem," *Naval research logistics quarterly*, vol. 2, no. 1-2, pp. 83–97, 1955.

[31] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11621–11631, 2020.