智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

面向智能网联汽车的 BEV 感知技术与发展趋势

宫彦, 王乃棒, 张新钰, 苏纳宇, 赵红飞, 袁云, 鲁建丽, 胡小溪, 刘华平

引用本文:
宫彦, 王乃棒, 张新钰, 等. 面向智能网联汽车的 BEV 感知技术与发展趋势[J]. 智能系统学报, 2026, 21(1): 41-59.
GONG Yan, WANG Naibang, ZHANG Xinyu, et al. BEV perception technologies and development trends for intelligent connected vehicles[J]. *CAAI Transactions on Intelligent Systems*, 2026, 21(1): 41-59.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202505027

您可能感兴趣的其他文章

融合视觉显著性再检测的孪生网络无人机目标跟踪算法
Siamese network combined with visual saliency re-detection for UAV object tracking
智能系统学报. 2021, 16(3): 584-594 https://dx.doi.org/10.11992/tis.202101035

面向观测融合和吸引因子的多机器人主动SLAM
Multi-robot active SLAM for observation fusion and attractor
智能系统学报. 2021, 16(2): 371-377 https://dx.doi.org/10.11992/tis.202006019

一种改进的深度学习的道路交通标识识别算法
An improved deep learning algorithm for road traffic identification
智能系统学报. 2020, 15(6): 1121-1130 https://dx.doi.org/10.11992/tis.201811009

基于邻域系统的智能车辆最优轨迹规划方法
Optimal trajectory planning method of intelligent vehicles based on neighborhood system
智能系统学报. 2019, 14(5): 1040-1047 https://dx.doi.org/10.11992/tis.201805004

基于图像聚类的交通标志CNN快速识别算法
CNN-based image clustering algorithm for fast recognition of traffic signs
智能系统学报. 2019, 14(4): 670-678 https://dx.doi.org/10.11992/tis.201806026

多移动机器人协同搬运技术综述
Technologies for cooperative transportation by multiple mobile robots
智能系统学报. 2019, 14(1): 20-27 https://dx.doi.org/10.11992/tis.201801038

DOI: 10.11992/tis.202505027
网络出版地址: https://link.cnki.net/urlid/23.1538.tp.20260104.0935.002

面向智能网联汽车的 BEV 感知技术与发展趋势

宫彦[1,2,3], 王乃棒[1,2], 张新钰[1,2], 苏纳宇[1,2,4], 赵红飞[1,2], 袁云[1,2], 鲁建丽[1,2], 胡小溪[1,2], 刘华平[5]
(1. 清华大学 智能绿色车辆与交通全国重点实验室, 北京 100084; 2. 清华大学 车辆与运载学院, 北京 100084; 3. 哈尔滨工业大学 机器人技术与系统国家重点实验室, 黑龙江 哈尔滨 150001; 4. 燕山大学 电气工程学院, 河北 秦皇岛 066004; 5. 清华大学 计算机科学与技术系, 北京 100084)

摘 要: 鸟瞰视图 (bird's eye view, BEV) 感知因其统一且可解释的空间表达能力, 已成为自动驾驶环境理解的核心技术。本文旨在全面阐述面向智能网联汽车的 BEV 感知技术, 总结相关公开数据集, 探讨相关挑战及发展趋势, 为该领域提供系统的理论支持与实践指导。本文系统梳理了 BEV 感知技术在自动驾驶中的研究进展, 围绕路端及车路协同应用场景, 构建了涵盖纯视觉、纯点云与多模态融合的技术框架, 深入分析了代表性方法的核心思想与实现机制。本文首次在数据层面进行系统整理, 并比较了现有 BEV 感知相关的数据集, 包括规模、传感器配置与标注类型。本文聚焦 BEV 感知在开放类别识别、大规模无监督数据利用、传感器不确定性等关键挑战, 并探讨其与端到端自动驾驶、具身智能、大模型协同感知架构的融合趋势。

关键词: 智能网联汽车; 车路协同; 协同感知; 鸟瞰视图; 自动驾驶; 数据集; 车联万物; 多模态融合
中图分类号: TP391.41; U463.6; U495 **文献标志码:** A **文章编号:** 1673−4785(2026)01−0041−19

中文引用格式: 宫彦, 王乃棒, 张新钰, 等. 面向智能网联汽车的 BEV 感知技术与发展趋势 [J]. 智能系统学报, 2026, 21(1): 41–59.
英文引用格式: GONG Yan, WANG Naibang, ZHANG Xinyu, et al. BEV perception technologies and development trends for intelligent connected vehicles[J]. CAAI transactions on intelligent systems, 2026, 21(1): 41–59.

BEV perception technologies and development trends for intelligent connected vehicles

GONG Yan[1,2,3], WANG Naibang[1,2], ZHANG Xinyu[1,2], SU Nayu[1,2,4], ZHAO Hongfei[1,2], YUAN Yun[1,2], LU Jianli[1,2], HU Xiaoxi[1,2], LIU Huaping[5]
(1. State Key Laboratory of Intelligent Green Vehicle and Mobility, Tsinghua University, Beijing 100084, China; 2. School of Vehicle and Mobility, Tsinghua University, Beijing 100084, China; 3. the State Key Laboratory of Robotics and System, Harbin Institute of Technology, Harbin 150001, China; 4. Institute of Electrical Engineering, Yanshan University, Qinhuangdao 066004, China; 5. Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China)

Abstract: Bird's eye view (BEV) perception has become a fundamental technique for environmental understanding in autonomous driving, due to its unified and interpretable spatial representation. This survey provides a comprehensive review of BEV perception technologies tailored for intelligent connected vehicles. It systematically categorizes existing approaches based on sensor modality and deployment configuration, covering vehicle-side, infrastructure-side, and vehicle-infrastructure cooperative scenarios. The review introduces a multi-dimensional framework encompassing vision-only, LiDAR-only, and multimodal fusion methods, and analyzes representative techniques in terms of their design principles and implementation strategies. In addition, this work presents the first consolidated comparison of BEV-related datasets, detailing their sensor setups, task types, and annotation schemes to support standardized benchmarking. Finally, the survey outlines key challenges—such as open-category recognition, unsupervised learning from large-scale data, and robustness under sensor uncertainty—and explores future directions involving end-to-end autonomous driving, embodied intelligence, and large-model-based cooperative BEV perception systems.

Keywords: intelligent connected vehicles; vehicle-infrastructure cooperation; cooperative perception; BEV; autonomous driving; dataset; vehicle-to-everything (V2X); multimodal fusion

环境感知是自动驾驶系统实现安全驾驶的基础, 鸟瞰视图 (bird's eye view, BEV) 感知因其具备统一、可解释的空间表达能力, 正成为新一代自动驾驶感知系统的核心范式。通过将多源传感器 (如摄像头、激光雷达 (light detection and ranging, LiDAR)、毫米波雷达 (radio detection and ranging, Radar)) 在三维空间中的观测信息投影至统一的二维俯视平面, BEV 表征能够整合丰富的语义和高精度的空间几何信息, 为多传感器融合、跨模态协同及下游任务提供统一的信息中介。

相较于传统感知范式, BEV 感知在处理目标遮挡、尺度不一致等问题上具有天然优势, 能够更清晰地表征目标间的相对空间关系。其空间一

收稿日期: 2025−05−27. **网络出版日期:** 2026−01−04.
基金项目: 国家自然科学基金项目 (62273198); 北京市自然科学基金项目 (L241017).
通信作者: 张新钰. E-mail: xyzhang@tsinghua.edu.cn.

©《智能系统学报》编辑部版权所有

致性使得 BEV 成为自动驾驶系统中三维目标检测、路径规划、车道线分割[1] 等关键任务的理想表达形式，尤其在交叉口、隧道、恶劣天气等复杂场景下展现出更强的鲁棒性。同时，BEV 不仅适用于车端的单车感知系统，也可扩展至路端感知系统，进而通过车路协同感知提升自动驾驶系统对车辆周围环境全面的感知与理解。

基于视觉的 BEV 感知主要依赖从前视图或环视图生成俯视视角，其关键在于高效准确地完成从二维图像到三维空间的映射。早期基于单应矩阵的几何方法[2]，在假设地面平坦的前提下实现视角变换，具备一定实用性，但在真实交通环境中常受地形变化限制。随着深度学习的发展，基于图像深度估计与可微分几何模块的视图变换方法应运而生，逐步替代传统投影方法。根据转换策略不同，主流视觉 BEV 方法可分为 3 类：基于深度估计的显式[3-7] 与隐式[4,8-10] 变换、基于多层感知机 (multi-layer perceptron, MLP) 的空间映射[11-14]，以及基于 Transformer 的查询对齐[14-17]。特别是 Transformer 架构，通过构建 BEV 查询并引入交叉注意力机制，在多视角图像中高效检索融合空间特征[18-19]，已成为视觉 BEV 感知的主流路径。

在点云感知方面，LiDAR 和 Radar 提供的三维点云可直接映射至 BEV 空间，构成基于前投影与后投影的两类方法：前者[3, 6, 20] 将原始点云栅格化后在 BEV 平面中处理，侧重实时性；后者[10, 20-21] 先在三维空间中提取特征再进行投影，保留更多空间几何信息，适用于高精度感知任务。随着多模态融合趋势的发展，图像与点云融合的 BEV 感知方法日益增多，旨在充分利用图像的纹理信息与点云的几何信息，实现更具鲁棒性和高精度的感知系统。

车端 BEV 感知系统作为自动驾驶的核心模块，通过统一的空间表征支撑感知、预测与规划任务。其核心优势在于将多模态传感器数据 (如前视摄像头、环视鱼眼相机、LiDAR、Radar) 映射至鸟瞰视角，消除透视畸变并建立跨传感器的空间一致性，从而为车辆决策提供全局环境理解。在车端部署中，BEV 感知需满足严苛的实时性与低算力约束，推动模型设计向轻量化演进：例如基于 Transformer 的方法通过可学习的 BEV Query 压缩特征维度[22-23]，而基于 MLP 的映射则利用参数共享降低计算复杂度[5, 24]。典型应用场景涵盖自动泊车中的障碍物检测、城区导航的交叉路口通行决策，以及高速场景下的长距离目标跟踪[25]。然而，车端系统仍面临动态障碍物遮挡、极端天气干扰及复杂地形适应性等挑战，需依赖多传感器时序融合与在线标定技术提升鲁棒性。

尽管近年来 BEV 感知相关研究成果不断涌现，现有综述工作仍主要聚焦于车端感知方法及纯视觉范式，尚缺乏对路侧感知与车路协同 BEV 感知系统的系统性梳理与分类。相较于黄德启等[26] 基于模型结构所提出的感知方法分类体系，本文进一步引入“感知部署位置”(包括车端、路端与车路协同) 与“感知模态类型”(包括图像、点云与多模态融合) 两个关键维度，构建了覆盖主流技术路径的多维度感知框架，拓展了分类粒度与系统性；相较于时培成等[27] 聚焦模态演进的回顾性分析，本文进一步引入“跨域特征对齐”这一尚未充分探讨的核心挑战，并提出统一分析范式以贯通感知输入、协同机制与输出空间的一致性建模；相较于肖荣春等[28] 所侧重的视觉 BEV 感知流程剖析，本文更加注重异构模态间的交叉融合机制，系统涵盖基于 LiDAR 及多传感器的多路径建图方案；此外，针对周松燃等[29] 从感知部署视角所构建的分域框架，本文通过引入“部署位置 × 感知模态”的二维分类体系，实现了对现有 BEV 感知范式更具结构性、可扩展性的全景化刻画。

基于上述分析，本文提出了一种兼具系统性与扩展性的 BEV 感知综述框架，旨在弥补现有研究在感知部署与模态融合两个维度上的不足。具体而言，本文从“感知部署位置 (路端、车路两端)”与“感知模态 (图像、点云、融合)”两个维度出发，首次构建了覆盖 BEV 感知主流技术路径的多维度、多模态技术框架。在此基础上，本文系统梳理了路侧及车路协同感知相关的真实与仿真数据集，涵盖其传感器配置、任务类型与标注形式，为后续研究奠定良好基础。同时，本文还聚焦 BEV 感知系统在开放类别识别、大规模无监督数据学习、传感器不确定性等现实挑战，并展望其在端到端闭环系统、具身智能、大模型驱动协同感知等方向的演进趋势。本文的主要贡献包括：

1) 首次构建了多维度、多模态 BEV 感知技术框架，系统总结 BEV 感知在路端与协同场景中的典型方法与关键机制；

2) 首次全面整理并对比了 BEV 感知相关数据集，提供涵盖多模态、多任务的数据资源汇总，为模型评估提供标准基线；

3) 深入分析了当前 BEV 感知在开放世界下的类别识别、无标签数据和传感器不确定性的挑战，并探索了其与大模型、具身智能新兴技术的结合。

1　基于路端的 BEV 感知方法

本文框架如图 1 所示，首先对路端感知相关的研究进行讨论分析，然后对协同感知及相关数

据集进行分析, 最后讨论相关挑战和趋势。基于车端的感知受车辆位置、传感器位置、物体遮挡等因素的影响, 难以始终获得全面的感知信息, 且感知范围有限, 无法满足自动驾驶中感知的安全要求。相比之下, 路端感知得益于更高的传感器位置, 受交通状况影响较小, 感知范围更广。最新的路端 BEV 技术[30-32] 通过整合摄像头、LiDAR 和 Radar 等多种传感器, 实现了对交通参与者的类别、位置、速度和姿态等信息的实时感知。本章将依次从纯视觉、纯点云和多模态融合 3 个方向对路端 BEV 感知技术展开讨论, 深入剖析各类方法的研究现状和技术原理, 如图 2 所示。

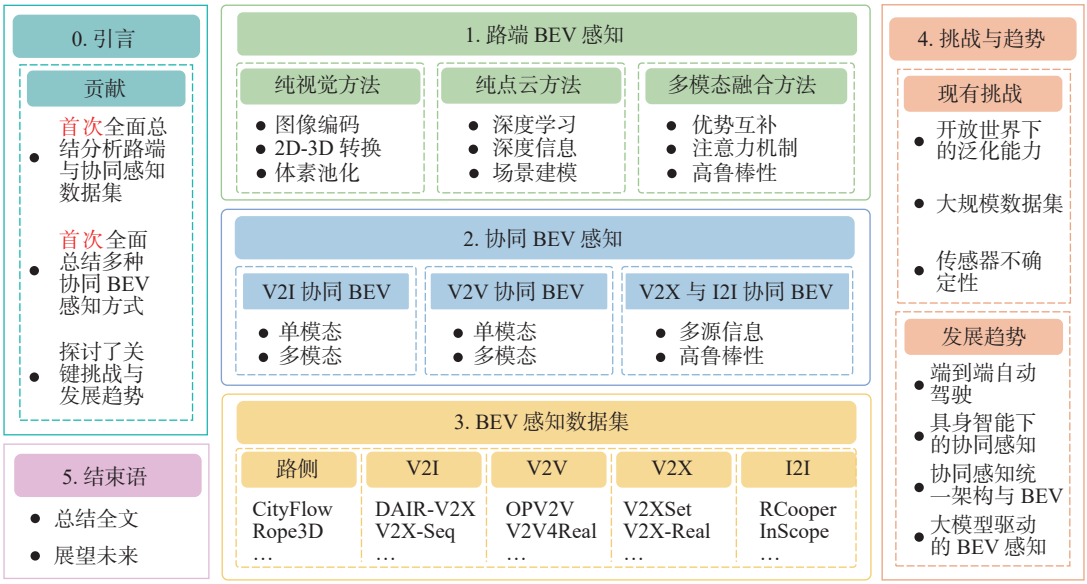

图 1　本文框架

Fig. 1　Framework for this paper

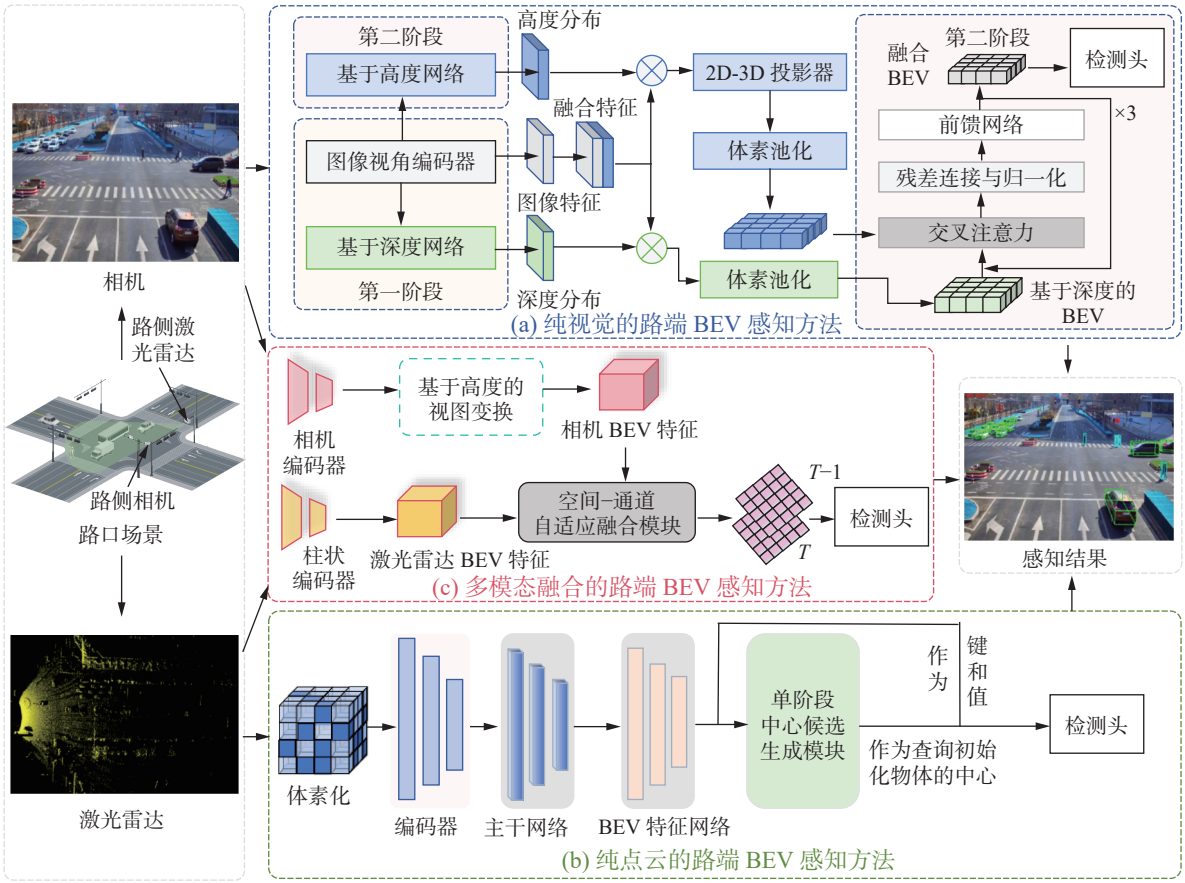

图 2　路端 BEV 感知技术方法

Fig. 2　Methodology for roadside BEV perception technology

1.1　纯视觉的路端 BEV 感知方法

城市路口的密集分布与交通流动态性，要求感知系统具备全局覆盖能力与高精度语义理解。视觉相机凭借其低成本、高分辨率及可应用于多任务处理的能力，成为构建大规模路端 BEV 感知网络的首选传感器，可在关键路段实现高密度、低成本的环境监控与实时感知。路端纯视觉解决方案通过路端基础设施获取图像和视频作为输入，最终通过图像编码、2D-3D 转换和体素池化的过程实现对象检测。

路端纯视觉感知方法可以缓解车端纯视觉感知易受遮挡、感知范围有限的问题，并且提升算法鲁棒性。但是，传统车端 BEV 感知系统依赖基于单目深度估计的 2D-3D 投影机制，其固有缺陷在于深度估计的歧义性随距离呈指数级增长。例如当目标距离超过 30 m 时，地面与车辆目标的相对高度差异迅速收敛至毫米级，导致投影误差急剧放大。

针对上述瓶颈，路端 BEV 感知方法通过固定安装高度引入强几何先验，衍生出多路径优化方法，如表 1 所示，其中，“🔗”表示论文提供的源码地址，“—”表示该论文暂未开源相关代码。BEVHeight[33] 创新性地构建像素级高度预测网络，通过解耦高度与深度估计有效缓解逆投影几何约束的敏感性。改进的 BEVHeight++[30] 在此基础上构建高度−深度联合优化目标函数，在 Rope3D 和 DAIR-V2X-I 数据集上大幅降低长距离目标定位误差。CoBEV[34] 支持端到端检测，提高了长距离场景和噪声相机干扰中的鲁棒性和场景、相机参数剧烈变化时的泛化能力。此外，BEVSpread[35] 基于距离和深度效应，提出了一种新的体素池化方法，在保证推理时间的同时降低了位置近似误差。为了解决各种相机安装位姿和标定噪声问题，CBR[36] 提出了一种免标定的 BEV 表示网络，有效实现了 3D 检测，而无需标定参数和额外的深度监督。针对目前缺乏路端多摄像头 BEV 解决方案的技术瓶颈，RopeBEV[37] 设计了 BEV Augmentation、CamMask、ROIMask 和 Camera Rotation Embedding，从 4 个方面改进了摄像头姿态多样、摄像头数量不确定性、感知区域稀疏、方向角模糊等带来的感知问题。

表 1　路端感知方法总结

Table 1　Summary of roadside perception methods

模态	方法	来源	感知任务	数据集	贡献	代码
C	BEVHeight++[30]	TPAMI 2025	目标检测	DAIR-V2X[91]	高度−深度联合优化目标函数	🔗
	BEVHeight[33]	CVPR 2023	目标检测	DAIR-V2X-I[91]	像素级高度预测	🔗
	CoBEV[34]	TIP 2024	目标检测	DAIR-V2X-I[91]	融合高度、深度特征端到端检测	🔗
	BEVSpread[35]	CVPR 2024	目标检测	DAIR-V2X-I[91]	兼顾距离、深度效应	🔗
	CBR[36]	IROS 2023	目标检测	DAIR-V2X[91]	无监督BEV表征	🔗
	RopeBEV[37]	arXiv 2024	车道分割	RoScenes[88]	优化多视角稀疏感知等问题	—
L	Center-Aware[31]	SUSTAINABIL 2023	目标检测	DAIR-V2X-I[91]	可变形交叉注意力	—
	DMD、CFTA[38]	JAT 2022	目标检测	PandaSet LiDAR	无监督强度−距离特征建模	—
	3D-DSF[39]	ITSC 2017	车道检测	自建数据集	三维点云优化、背景滤波	—
	Lin等[40]	JSEN 2021	车道检测	自建数据集	抗干扰地面识别优化	—
	Zhao等[41]	J.TRC 2019	检测跟踪	自建数据集	路端LiDAR全息感知系统	—
	Cui等[42]	MIS 2019	车辆跟踪	自建数据集	LiDAR-DSRC协同	—
	Wu等[43]	MITS 2018	车道检测	自建数据集	背景滤波驱动的弯道适应算法	—
C+L	BEVRoad[32]	EasyChair 2024	目标检测	DAIR-V2X-I[91]	融合感知	—
	HSRDet[46]	JSEN 2024	目标检测	DAIR-V2X[91]	高度感知场景重建、注意力机制	—
	Wang等[47]	TIM 2022	车辆跟踪	自建数据集	融合二维和三维轨迹信息	—

注：“C”表示相机，“L”表示激光雷达，“🔗”表示论文提供的源码地址，“—”表示该论文暂未开源相关代码。

综上，路端纯视觉感知通过高架或立柱式相机布局，有效克服了车端视觉易受遮挡和覆盖范围受限的先天瓶颈，特别契合城市复杂路口和大型交通枢纽场景。此方案不仅大幅拓展了 BEV 感知的空间视野，有效消减了遮挡造成的视觉盲点，还显著增强了目标检测与跟踪的精度与可靠性，从布局架构到性能提升层面实现了创新突破。

1.2　纯点云的路端 BEV 感知方法

尽管纯视觉技术已形成较为完整的解决方案，但 LiDAR 在三维空间感知方面仍展现出不可

替代的技术优势。相较于视觉传感器在深度感知方面的固有局限, LiDAR 通过主动式激光测距, 可获取厘米级精度的三维点云数据, 尤其是在复杂光照条件下, 其距离感知稳定性更为突出。然而受限于硬件成本与计算复杂度, 传统 LiDAR 方案曾面临规模化部署的挑战。近年来, 随着点云处理算法的突破性进展, 基于深度学习的三维目标检测技术有效提升了路端 LiDAR 系统的感知效能。

典型的路端点云处理流程可分为四大核心环节: 首先, 采用背景建模将动态目标与静态场景分离; 其次, 利用空间聚类算法对前景点云进行实例分割; 然后, 提取多维几何特征以支撑表征学习; 最后, 通过分类器对交通参与者进行精确识别与轨迹预测。鉴于纯点云 BEV 感知技术在路端场景的成熟应用程度尚不足够, 诸多研究聚焦于上述流程各环节, 分别提出针对性改进策略, 旨在提升整体感知性能。

表 1 给出了基于纯点云的路端感知方法, Zhang 等[38] 结合动态模态分解与粗细三角算法, 基于强度和距离信息自动提取背景特征, 而无需任何标注数据, 实现了场景建模的高效与鲁棒。Wu 等[39] 设计了一种特殊的三维点云格式及背景滤波方法, 既拓展了对车辆与行人的检测范围, 又显著降低了计算复杂度。针对低密度路端 LiDAR, Lin 等[40] 改进了地面识别算法以捕获更多的正常地面点, 从而减弱异常道路点对车道线提取的干扰。通过集成行人和车辆的存在状态、空间位置、运动速度及航向角度等实时信息, Zhao 等[41] 构建了一套用于基础设施级路端 LiDAR 数据处理与分析的综合系统, 可实现信息的深度融合与挖掘, 同时还具备强大的实时处理能力。Cui 等[42] 与 Wu 等[43] 分别通过基础设施数据关联和背景滤波优化, 进一步提升了车道检测精度。Shi 等[31] 提出基于可变形交叉注意力机制的中心感知检测器, 在路端点云感知任务中达到了领先性能。上述方法不仅验证了路端 LiDAR 的技术可行性, 更通过算法层面的创新推动 BEV 感知范式向更高精度、更强泛化方向发展。

虽然路端纯点云的 BEV 感知技术在实际应用场景中的规模化部署仍处于探索阶段, 但 LiDAR 在三维环境感知领域展现出的性能优势极为显著, 不仅能够突破视觉传感器在深度感知维度上的物理局限, 更在复杂光照干扰下的动态目标轨迹预测方面表现出更强的鲁棒性。

1.3　多模态融合的路端 BEV 感知方法

现阶段, 单模态感知体系在特定环境下面临着: 视觉感知系统对环境光照变化极为敏感, 其性能易受环境光照波动、色彩偏移等因素的干扰; 点云感知系统则受限于运动模糊效应与数据稀疏表征能力匮乏的制约, 难以全面捕捉目标物体的特征信息。综上原因, 融合可见光相机特征与 LiDAR 几何特征的多模态感知框架[44], 凭借其整合不同模态信息、优势互补的特性, 已经成为突破单一传感器性能极限、实现多源信息深度融合的关键路径。该范式通过对异构传感器数据进行精确的时空对齐与特征级融合[45], 可显著提升目标分类置信度与轨迹预测精度。该范式还具有较强的环境适应性, 即使是在雨雾、夜间低照度等极端工况下, 仍可为智能网联道路系统的可靠运行提供坚实的技术支撑。

跨模态融合感知技术通过异构传感器优势互补, 有效突破了单源感知系统的性能边界, 如表 1 所示。路端融合 BEV 感知将视觉语义信息与 LiDAR 几何特征进行多层次融合, 如 BEVRoad[32] 通过融合感知, 可以有效应对恶劣环境的干扰, 准确预测目标的速度和位置信息, 有效地缓解了目标遮挡问题。HSRDet[46] 结合高度感知场景重建和基于注意力的特征融合, 以生成强大的 BEV 表示。Wang 等[47] 通过引入融合注意力机制, 提出了一种融合二维和三维轨迹信息的新型跟踪方法, 以提高网络在速度计算、跟踪范围、物体损失率和断开轨迹修复率方面的性能。总而言之, 融合感知方案可以减少感知特征丢失, 提高检测鲁棒性, 将路端 BEV 感知能力和效果提升到了一个更高的水平, 减少了纯视觉和纯点云感知的局限性。

通过对视觉、LiDAR、Radar 等多源信息的时空对齐与特征级融合, BEV 多模态感知不仅能够在极端天气和弱光环境中保持高精度的目标检测与轨迹预测, 还将为智能交通信号优化、自动驾驶高精定位与辅助、智能安防应急响应、城市基础设施健康监测、智慧停车管理及交通能耗调度等多领域提供技术支撑[48], 推动车路协同与智慧城市建设迈向更高水平。

1.4　局限性和适用场景分析

基于路端的 BEV 感知可广泛应用于城市智慧路口、高速公路全天候监控、恶劣天气辅助驾驶等场景。然而, 视觉系统在透视图到 BEV 视图的转换过程中存在深度估计不确定性, 需依赖多视几何约束或先验知识补偿。虽然激光雷达能直接获取深度, 但其点云密度随距离呈指数衰减, 导致远距离目标细节丢失, 感知精度会衰减。现

有 BEV 模型的环境鲁棒性不足，在雨雾或强光条件下传感器会失效，且对训练数据未覆盖的交通异常情况适应能力有限。再者，路端 BEV 设备安装校准复杂，扩展性差，部署与维护成本较高。同时边缘计算设备难以满足实时 BEV 模型推理，算力与实时性存在瓶颈。此外，多模态同步也存在挑战，如摄像头与激光雷达时空未对齐会产生融合噪声。

2　基于协同的 BEV 感知方法

上一章系统梳理了路端 BEV 感知的核心技术路径，包括纯视觉、纯点云以及多模态融合等典型方法，展示了其在克服遮挡、提升感知精度等方面的显著优势。然而，单一的路端或车端感知系统在复杂动态环境下仍面临感知范围有限、视角受限等挑战。当传感器发生故障或数据异常时，系统的鲁棒性将显著下降。而协同感知技术通过多终端信息交互，融合来自不同视角的感知数据，能够有效扩展感知范围，降低盲区风险[49]，并提升目标检测精度和系统鲁棒性。因此，协同感知技术逐渐成为 BEV 感知发展的重要方向。本章将深入探讨基于车路协同 (vehicle-to-infrastructure, V2I)、车车协同 (vehicle-to-vehicle, V2V)、车与万物协同 (vehicle-to-everything, V2X) 和路侧基础设施间的协同 (infrastructure-to-infrastructure, I2I) 的 BEV 感知方法，分析其在多终端信息融合下的应用机制与研究进展。

从感知的输入模态角度来看，每类协同感知方法可以进一步细分为单模态和多模态融合两种感知方式。单模态感知侧重于同源传感器数据的融合，而多模态感知则通过整合异构传感器的互补信息，实现更全面的环境理解。常见的以 3D ODet 为主要任务的协同感知方法的具体信息如表 2 所示。

表 2　BEV 协同感知方法总结

Table 2　Summary of BEV cooperative perception methods

类型	方法	来源	模态	数据集	贡献	代码
V2I	VIMI[52]	arXiv 2023	C	DAIR-V2X-C	车路协同任务中间融合方法	—
	CenterCoop[54]	RA-L 2023	L	DAIR-V2X	基于中心的特征聚合框架	—
	3D HL[56]	TVT 2023	L	DAIR-V2X-I	3D谐波损失函数	—
	CoFormerNet[53]	Sensors 2024	L	DAIR-V2X, V2XSet	时间与空间信息的特征融合	—
	VICOD[60]	WCMEIM 2022	C, L	DAIR-V2X-C	区域建议网络和第二阶段检测网络	—
	V2I-BEVF[58]	ITSC 2023	C, L	DAIR-V2X-I	可变形注意力解码器	—
	CO^3[59]	ICLR 2023	C, L	Once, KITTI	协同对比学习与上下文形状预测	🔗
	MSMDFusion[61]	CVPR 2023	C, L	nuScenes	多深度反投影策略和GMA-Conv	🔗
	FFNet[62]	ICLR 2023	C, L	DAIR-V2X	自监督方法和特征流预测模块	🔗
V2V	CoBEVT[65]	CoRL 2022	C	OPV2V, nuScenes	稀疏注意力模块FAX	🔗
	CoCa3D[64]	arXiv 2023	C	DAIR-V2X, OPV2V+	核心通信高效协作技术	🔗
	TempCoBEV[66]	IV 2024	C	OPV2V	融合当前和历史 BEV 表征	🔗
	V2VNet[68]	ECCV 2020	L	V2V-Sim	空间感知图神经网络	—
	CoBEVFlow[67]	NIPS 2023	L	DAIR-V2X	运动补偿对齐代理感知信息	🔗
	LCRN-V2VAM[69]	TIV 2023	L	OPV2V	感知修复网络和 V2V 注意力	🔗
	HM-ViT[71]	ICCV 2023	C, L	OPV2V	异构3D图注意力	🔗
	MCoT[72]	ICPADS 2023	C, L	OPV2V	刚性对齐与交叉注意力	—
	CoBEVFusion[70]	DICTA 2024	C, L	OPV2V	双窗口交叉注意力	—
	V2VFormer++[73]	TITS 2024	C, L	OPV2V, V2X-Sim 2.0	动态通道融合与全局局部变换策略	—
V2X	BEV-V2X[76]	IV 2024	C	INTERACTION	多车BEV 融合和占用图预测	—
	V2X-BGN[77]	TIV 2023	L	V2X-Set	全局非最大值抑制和后期融合	—

注："C"表示相机，"L"表示激光雷达，"🔗"表示论文提供的源码地址，"—"表示该论文暂未开源相关代码。

2.1　路端和车端协同的 BEV 感知方法

车路协同技术通过车端与路端基础设施之间的信息交互，实现多源感知数据的融合处理，整合车端的局部高精度环境感知能力与路端的广域覆盖能力。与单端感知相比，车路协同不仅能够有效提升对遮挡区域、远距离目标等关键区域的

感知性能, 还能在复杂动态交通环境中增强系统对突发情况的响应能力与整体鲁棒性, 从而弥补传统感知方法在精度和视野范围上的不足[50-51]。

2.1.1 单模态融合

车路单模态融合技术聚焦于同源传感器数据的协同处理, 当前研究热点主要集中在视觉与点云数据的融合方法。在视觉融合方面, 主要挑战在于多视角图像的校准误差以及二维特征投影至三维空间过程中的信息损失。针对上述问题, VIMI 框架[52] 提出了一种动态图像特征增强机制, 通过多视图中间特征的协同融合, 有效提升了三维目标检测的精度与鲁棒性。

在点云融合方面, 代表性方法 CoFormerNet[53] 通过时空聚合模块与空间调制交叉注意力机制, 有效缓解了通信延迟与空间错位对融合性能的影响; CenterCoop[54] 采用中心化特征编码策略, 将局部上下文压缩为紧凑表示, 显著降低了通信带宽需求, 同时保留关键感知信息。此外, 3D Harmonic Loss[56] 引入谐波损失函数, 在点云域实现了检测精度与计算效率之间的良好平衡, 从而进一步优化了点云协同感知性能。

综上, V2I 场景下的单模态融合技术通过特征对齐优化与通信机制设计, 在扩展感知视野、提升鲁棒性方面取得了显著进展。然而, 由于感知能力依赖于单一传感器模态, 其性能仍受限于该模态固有的缺陷。相比之下, 多模态融合技术通过综合多传感器间互补信息, 具备更强的环境建模能力与冗余鲁棒性, 展现出突破协同感知性能瓶颈的广阔前景。

2.1.2 多模态融合

多模态融合技术通过整合车端与路端的视觉、激光雷达等异构感知数据, 为环境感知提供更为丰富的信息。这种方法有效弥补了单一传感器的局限性, 提升了环境理解的准确性, 帮助构建更全面的环境表征, 从而增强了自动驾驶系统在复杂场景中的感知能力[55,57,63]。

V2I-BEVF[58] 提出了双分支特征提取架构, 分别处理 2D 图像与 3D 点云数据, 并利用可变形注意力机制实现 BEV 特征对齐, 显著提升了交通参与者的实时检测精度。CO^3[59] 基于协同对比学习与上下文形状预测, 采用无监督方式学习点云的三维表示, 增强了模型在跨设备应用中的泛化能力。VICOD[60] 通过检测框融合网络优化车端与路端检测结果的匹配, 扩展了感知范围的同时降低了误检率。MSMDFusion[61] 采用多深度种子投影与门控模态感知卷积, 实现了多尺度特征融合, 在 3D 对象检测和跟踪任务中取得了最先进的结果。FFNet[62] 利用时序相干性预测未来特征, 通过特征流补偿通信延迟, 即使在时间不同步超过 200 ms 的场景中仍能保持稳定的检测能力。

多模态融合通过利用异构数据的互补性, 将 2D 图像感知与 3D 点云感知有效结合, 弥补了单一传感器的局限性, 减少了误差影响。这种融合方式在复杂交通场景中显著提升了感知精度和鲁棒性, 使自动驾驶系统能够更准确地理解环境变化, 从而提高了系统的安全性与可靠性。

2.2 车端和车端协同 BEV 感知方法

V2V 技术通过车辆间实时通信共享感知信息, 提升无路端基础设施场景下的感知能力。相比静态路端设备, V2V 适用于远距离目标、遮挡区域及高速目标的检测与跟踪, 扩展感知范围并增强系统在复杂交通环境下的鲁棒性。

2.2.1 单模态融合

V2V 单模态融合聚焦同源传感器数据的协同优化, 常见的有视觉信息与点云信息融合。CoCa3D[64] 提出多智能体协作框架, 通过深度估计优化与通信提升, 解决遮挡与远距离检测问题; CoBEVT[65] 基于轴向注意力 Transformer, 实现多车 BEV 语义分割的全局特征融合; TempCoBEV[66] 引入历史时序线索, 增强 BEV 地图分割的稳定性, 在通信中断场景下表现优异。在点云融合中, CoBEVFlow[67] 通过运动补偿对齐异步点云数据, 缓解时空错位问题; V2VNet[68] 基于图神经网络聚合多车信息, 优化预测能力; LCRN-V2VAM 框架[69] 通过修复网络与不确定性感知注意力机制, 提升检测鲁棒性。单模态融合通过注意力机制与时空对齐策略, 有效降低语义歧义与通信延迟影响, 但单传感器协作难以覆盖多车协同的全场景需求, 需结合多模态数据增强环境表征能力。

2.2.2 多模态融合

V2V 多模态融合整合视觉、激光雷达等异构数据, 构建鲁棒的协同感知框架。CoBEVFusion[70] 提出双窗口交叉注意力模型, 通过 3D CNN(convolutional neural network) 聚合多车 LiDAR-相机特征, 优化 BEV 语义分割与 3D 检测性能。HM-ViT[71] 基于异构模态 Transformer 框架, 支持动态异构交通场景下的灵活协作, 显著提升泛化能力。MCoT[72] 采用刚性关联对齐 BEV 特征, 结合交叉注意力机制实现相机与 LiDAR 数据的软融合, 增强特征互补性。V2VFormer++[73] 通过全局−局部 Transformer 模块融合多车 BEV 地图[74], 结合动态通道融合策略, 实现高效多模态特征聚

合。GraphBEV[75] 通过图结构建模邻域深度关系，并联合局部−全局对齐策略，减少了多模态 BEV 融合中的标定误差。多模态融合通过跨模态特征交互与高效融合架构设计，突破单一传感器物理限制，为复杂动态场景提供高精度感知支持。

2.3　车端和万物以及路端和路端协同 BEV 感知方法

V2X 与 I2I 协同 BEV 感知通过车与基础设施的多源信息交互，提供更全面、精准的环境理解，提升自动驾驶系统的稳定性与鲁棒性。如图 3 所示，V2X 通信包括多种子类型：V2V、V2I、V2P (vehicle-to-pedestrian)、V2N(vehicle-to-network)、I2I、V2X 等。这些子模块共同构成了复杂的协同感知网络系统，用于支持多种自动驾驶和智能交通场景。

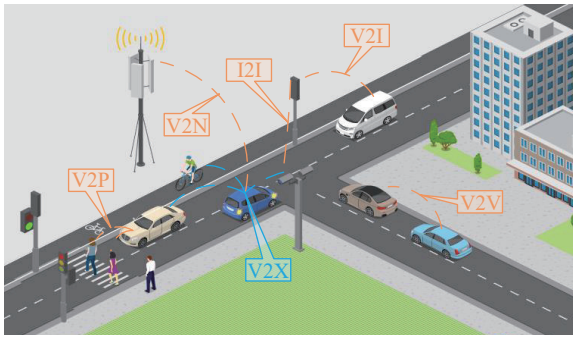

图 3　V2X 通信模式示意

Fig. 3　Schematic of V2X communication mode

近年来，多种协同 BEV 感知框架相继提出。BEV-V2X[76] 利用云端或路端融合多车 BEV 视图，优化全局占用图预测。V2X-BGN[77] 结合全局非极大值抑制与后融合技术，在遮挡场景下表现优越。H-V2X[78] 和 InScope[79] 数据集填补了高速公路与 I2I 遮挡评估领域的空白。V2X 协同 BEV 感知仍面临遮挡、通信延迟、远距协作等挑战，未来应重点发展多模态融合与目标距离自适应的动态协作机制[80]，以提升系统在复杂环境下的实时性、泛化能力与协作效率。

2.4　总结与分析

尽管当前协同 BEV 感知在多终端信息融合、感知盲区补偿与系统鲁棒性提升方面取得了一定进展，但在实际部署过程中仍面临若干关键挑战。首先，时空对齐误差问题依然存在。由于协同主体间存在异步采样、网络延迟与视角偏移等因素干扰，协同系统难以实现高精度、低误差的 BEV 特征融合。尽管已有如 CoBEVFlow[67]、FFNet[62] 等方法引入运动补偿与时序建模机制以缓解该问题，但在高动态、多目标密集的交通场景中，其稳定性与实时性仍难以保障。其次，多模态融合机制尚缺乏统一标准。现有方法多聚焦于 V2I、V2V 等特定协作模式和 3D 检测、语义分割等单一任务，在通用 BEV 表示构建、模态选择与融合策略设计方面仍缺乏规范化流程，限制了跨模态、跨平台的协同部署与扩展能力。此外，协同方法的适用性在复杂场景中仍然不足。当前大多数方法基于有限规模数据集进行验证，尚缺乏对施工区域、多车道拥堵等复杂道路结构下协同感知性能的系统性评估，泛化能力仍待提升。综上所述，未来协同 BEV 感知的发展应聚焦于构建鲁棒的异步对齐机制、形成统一的多模态融合表示框架，并提升算法在真实复杂环境下的泛化与部署能力，以支撑高可靠、高一致性的 BEV 感知体系在车路协同自动驾驶中的落地应用。

3　感知数据集

第二章重点探讨了协同 BEV 感知在多种通信模式与融合方式下的关键技术与典型应用，展示了其在拓展感知范围、提升目标检测精度及增强系统鲁棒性方面的显著优势。然而，要推动相关技术持续发展并实现有效对比评估，标准化、高质量的数据集是不可或缺的基础支撑。

当前，BEV 感知涉及的应用场景日益丰富，包括路端 (infrastructure, I)、V2I、V2V、以及 I2I 等多种形态。不同场景下，感知系统在传感器配置、任务类型、数据标注方式等方面存在显著差异，这直接影响到 BEV 感知方法的设计选择、性能上限与适用范围。因此，有必要对已有数据集进行系统整理与分类分析，以构建统一的研究基准，推动该领域模型性能的可比性与可复现性[81]。

本章将围绕 BEV 感知中的数据资源展开，系统梳理与分析典型的数据集，内容结构如下：首先介绍路端感知数据集；随后依次探讨 V2I、V2V、V2X, 以及 I2I 场景下的数据集资源；其次，进一步探讨 BEV 感知任务中常用的算法评估指标体系；最后，总结当前主流数据集在标注策略与标签类型方面的构成，为后续研究提供统一的基准支撑与数据方向指引。

3.1　路端数据集

路端感知通过固定部署的路端单元 (roadside unit, RSU) 实现高精度、大范围的环境感知，与车端多模态感数据集[82] 相比，路端感知数据集在遮挡和远距目标检测中表现出显著优势，成为 BEV 感知体系的重要组成部分。表 3 总结了主要路端感知数据集。

表 3　协同感知数据集统计

Table 3　Statistics of cooperative perception datasets

场景	数据集	年份	期刊/会议	来源	传感器	任务	图像数/10^3	点云数/10^3	3D框数/10^3	地址
I	CityFlow[83]	2019	CVPR	实际	C	MTMCT, ReID	118.0	—	22.9	
	INTERACTION[84]	2019	IROS	实际	C, L	ODet, TP	1 400.0	—	1 400.0	
	Coop3DInf[90]	2020	TITS	仿真	C	ODet	10.0	—	121.2	
	A9-Dataset[87]	2022	IV	实际	C, L	ODet	1.1	1.1	14.4	
	IPS300+[85]	2022	ICRA	实际	C, L	ODet	56.7	14.2	4 500.0	
	Rope3D[89]	2022	CVPR	实际	C, L	ODet	50.0	—	1 500.0	
	TUMTraf-I[86]	2023	ITSC	实际	C, L	ODet	4.8	4.8	57.4	
	RoScenes[88]	2024	ECCV	实际	C	ODet	1 300.0	—	21 130.0	
	H-V2X[78]	2024	ECCV	实际	C, R	BEVDet, MOT, TP	1 940.0	—	—	
V2I	DAIR-V2X-C[91]	2022	CVPR	实际	C, L	ODet	39.0	39.0	464.0	
	V2X-Seq[92]	2023	CVPR	实际	C, L	ODet, MOT, TP	71.0	15.0	464.0	
	HoloVIC[93]	2024	CVPR	实际	C, L	ODet, MOT	100.0	100.0	11 470.0	
	OTVIC[94]	2024	IROS	实际	C, L	ODet	15.0	15.0	24.4	
	DAIR-V2XReid[95]	2024	TITS	实际	C, L	ODet, ReID	2.5	—	—	
	TUMTraf V2X[96]	2024	CVPR	实际	C, L	ODet, MOT	5.0	2.0	30.0	
	V2X-Radar[97]	2024	arXiv	实际	C, L, R	ODet	40.0	20.0	350.0	
V2V	OPV2V[98]	2022	ICRA	仿真	C, L, R	ODet, MOT, SS	44.0	11.4	232.9	
	OPV2V+[62]	2023	CVPR	仿真	C, L, R	ODet	11.4	11.4	232.9	
	V2V4Real[100]	2023	CVPR	实际	C, L	ODet, MOT, S2R	40.0	20.0	240.0	
	MARS[102]	2024	CVPR	实际	C, L	MAP, UODet	15.0	15.0	—	
	OPV2V-H[99]	2024	ICLR	仿真	C, L, R	ODet	79.0	79.0	232.9	
	V2V-QA[101]	2025	arXiv	实际	C, L	ODet, PQA	—	18.0	—	
V2X	V2X-Sim 2.0[103]	2022	RA-L	仿真	C, L	ODet, MOT, SS	60.0	10.0	26.6	
	V2XSet[106]	2022	ECCV	仿真	C, L	Odet	44.0	11.4	233.0	
	DOLPHINS[104]	2022	ACCV	仿真	C, L	ODet	42.3	42.3	292.5	
	DeepAccident[108]	2024	AAAI	仿真	C, L	ODet, MOT, SS, MP	—	57.0	285.0	
	V2X-Real[110]	2024	ECCV	实际	C, L	ODet	171.0	33.0	1 200.0	
	Multi-V2X[109]	2024	arXiv	仿真	C, L	ODet, MOT	549.0	146.0	4 200.0	
	Adver-City[105]	2024	arXiv	仿真	C, L	ODet, MOT, SS	24.0	24.0	890.0	
	V2X-Traj[112]	2024	NIPS	实际	C, L	TP	808.0	808.0	1 400.0	
	V2X-R[107]	2024	arXiv	仿真	C, L, R	Odet	150.9	37.7	170.8	
	V2XPnP[111]	2024	arXiv	实际	C, L	PnP, TP	208.0	40.0	1 450.0	
	Mixed Signals[113]	2025	arXiv	实际	L	ODet	—	45.1	240.6	
I2I	Rcooper[114]	2024	CVPR	实际	C, L	ODet, MOT	50.0	30.0	30.0	
	InScope[79]	2024	arXiv	实际	L	ODet, MOT	—	21.3	188.0	

注: 模态——毫米波雷达(radar, R); 任务——多智能体感知(MAP)、多目标跟踪(multi-object tracking, MOT)、轨迹预测(trajectory prediction, TP)、多目标多摄像头跟踪(multi-target multi-camera tracking, MTMCT)、规划问答(planning and question answering, PQA)、感知与预测(perception and prediction, PnP)、重识别(re-identification, ReID)、语义分割(semantic segmentation, SS)、无监督目标发现(unsupervised object discovery, UOD)、传感器到真实域自适应(sensor-to-real domain adaptation, S2R)、鸟瞰图检测(bird's eye view detection, BEVD); "🔗"表示论文提供的源码地址, "—"表示该论文暂未开源相关代码。

纯路端数据集基于交叉路口的真实基础设施, 能有效捕捉城市交通环境的动态特征。早期的 CityFlow[83] 为路端感知的初步尝试, 但未满足三维感知需求。为提升适用性, INTERACTION[84] 融合无人机与道路摄像头采集, 广泛应用于 BEV 交互分析与临界事件建模。IPS300+[85] 提供完整

的 3DBEV 视图，但对大型车辆识别支持不足。TUMTraf-I[86] 引入多传感器与多时段采集，增强在恶劣条件下的 BEV 感知能力。

在高速公路场景中，TUMTraf-A9[87] 扩展了道路与天气种类，但数据标注有限。RoScenes[88] 覆盖面广，模拟了拥堵状态。H-V2X[78] 融合雷达与摄像头，适用于多模态 BEV 感知研究。部分数据集探索了新技术路径，如 Rope3D[89] 专注单目三维检测，评估图像到三维空间的推理能力。Coop3-DInf[90] 通过仿真生成交叉口场景，模拟点云，但在真实感与通用性上仍有限。纯路端数据集依托固定 RSU，具备稳定视角与高精度传感器配置，适合长时间、多时段采集，适用于 BEV 感知预测任务。

3.2　车端和路端协同的数据集

V2I 数据集融合路端与车载多模态传感器数据，支持复杂交通场景和昼夜等多变环境下的 BEV 协同感知，显著提升系统在多智能体协同任务中的鲁棒性和准确性，表 3 总结了主要的 V2I 数据集。

DAIR-V2X-C[91] 是首个面向车路协作的大规模真实数据集，提供了多模态传感器在不同融合阶段的 BEV 感知基准。V2X-Seq[92] 在 DAIR-V2X 基础上扩展了 BEV 感知任务维度，捕捉了丰富的轨迹信息。HoloVIC[93] 构建了不同布局的全息交叉路口，通过多传感器融合克服遮挡和盲点。OTVIC[94] 针对高速和嘈杂同虚拟场景，提供多模态，多视角数据。DAIR-V2XReID[95] 通过车路端目标标签关联，实现 BEV 场景下的车辆重识别，TUMTraf-V2X[96] 支持跨模态深度融合，为 BEV 感知提供了更贴近实际的基准。V2X-Radar[97] 引入 4D 雷达，增强恶劣天气下的鲁棒性。

V2I 数据集具备高精度和广泛覆盖，适用于动态场景，在遮挡检测、路径预测等任务中表现出色。未来应扩大数据规模，增加极端天气和道路类型，采用虚实结合的训练方式。

3.3　车端和车端协同的数据集

V2V 通过多车信息共享提升了复杂交通环境中的感知范围、鲁棒性和准确性。相关数据集通常具备多车同步、多模态融合与复杂场景覆盖，支持 BEV 感知任务，如表 3 所示。本节从仿真与真实角度阐述 V2V 数据集。

3.3.1　仿真数据集

仿真 V2V 数据集因灵活性高、成本低、易扩展，成为协同感知算法验证的重要平台。OPV2V 系列[64][98-99] 目前成为主流的 V2V 数据集，OPV2V[98] 提供了 16 种 BEV 融合算法的基准测试，非常适合多模态融合的 3D Det 方法评估。OPV2V+[64] 通过增加智能体数量，提升协同感知能力；OPV2V-H[99] 引入数据异构性和新智能体，增强激光雷达与图像模式间的融合适应性。仿真 V2V 数据集具有高可控性和可扩展性，但域差异限制其泛化能力。未来需在模态融合、任务覆盖与真实迁移等方面完善，推动多车感知向更真实、多样的方向演进。

3.3.2　真实数据集

真实 V2V 数据集提供真实场景下的车辆交互数据，支持 BEV 场景下多模态数据融合。V2V4-Real[100] 是首个支持三维检测与跟踪的多模态大规模数据集，推动了 BEV 跨域感知研究。V2V-QA[101] 数据集在此基础上引入大语言模型，支持车辆间问答任务。MARS[102] 数据集扩展至四车协作，提供丰富的多模态轨迹信息。

真实 V2V 数据集具有高精度的实地数据，能够真实反映协同感知中的交互与动态变化，但数据采集成本高、场景有限。未来应加强多样性、任务拓展与仿真和真实融合，推动 V2V 感知技术落地。

3.4　车端和万物协同的数据集

V2X 感知融合 V2I 和 V2V 协作，融合多模态传感器与交通环境信息，提升 BEV 感知任务的精度与时序一致性，扩展感知范围，增强动态环境理解能力。如表 3 所示，本节梳理了仿真与真实两类数据集的现状与特点。

3.4.1　仿真数据集

仿真 V2X 数据集具备高灵活性、低成本与可控性，适用于复杂场景与极端天气模拟。V2X-Sim 2.0[103] 是首个面向多任务的仿真数据集，但目标类别与传感器配置仍有不足。DOLPHINS[104] 引入三车协作机制与多种恶劣天气。Adver-City[105] 扩展至六类极端气象，模拟强光环境。V2XSet[106] 融合真实噪声模拟，构建多模态多智能体仿真场景。V2X-R[107] 整合相机、激光雷达与 4D 雷达，应对恶劣气候下的感知退化问题。DeepAccident[108] 专注交叉口 BEV 感知事故预测。Multi-V2X[109] 探讨了不同渗透率下的智能体调度策略，支持多样化智能体的协同感知。

仿真 V2X 数据集覆盖长尾事件与复杂场景，但实际部署仍有局限。未来需提升可信度和适应性，推动标准化评估与算法开发。

3.4.2　真实数据集

真实 V2X 数据集记录了 V2V 与 V2I 等多类

型的感知信息, 涵盖多样交通场景与驾驶行为, 支持 BEV 动态环境与不确定性建模。V2X-Real[110] 整合多类型车辆与基础设施, 包含大量 3D 标签, 推动多模态 BEV 感知发展。V2XPnP[111] 与 V2X-Traj[112] 提供多智能体轨迹建模, 提升 BEV 任务复杂度。Mixed Signal[113] 数据集采用异构激光雷达配置, 反映特殊交通特征。真实 V2X 数据集在 BEV 感知融合、传感器同步与交互行为分析中具有关键价值, 但受限于标注复杂性与区域交通行为差异。未来应兼顾数据真实性与仿真灵活性, 推动标准化与共享, 提升对长尾场景的覆盖。

3.5　路端和路端协同的数据集

现有路端数据集多聚焦单一 RSU, 难以支撑多 RSU 协同感知研究。表 3 总结了现有的 I2I 数据集, RCooper[114] 数据集覆盖交叉口与走廊场景, 支持 BEV 协作策略与融合算法研究。InScope[79] 聚焦多 RSU 协作, 通过雷达实现盲区覆盖, 设有 4 项 BEV 基准任务。I2I 数据集提升盲区覆盖与感知连续性, 适用于异构传感器融合与 BEV 协同策略研究。

未来, 路端感知数据构建将注重融合性、高覆盖率与复杂场景建模, 结合真实与仿真手段, 推动多任务学习与动态协作机制, 为智能交通体系发展提供支撑。

3.6　BEV 感知任务算法的评估指标

在自动驾驶系统中, BEV 作为环境建模的关键方法, 其评估体系已从单一精度指标拓展为融合鲁棒性、安全性与可信性的多维框架。

三维目标检测任务通常采用平均精度 (average precision, AP) 与平均交并比 (mean intersection over union, mIoU) 进行评估。对于包含 C 个类别的检测任务, mIoU 定义为

$$I_{\mathrm{oU}}^{\mathrm{m}}=\frac{1}{C}\sum_{c=1}^{C}I_{\mathrm{oU}}^{c}=\frac{1}{C}\sum_{c=1}^{C}\frac{N_{\mathrm{TP}}^{c}}{N_{\mathrm{TP}}^{c}+N_{\mathrm{FP}}^{c}+N_{\mathrm{FN}}^{c}}$$

式中 N_{TP}^{c}、N_{FP}^{c}、N_{FN}^{c} 分别代表类别 C 的真阳性、假阳性与假阴性检测数。OPV2V[98] 采用多阈值 mIoU 下的 AP, DAIR-V2X[91] 则引入通信成本衡量延迟鲁棒性。语义分割任务常用 mIoU 与平均像素精度 (mean pixel accuracy, mPA) 指标评估。DeepAccident[108]、DOLPHINS[104] 和 Adver-City[105] 通过使用 mPA 结合不同阈值的 mIoU 来评估图像和点云分割的鲁棒性。在事故预测任务中, DeepAccident[108] 引入体积预测质量 (volume prediction quality, VPQ) 和事故预测准确性 (accident prediction accuracy, APA) 细化事故预测误差来源。VPQ 计算公式为

$$I_{\mathrm{VPQ}}=\frac{1}{N}\sum_{i=1}^{N}\frac{\left|V_{\mathrm{pred}}^{i}/N_{\mathrm{gt}}^{i}\right|}{\left|V_{\mathrm{pred}}^{i}\cup V_{\mathrm{gt}}^{i}\right|}$$

它将预测时间范围划分为 N 个窗口, 计算每个窗口内预测事故体积 V_{pred}^{i} 与真实事故体积 V_{gt}^{i} 的 3D IoU, 并求平均。APA 计算公式为

$$I_{\mathrm{APA}}=\frac{N_{\mathrm{TP}}+N_{\mathrm{TN}}}{N_{\mathrm{TP}}+N_{\mathrm{TN}}+N_{\mathrm{FP}}+N_{\mathrm{FN}}}$$

通过标准的二分类准确率来评估模型的事故情况, N_{TP}、N_{TN}、N_{FP}、N_{FN} 分别代表真正例、真负例、假正例与假负例的数量。在轨迹预测任务中主要通过平均位移误差 (average displacement error, ADE)、最终位移误差 (final displacement error, FDE) 进行评估。对于 N 条预测轨迹, 其 ADE 与 FDE 计算公式分别为

$$I_{\mathrm{ADE}}=\frac{1}{N\cdot T}\sum_{n=1}^{N}\sum_{t=1}^{T}\|\hat{p}_{t}^{(n)}-p_{t}^{(n)}\|_{2}$$

$$I_{\mathrm{FDE}}=\frac{1}{N}\sum_{n=1}^{N}\|\hat{p}_{T}^{(n)}-p_{T}^{(n)}\|_{2}$$

式中 $\hat{p}_{t}^{(n)}$ 和 $p_{t}^{(n)}$ 分别表示第 n 条轨迹在第 t 个时间步的预测位置与真实位置, T 为轨迹总时间步长。V2X-Traj[112] 和 V2X-Seq[92] 还引入轨迹缺失率 (missing rate, MR) 准确评估了预测的准确性和鲁棒性, 定义为预测轨迹终点误差超过设定阈值 δ 的比例:

$$I_{\mathrm{MR}}=\frac{1}{N}\sum_{n=1}^{N}\mathbb{I}\left(\|\hat{p}_{T}^{(n)}-p_{T}^{(n)}\|_{2}>\delta\right)$$

式中 $\mathbb{I}(\cdot)$ 为指示函数。对于跟踪任务的评价, 代表指标包括多目标跟踪准确率 (multiple object tracking accuracy, MOTA)、精确率 (multiple object tracking precision, MOTP), V2X-Seq[92] 在此基础上引入身份切换次数 (identity switches, IDS) 和每秒传输的字节数 (bytes per second, BPS), 全面评估跟踪效果的鲁棒性和稳定性。

总体而言, BEV 感知系统的评价指标正向多任务、多尺度与多维性能融合方向演进, 为自动驾驶模型的稳健性与实用性提供了系统化评估依据。

3.7　数据集标注

3.7.1　标注策略

主流 BEV 感知数据集标注策略可分为自动标注与人工标注两类。自动标注依托传感器配准与轨迹推理生成高质量标签, 适用于仿真或大规模采集数据; 人工标注则多用于关键帧和目标行为, 确保标注精度, 常见于真实数据集。部分数据集还采用多传感器互补校正、弱监督或大模型

辅助标注策略，提升标签一致性与标注效率。

3.7.2　标注类型

目前 BEV 感知相关数据集中采用的标注类型主要包括以下几类：三维边界框是最常见的形式，用于描述车辆、行人等交通参与者的空间位置与类别标签，广泛应用于如 DAIR-V2X[91]、V2X-Seq[92]、V2V4Real[100] 等数据集中；轨迹标注记录目标在连续帧中的时空坐标与身份 ID，支撑运动预测与多目标跟踪任务，典型数据集有 V2XPnP[111]、V2X-Traj[112] 和 MARS[102]；占用图标注则以 BEV 栅格或语义地图形式表达场景中空间占据情况，适用于场景建模与语义预测，代表数据集如 INTERACTION[84]；此外，部分数据集还包含语义分割标签如 OPV2V[98] 和 Adver-City[105]，用于细粒度环境理解；重识别标签用于跨视角或跨设备目标一致性匹配，如 DAIR-V2XReID[95] 和 CityFlow[83]；在更高层次的任务中，也有数据集提供行为与决策标签，用于支撑轨迹规划与行为预测，代表性数据集包括 DeepAccident[108] 与 V2V-QA[101] 等。

这些标签形式为 BEV 感知系统的多任务训练提供支撑，覆盖检测、分割、跟踪、预测、识别等多类任务，推动了多模态、多智能体、多任务融合的研究发展。

4　现有挑战和发展趋势

本节梳理 BEV 感知算法面临的挑战与发展趋势。第 4.1 节探讨 BEV 感知算法对开放世界对象类别的泛化能力、大规模无监督数据以及传感器不确定性对性能的影响。第 4.2 节讨论端到端自动驾驶、具身智能、协同感知架构和大模型与 BEV 架构的关系。

4.1　挑战

随着自动驾驶技术进展，BEV 感知作为环境理解的重要方法，正面临开放世界场景下的诸多挑战。BEV 感知系统不仅需具备对新颖类别的识别与适应能力，还需在缺乏标注数据的条件下实现高效学习，并能够应对传感器噪声、性能退化等带来的不确定性，保持稳定、可靠的感知性能[115]。上述问题限制了现有 BEV 感知模型在真实复杂场景中的泛化能力与可扩展性，也对其在自动驾驶实际部署中的鲁棒性提出了更高要求。围绕上述问题，当前研究面临的主要挑战可归纳为 3 个方面：

1) 在实际应用中，自动驾驶车辆不可避免地遇到训练阶段未见的新类别或长尾目标。由于当前大多数 BEV 感知系统构建于封闭集假设之上，其语义识别能力受限于已有标注空间，缺乏对未知类别或未标注对象的认知机制。此外，场景中对象的类别分布往往具有高度不平衡性，进一步加剧模型对长尾类或新颖类的感知偏差。这种类别空间的开放性使得系统在运行中存在认知盲区，显著削弱其泛化能力与安全性保障。

2) 尽管车载系统每天可采集海量多模态原始数据，但现阶段 BEV 感知模型的训练仍严重依赖人工精标的大规模数据集。而高质量标注数据昂贵、稀缺，尤其在边缘场景、极端天气或新地域中难以覆盖，导致模型难以适应数据分布变化，泛化能力受限。同时，多模态感知数据间存在视角差异、语义不一致等跨模态差异，进一步增加了数据利用的复杂性，造成“数据足够、监督不足”的结构性矛盾。

3) 在真实世界中，传感器性能易受环境因素干扰，如强光、雨雪、遮挡、模糊等都会影响图像与点云质量，导致感知输入在时空维度产生漂移。此外，车路协同系统中的通信延迟、丢包或同步误差也可能引发感知不一致现象。更重要的是，随着多模态融合的发展，不同模态间的可靠性差异成为系统稳定性的隐患，错误信息在感知链中易被放大，进而影响 BEV 表征的空间一致性和下游决策可靠性。

4.2　发展趋势

随着自动驾驶系统对协同感知与智能决策能力的需求不断提升，传统模块化架构逐渐暴露出响应滞后、任务耦合松散等问题。相比之下，端到端方法通过将传感器输入直接映射为控制输出，有效简化了感知—决策流程，提高了系统的整体效率，成为协同感知研究的重要方向之一。尽管当前 BEV 感知在多模态融合和空间统一表达方面取得了显著进展，但在动态交互建模与时空语义理解方面仍存在不足。而具身智能通过感知—决策—控制的闭环反馈[116-117]，为系统带来了更强的情境适应能力。此外，打破感知、通信、决策等模块边界，构建跨模态协同的统一架构，也正成为协同自动驾驶系统演进的关键趋势。当前 BEV 感知与协同自动驾驶的发展趋势主要体现在 4 个方面：

1) 协同场景下的端到端自动驾驶变体。现有端到端自动驾驶方法多采用 BEV 表征结构，融合自车运动状态、地图先验与任务指令等信息，以增强整体场景理解能力。然而，该类模型仍存在收敛缓慢、行为不可控、泛化能力弱等问题。为此，研究引入稀疏航点与候选轨迹作为辅助监督

信号, 增强策略的可解释性与训练稳定性。同时, 结合闭环强化学习机制, 系统可基于环境反馈持续优化决策策略, 从而提升在长尾与异常场景下的鲁棒性与反应能力。由于极端场景数据稀缺且采集成本高昂, 神经辐射场 (neural radiance fields, NeRF) 与扩散模型等生成式技术也被广泛用于合成高保真多模态感知数据, 辅助模型训练, 提升系统的泛化性能。

2) 具身智能具有高效的协同场景理解能力。尽管 BEV 感知在统一表达环境几何与语义信息方面优势显著, 但对于动态变化与交互建模仍显不足。具身智能通过感知—决策—行动的闭环机制, 使系统具备情境感知与动态适应能力, 为 BEV 感知注入更强的交互推理能力。其集成方式主要包括: 一是基于 BEV 表征进行场景理解与意图预测, 预判目标行为与潜在风险; 二是结合实时感知结果进行策略生成, 实现精细化的变道、避障等动态决策; 三是开展集体智能交互建模, 模拟交通参与者间的隐性沟通行为, 提升整体协作效率与交通安全性。

3) 协同统一架构与 BEV 协同感知的融合发展。 面对高动态、多参与体、任务复杂的协同驾驶场景, 单一 BEV 感知模块难以独立应对所有任务需求。协同统一架构强调打破传统“感知—决策—控制”的链式结构, 转向融合式的跨模态协同机制, 通过共享状态空间与任务驱动协同感知, 实现系统级的信息闭环与智能决策。BEV 感知作为统一表达多模态数据的重要支撑, 可为统一架构提供清晰、标准化的空间基础, 提升系统对时空动态场景的响应能力与适应性。两者的深度融合, 有望构建更高效、更稳健的协同自动驾驶系统。

4) 近年来, 大规模视觉语言模型[118](如 GPT-4V、SAM、SEEM、DriveLM 等) 在认知理解与跨模态推理方面展现出强大能力, 为 BEV 感知系统引入了全新的发展路径。现有方法如 BEV-CLIP 通过将 CLIP 的语义先验引入 BEV 空间, 实现了类开放词汇的语义感知, 为多模态融合开辟了方向。在此基础上, 未来可进一步探索基于大模型的跨模态 BEV 表征生成机制, 实现从图像−文本−点云等多源信息直接构建结构化 BEV 表征; 同时, 结合指令驱动的交互式感知范式, 推动 BEV 感知由被动建图向主动理解演进。此外, 可借助大模型丰富的知识库和上下文推理能力, 设计零样本泛化机制以支持长尾目标识别与决策, 并通过知识蒸馏与剪枝等方式部署轻量化模型于边缘设备, 构建“中心−边缘”协同的大模型 BEV 感知框架。这一方向不仅融合了语言理解与空间建模, 更为 BEV 感知系统在开放世界中的通用性与智能性提供了新范式。

5　结束语

本文系统性地梳理了 BEV 感知在多模态、多场景与多协同条件下的发展脉络, 首次构建了涵盖路端感知、车路协同感知与车车协同感知的三维技术框架。通过对典型方法和公开数据集的综合分析, 本文揭示了 BEV 感知在统一空间表达、多源信息融合及跨智能体协同处理方面的独特优势, 特别是在复杂交通环境中, BEV 感知显著提升了自动驾驶系统的感知精度与鲁棒性。

在理论层面, 本文拓展了 BEV 感知的技术体系, 深入剖析其在多模态融合、开放环境适应性和协同感知机制中的关键原理; 在应用层面, 本文对现有方法进行了分类总结与性能比较, 为未来自动驾驶感知系统的高精度、高可靠性与高安全性提供了坚实的技术支撑。

此外, 本文还前瞻性地探讨了 BEV 感知在具身智能、大模型与端到端架构中的潜力与挑战, 为后续研究提供了全新的视角。目前, BEV 感知仍面临若干亟待突破的难题, 如新颖类别的识别能力不足、大规模无监督数据利用效率有限以及传感器不确定性对系统稳定性的影响等。未来, BEV 感知的发展应重点关注: 端到端闭环感知决策体系的优化、具身智能与 BEV 表示的深度融合以及跨模态统一协同架构的构建与部署。

综上所述, BEV 感知正逐步成为自动驾驶感知范式的核心支撑技术, 其在人工智能、大模型驱动感知以及智能体决策协同等关键领域具有广阔的发展前景。希望本综述能为相关研究提供理论启发与技术参考, 推动 BEV 感知在未来自动驾驶系统中实现更高层次的智能化发展。

参考文献:

[1] GONG Yan, ZHANG Xinyu, LU Jianli, et al. Steering angle-guided multimodal fusion lane detection for autonomous driving[J]. IEEE transactions on intelligent transportation systems, 2025, 26(2): 1470−1481.

[2] BERTOZZ M, BROGGI A, FASCIOLI A. Stereo inverse perspective mapping: theory and applications[J]. Image and vision computing, 1998, 16(8): 585−590.

[3] PHILION J, FIDLER S. Lift, splat, shoot: encoding images from arbitrary camera rigs by implicitly unproject-

ing to 3D[C]//Computer Vision – ECCV 2020. Cham: Springer, 2020: 194−210.

[4] MA Qihang, TAN Xin, QU Yanyun, et al. COTR: compact occupancy Transformer for vision-based 3D occupancy prediction[C]//2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2024: 19936−19945.

[5] LI Yinhao, GE Zheng, YU Guanyi, et al. BEVDepth: acquisition of reliable depth for multi-view 3D object detection[J]. Proceedings of the AAAI conference on artificial intelligence, 2023, 37(2): 1477−1485.

[6] HUANG Junjie, HUANG Guan, ZHU Zheng, et al. BEVDet: high-performance multi-camera 3D object detection in bird-eye-view[EB/OL]. (2021−12−22)[2025−05−27]. https://arxiv.org/abs/2112.11790.

[7] ZHANG Yunpeng, ZHU Zheng, ZHENG Wenzhao, et al. BEVerse: unified perception and prediction in birds-eye-view for vision-centric autonomous driving[EB/OL]. (2022−05−19)[2025−05−27]. https://arxiv.org/abs/2205.09743.

[8] WANG Yan, CHAO Weilun, GARG D, et al. Pseudo-LiDAR from visual depth estimation: bridging the gap in 3D object detection for autonomous driving[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2020: 8437−8445.

[9] GUPTA P, RENGARAJAN R, BANKAPUR V, et al. CVCP-fusion: on implicit depth estimation for 3D bounding box prediction[EB/OL]. (2024−10−16)[2025−05−27]. https://arxiv.org/abs/2410.11211.

[10] LI Zhiqi, WANG Wenhai, LI Hongyang, et al. BEVFormer: learning bird's-eye-view representation from LiDAR-camera via spatiotemporal transformers[J]. IEEE transactions on pattern analysis and machine intelligence, 2025, 47(3): 2020−2036.

[11] LIU Wenxi, LI Qi, YANG Weixiang, et al. Monocular BEV perception of road scenes via front-to-top view projection[J]. IEEE transactions on pattern analysis and machine intelligence, 2024, 46(9): 6109−6125.

[12] LU Siyi, HE Lei, LI S E, et al. Hierarchical end-to-end autonomous driving: integrating BEV perception with deep reinforcement learning[C]//2025 IEEE International Conference on Robotics and Automation. Atlanta: IEEE, 2025: 8856−8863.

[13] ZHANG Zhihuang, XU Meng, ZHOU Wenqiang, et al. BEV-Locator: an end-to-end visual semantic localization network using multi-view images[J]. Science China information sciences, 2025, 68(2): 122106.

[14] JUN W, LEE S, JUN W, et al. A comparative study and optimization of camera-based BEV segmentation for real-time autonomous driving[J]. Sensors, 2025, 25(7): 2300.

[15] LUO Zhipeng, ZHOU Changqing, PAN Liang, et al. Exploring point-BEV fusion for 3D point cloud object tracking with transformer[J]. IEEE transactions on pattern analysis and machine intelligence, 2024, 46(9): 5921−5935.

[16] YANG C, LIN Tianwei, HUANG Lichao, et al. WidthFormer: toward efficient transformer-based BEV view transformation[C]//2024 IEEE/RSJ International Conference on Intelligent Robots and Systems. Abu Dhabi: IEEE, 2024: 8457−8464.

[17] DONG Peiyan, KONG Zhenglun, MENG Xin, et al. HotBEV: hardware-oriented transformer-based multi-view 3D detector for BEV perception[J]. Advances in neural information processing systems, 2023, 36: 2824−2836.

[18] GONG Yan, LU Jianli, LIU Wenzhuo, et al. SIFDriveNet: speed and image fusion for driving behavior classification network[J]. IEEE transactions on computational social systems, 2024, 11(1): 1244−1259.

[19] LI Zhiwei, ZHANG Xinyu, TIAN Chi, et al. TVG-ReID: transformer-based vehicle-graph re-identification[J]. IEEE transactions on intelligent vehicles, 2023, 8(11): 4644−4652.

[20] LI Zhiqi, YU Zhiding, WANG Wenhai, et al. FB-BEV: BEV representation from forward-backward view transformations[C]//2023 IEEE/CVF International Conference on Computer Vision. Porte de Versailles: IEEE, 2024: 6896−6905.

[21] LANG Bo, LI Xin, CHUAH M C. BEV-TP: end-to-end visual perception and trajectory prediction for autonomous driving[J]. IEEE transactions on intelligent transportation systems, 2024, 25(11): 18537−18546.

[22] LI Zhiqi, WANG Wenhai, LI Hongyang, et al. Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal Transformers [EB/OL]. (2022−03−31)[2025−05−27]. https://arxiv.org/abs/2203.17270.

[23] LIU Yingfei, WANG Tiancai, ZHANG Xiangyu, et al. PETR: position embedding transformation for multi-view 3D object detection[C]//Computer Vision–ECCV 2022. Cham: Springer, 2022: 531−548.

[24] LI Yangguang, HUANG Bin, CHEN Zeren, et al. Fast-BEV: a fast and strong bird's-eye view perception baseline [J]. IEEE transactions on pattern analysis and machine intelligence, 2024, 46(12): 8665−8679.

[25] PENG Lang, CHEN Zhirong, FU Zhangjie, et al. BEVSegFormer: bird's eye view semantic segmentation

from arbitrary camera rigs[C]//2023 IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa: IEEE, 2023: 5924−5932.

[26] 黄德启, 黄海峰, 黄德意, 等. BEV 感知学习在自动驾驶中的应用综述[J]. 计算机工程与应用, 2025, 61(6): 1−21.
HUANG Deqi, HUANG Haifeng, HUANG Deyi, et al. Review of application of BEV perceptual learning in autonomous driving[J]. Computer engineering and applications, 2025, 61(6): 1−21.

[27] 时培成, 董心龙, 杨爱喜, 等. 面向自动驾驶的 BEV 感知算法研究进展[J]. 华中科技大学学报 (自然科学版), 2025, 53(5): 104−127.
SHI Peicheng, DONG Xinlong, YANG Aixi, et al. Research progress on BEV perception algorithms for autonomous driving: a review[J]. Journal of Huazhong University of Science and Technology (nature science edition), 2025, 53(5): 104−127.

[28] 肖荣春, 刘元盛, 张军, 等. BEV 融合感知算法综述[C]//中国计算机用户协会网络应用分会 2023 年第二十七届网络新技术与应用年会论文集. 北京: 北京市信息服务工程重点实验室, 2023: 68−72.
XIAO Rongchun, LIU Yuansheng, ZHANG Jun, et al. Comprehensive review of bird's eye view fusion perception algorithms[C]//Proceedings of the 27th Annual Conference on Network New Technologies and Applications of China Computer Users Association Network Application Branch, 2023. Beijing: Beijing Key Laboratory of Information Service Engineering, 2023: 68−72.

[29] 周松燃, 卢烨昊, 励雪巍, 等. 车路两端纯视觉鸟瞰图感知研究综述[J]. 中国图象图形学报, 2024, 29(5): 1169−1187.
ZHOU Songran, LU Yehao, LI Xuewei, et al. Pure camera-based bird's-eye-view perception in vehicle side and infrastructure side: a review[J]. Journal of image and graphics, 2024, 29(5): 1169−1187.

[30] YANG Lei, TANG Tao, LI Jun, et al. Bevheight++: toward robust visual centric 3D object detection[J]. IEEE transactions on pattern analysis and machine intelligence, 2025, 47(6): 5094−5111.

[31] SHI Haobo, HOU Dezao, LI Xiyao, et al. Center-aware 3D object detection with attention mechanism based on roadside LiDAR[J]. Sustainability, 2023, 15(3): 2628.

[32] LI Xiaohai, ZHANG Jieyao, GU Jiaming, et al. BEV-Road: a cross-modal and temporary-recurrent 3D object detector for infrastructure perception[C]//Neural Information Processing. Singapore: Springer, 2026: 270−284.

[33] YANG Lei, YU Kaicheng, TANG Tao, et al. BEVHeight: a robust framework for vision-based roadside 3D object detection[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023: 21611−21620.

[34] SHI Hao, PANG Chengshan, ZHANG Jiaming, et al. CoBEV: elevating roadside 3D object detection with depth and height complementarity[J]. IEEE transactions on image processing, 2024, 33: 5424−5439.

[35] WANG Wenjie, LU Yehao, ZHENG Guangcong, et al. BEVSpread: spread voxel pooling for bird's-eye-view representation in vision-based roadside 3D object detection[C]//2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2024: 14718−14727.

[36] FAN Siqi, WANG Zhe, HUO Xiaoliang, et al. Calibration-free BEV representation for infrastructure perception[C]//2023 IEEE/RSJ International Conference on Intelligent Robots and Systems. Detroit: IEEE, 2023: 9008−9013.

[37] JIA Jinrang, YI Guangqi, SHI Yifeng. RopeBEV: a multi-camera roadside perception network in bird's-eye-view[EB/OL]. (2024−09−18)[2025−05−27]. https://arxiv.org/abs/2409.11706.

[38] ZHANG Tianya, JIN P J. Roadside LiDAR vehicle detection and tracking using range and intensity background subtraction[J]. Journal of advanced transportation, 2022, 2022: 2771085.

[39] WU Jianqing, XU Hao, ZHENG Jianying. Automatic background filtering and lane identification with roadside LiDAR data[C]//2017 IEEE 20th International Conference on Intelligent Transportation Systems. Mielparque Yokohama: IEEE, 2018: 1−6.

[40] LIN Ciyun, GUO Yingzhi, LI Wenjun, et al. An automatic lane marking detection method with low-density roadside LiDAR data[J]. IEEE sensors journal, 2021, 21(8): 10029−10038.

[41] ZHAO Junxuan, XU Hao, LIU Hongchao, et al. Detection and tracking of pedestrians and vehicles using roadside LiDAR sensors[J]. Transportation research part C: emerging technologies, 2019, 100: 68−87.

[42] CUI Yuepeng, XU Hao, WU Jianqing, et al. Automatic vehicle tracking with roadside LiDAR data for the connected-vehicles system[J]. IEEE intelligent systems, 2019, 34(3): 44−51.

[43] WU Jianqing, XU Hao, ZHAO Junxuan. Automatic lane identification using the roadside LiDAR sensors[J]. IEEE intelligent transportation systems magazine, 2020, 12(1): 25−34.

[44] XU Shaoqing, LI Fang, HUANG Peixiang, et al. TiG-Distill-BEV: multi-view BEV 3D object detection via

target inner-geometry learning distillation[J]. IEEE transactions on circuits and systems for video technology, 2026, 36(1): 846−860.

[45] ZHANG Xinyu, GONG Yan, LU Jianli, et al. Multi-modal fusion technology based on vehicle information: a survey[J]. IEEE transactions on intelligent vehicles, 2023, 8(6): 3605−3619.

[46] CHEN Yaqing, WANG Huaming. Accurate and robust roadside 3-D object detection based on height-aware scene reconstruction[J]. IEEE sensors journal, 2024, 24(19): 30643−30653.

[47] WANG Shujian, PI Rendong, LI Jian, et al. Object tracking based on the fusion of roadside LiDAR and camera data[J]. IEEE transactions on instrumentation and measurement, 2022, 71: 7006814.

[48] SONG Ziying, LIU Lin, JIA Feiyang, et al. Robustness-aware 3D object detection in autonomous driving: a review and outlook[J]. IEEE transactions on intelligent transportation systems, 2024, 25(11): 15407−15436.

[49] 周一青, 张浩岳, 齐彦丽, 等. 基于感通算融合和信息年龄优化的车联网多节点协同感知[J]. 通信学报, 2024, 45(3): 1−16.
ZHOU Yiqing, ZHANG Haoyue, QI Yanli, et al. AoI-enabled multi-node cooperative sensing based on integration of sensing, communication, and computing in vehicular networks[J]. Journal on communications, 2024, 45(3): 1−16.

[50] 夏春星, 刘建航, 狄永锟, 等. 基于网络孪生的车路协同感知共享方案[J]. 计算机应用研究, 2025, 42(5): 1363−1369.
XIA Chunxing, LIU Jianhang, DI Yongkun, et al. Perception sharing scheme of vehicle-road cooperation based on cybertwin[J]. Application research of computers, 2025, 42(5): 1363−1369.

[51] 张新钰, 卢毅果, 高鑫, 等. 面向智能网联汽车的车路协同感知技术及发展趋势[J]. 自动化学报, 2025, 51(2): 233−248.
ZHANG Xinyu, LU Yiguo, GAO Xin, et al. Vehicle-road collaborative perception technology and development trend for intelligent connected vehicles[J]. Acta automatica sinica, 2025, 51(2): 233−248.

[52] WANG Zhe, FAN Siqi, HUO Xiaoliang, et al. VIMI: vehicle-infrastructure multi-view intermediate fusion for camera-based 3D object detection[EB/OL]. (2023−03−20)[2025−05−27]. https://arxiv.org/abs/2303.10975.

[53] LI Bin, ZHAO Yanan, TAN Huachun, et al. CoFormerNet: a Transformer-based fusion approach for enhanced vehicle-infrastructure cooperative perception[J]. Sensors, 2024, 24(13): 4101.

[54] ZHOU Linyi, GAN Zhongxue, FAN Jiayuan. CenterCoop: center-based feature aggregation for communication-efficient vehicle-infrastructure cooperative 3D object detection[J]. IEEE robotics and automation letters, 2024, 9(4): 3570−3577.

[55] GONG Yan, ZHANG Xinyu, LIU Hao, et al. SkipcrossNets: adaptive skip-cross fusion for road detection[J]. Automotive innovation, 2025, 8(2): 368−384.

[56] ZHANG Haolin, MEKALA M S, YANG Dongfang, et al. 3D harmonic loss: towards task-consistent and time-friendly 3D object detection on edge for V2X orchestration[J]. IEEE transactions on vehicular technology, 2023, 72(12): 15268−15279.

[57] GONG Yan, WANG Lu, XU Lisheng. A feature aggregation network for multispectral pedestrian detection[J]. Applied intelligence, 2023, 53(19): 22117−22131.

[58] XIANG Chao, XIE Xiaopo, FENG Chen, et al. V2I-BEVF: multi-modal fusion based on BEV representation for vehicle-infrastructure perception[C]//2023 IEEE 26th International Conference on Intelligent Transportation Systems. Bilbao: IEEE, 2024: 5292−5299.

[59] CHEN Runjian, MU Yao, XU Runsen, et al. CO^3: cooperative unsupervised 3D representation learning for autonomous driving[EB/OL]. (2022−06−08)[2025−05−27]. https://arxiv.org/abs/2206.04028.

[60] YU Hang, ZHAO Yongsheng, ZOU Ying, et al. Multistage fusion approach of lidar and camera for vehicle-infrastructure cooperative object detection[C]//2022 5th World Conference on Mechanical Engineering and Intelligent Manufacturing. Maanshan: IEEE, 2023: 811−816.

[61] JIAO Yang, JIE Zequn, CHEN Shaoxiang, et al. MSMDFusion: fusing LiDAR and camera at multiple scales with multi-depth seeds for 3D object detection[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023: 21643−21652.

[62] YU Haibao, TANG Yingjuan, XIE Enze, et al. Vehicle-infrastructure cooperative 3D object detection *via* feature flow prediction[EB/OL]. (2023−03−19)[2025−05−27]. https://arxiv.org/abs/2303.10552.

[63] GONG Yan, JIANG Xinmin, WANG Lu, et al. TCLaneNet: task-conditioned lane detection network driven by vibration information[J]. IEEE transactions on intelligent vehicles, 2024, 9(9): 5680−5693.

[64] HU Yue, LU Yifan, XU Runsheng, et al. Collaboration helps camera overtake LiDAR in 3D detection[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023: 9243−9252.

[65] XU Runsheng, TU Zhengzhong, XIANG Hao, et al.

CoBEVT: cooperative bird's eye view semantic segmentation with sparse transformers[EB/OL]. (2022−07−05)[2025−05−27]. https://arxiv.org/abs/2207.02202.

[66] RÖßLE D, GERNER J, BOGENBERGER K, et al. Unlocking past information: temporal embeddings in cooperative bird's eye view prediction[C]//2024 IEEE Intelligent Vehicles Symposium. Jeju Island: IEEE, 2024: 2220−2225.

[67] WEI Sizhe, WEI Yuxi, HU Yue, et al. Asynchrony-robust collaborative perception via bird's eye view flow[J]. Advances in neural information processing systems, 2023, 36: 28462−28477.

[68] WANG T H, MANIVASAGAM S, LIANG Ming, et al. V_2VNet: vehicle-to-vehicle communication for joint perception and prediction[C]//Computer Vision – ECCV 2020. Cham: Springer International Publishing, 2020: 605−621.

[69] LI Jinlong, XU Runsheng, LIU Xinyu, et al. Learning for vehicle-to-vehicle cooperative perception under lossy communication[J]. IEEE transactions on intelligent vehicles, 2023, 8(4): 2650−2660.

[70] QIAO Donghao, ZULKERNINE F, ANAND A. CoBEVFusion cooperative perception with LiDAR-camera bird's eye view fusion[C]//2024 International Conference on Digital Image Computing: Techniques and Applications. Perth: IEEE, 2025: 389−396.

[71] XIANG Hao, XU Runsheng, MA Jiaqi. HM-ViT: hetero-modal vehicle-to-vehicle cooperative perception with vision Transformer[C]//2023 IEEE/CVF International Conference on Computer Vision. Paris: IEEE, 2024: 284−295.

[72] SHI Shanwei, ZHANG Chaokun, LV Aojia, et al. MCoT: multi-modal vehicle-to-vehicle cooperative perception with Transformers[C]//2023 IEEE 29th International Conference on Parallel and Distributed Systems. Ocean Flower Island: IEEE, 2024: 1612−1619.

[73] YIN Hongbo, TIAN Daxin, LIN Chunmian, et al. V2vformer++: multi-modal vehicle-to-vehicle cooperative perception via global-local transformer[J]. IEEE transactions on intelligent transportation systems, 2024, 25(2): 2153−2166.

[74] LIU Wenzhuo, GONG Yan, ZHANG Guoying, et al. GLMDriveNet: global–local multimodal fusion driving behavior classification network[J]. Engineering applications of artificial intelligence, 2024, 129: 107575.

[75] SONG Ziying, YANG Lei, XU Shaoqing, et al. GraphBEV: towards robust BEV feature alignment for multi-modal 3D object detection[C]//Computer Vision–ECCV 2024. Cham: Springer, 2025: 347−366.

[76] CHANG Cheng, ZHANG Jiawei, ZHANG Kunpeng, et al. BEV-V2X: cooperative birds-eye-view fusion and grid occupancy prediction via V2X-based data sharing[J]. IEEE transactions on intelligent vehicles, 2023, 8(11): 4498−4514.

[77] ZHANG Caiji, TIAN Bin, MENG Shi, et al. V2X-BGN: camera-based V2X-collaborative 3D object detection with BEV global non-maximum suppression[C]//2024 IEEE Intelligent Vehicles Symposium. Jeju Island: IEEE, 2024: 602−607.

[78] LIU Chang, ZHU Mingxu, MA Cong. H-V2X: a large scale highway dataset for BEV perception[C]//Computer Vision – ECCV 2024. Cham: Springer, 2025: 139−157.

[79] ZHANG Xiaofei, LI Yining, WANG Jinping, et al. InScope: a new real-world 3D infrastructure-side collaborative perception dataset for open traffic scenarios[J]. Information fusion, 2025, 128: 103951.

[80] BI Jiangfeng, WEI Haiyue, ZHANG Guoxin, et al. DyFusion: cross-attention 3D object detection with dynamic fusion[J]. IEEE Latin America transactions, 2024, 22(2): 106−112.

[81] WANG Naibang, SHANG Deyong, GONG Yan, et al. Collaborative perception datasets for autonomous driving: a review[EB/OL]. (2025−04−17)[2025−05−27]. https://arxiv.org/abs/2504.12696.

[82] ZHANG Xinyu, LI Zhiwei, GONG Yan, et al. OpenMPD: an open multimodal perception dataset for autonomous driving[J]. IEEE transactions on vehicular technology, 2022, 71(3): 2437−2447.

[83] TANG Zheng, NAPHADE M, LIU Mingyu, et al. CityFlow: a city-scale benchmark for multi-target multi-camera vehicle tracking and re-identification[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2020: 8789−8798.

[84] ZHAN Wei, SUN Liting, WANG Di, et al. INTERACTION dataset: an international, adversarial and cooperative motion dataset in interactive driving scenarios with semantic maps[EB/OL]. (2019−09−30)[2025−05−27]. https://arxiv.org/abs/1910.03088.

[85] WANG Huanan, ZHANG Xinyu, LI Zhiwei, et al. IPS300+: a challenging multi-modal data sets for intersection perception system[C]//2022 International Conference on Robotics and Automation. Philadelphia: IEEE, 2022: 2539−2545.

[86] ZIMMER W, CREß C, NGUYEN H T, et al. TUMTraf intersection dataset: all you need for urban 3D camera-LiDAR roadside perception[C]//2023 IEEE 26th International Conference on Intelligent Transportation Systems.

Bilbao: IEEE, 2024: 1030−1037.

[87] CREß C, ZIMMER W, STRAND L, et al. A9-dataset: multi-sensor infrastructure-based dataset for mobility research[C]//2022 IEEE Intelligent Vehicles Symposium. Aachen: IEEE, 2022: 965−970.

[88] ZHU Xiaosu, SHENG Hualian, CAI Sijia, et al. RoScenes: a large-scale multi-view 3D dataset for roadside perception[C]//Computer Vision – ECCV 2024. Cham: Springer, 2025: 331−347.

[89] YE Xiaoqing, SHU Mao, LI Hanyu, et al. Rope3D: the roadside perception dataset for autonomous driving and monocular 3D object detection task[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022: 21309−21318.

[90] ARNOLD E, DIANATI M, DE TEMPLE R, et al. Cooperative perception for 3D object detection in driving scenarios using infrastructure sensors[J]. IEEE transactions on intelligent transportation systems, 2022, 23(3): 1852−1864.

[91] YU Haibao, LUO Yizhen, SHU Mao, et al. DAIR-V2X: a large-scale dataset for vehicle-infrastructure cooperative 3D object detection[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022: 21329−21338.

[92] YU Haibao, YANG Wenxian, RUAN Hongzhi, et al. V2X-seq: a large-scale sequential dataset for vehicle-infrastructure cooperative perception and forecasting[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023: 5486−5495.

[93] MA Cong, QIAO Lei, ZHU Chengkai, et al. HoloVic: large-scale dataset and benchmark for multi-sensor holographic intersection and vehicle-infrastructure cooperative[C]//2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2024: 22129−22138.

[94] ZHU He, WANG Yunkai, KONG Quyu, et al. OTVIC: a dataset with online transmission for vehicle-to-infrastructure cooperative 3D object detection[C]//2024 IEEE/RSJ International Conference on Intelligent Robots and Systems. Abu Dhabi: IEEE, 2024: 10732−10739.

[95] WANG Hai, NIU Yaqing, CHEN Long, et al. DAIR-V2XReid: a new real-world vehicle-infrastructure cooperative re-ID dataset and cross-shot feature aggregation network perception method[J]. IEEE transactions on intelligent transportation systems, 2024, 25(8): 9058−9068.

[96] ZIMMER W, WARDANA G A, SRITHARAN S, et al. TUMTraf V2X cooperative perception dataset[C]//2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2024: 22668−22677.

[97] YANG Lei, ZHANG Xinyu, LI Jun, et al. V2X-radar: a multi-modal dataset with 4D radar for cooperative perception[EB/OL]. (2024−11−17)[2025−05−27]. https://arxiv.org/abs/2411.10962.

[98] XU Runsheng, XIANG Hao, XIA Xin, et al. Opv2v: an open benchmark dataset and fusion pipeline for perception with vehicle-to-vehicle communication[C]//2022 International Conference on Robotics and Automation. Philadelphia: IEEE, 2022: 2583−2589.

[99] LU Yifan, HU Yue, ZHONG Yiqi, et al. An extensible framework for open heterogeneous collaborative perception[EB/OL]. (2024−01−25)[2025−05−27]. https://arxiv.org/abs/2401.13964.

[100] XU Runsheng, XIA Xin, LI Jinlong, et al. V_2V_4Real: a real-world large-scale dataset for vehicle-to-vehicle cooperative perception[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023: 13712−13722.

[101] CHIU H K, HACHIUMA R, WANG C Y, et al. V_2V-LLM: vehicle-to-vehicle cooperative autonomous driving with multi-modal large language models[EB/OL]. (2024−02−14)[2025−05−27]. https://arxiv.org/abs/2502.09980.

[102] LI Yiming, LI Zhiheng, CHEN Nuo, et al. Multiagent multitraversal multimodal self-driving: open MARS dataset[C]//2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2024: 22041−22051.

[103] LI Yiming, MA Dekun, AN Ziyan, et al. V2X-sim: multi-agent collaborative perception dataset and benchmark for autonomous driving[J]. IEEE robotics and automation letters, 2022, 7(4): 10914−10921.

[104] MAO Ruiqing, GUO Jingyu, JIA Yukuan, et al. Dolphins: dataset for collaborative perception enabled harmonious and interconnected self-driving[C]//Computer Vision – ACCV 2022. Cham: Springer, 2023: 495−511.

[105] KARVAT M, GIVIGI S. Adver-city: open-source multi-modal dataset for collaborative perception under adverse weather conditions[EB/OL]. (2024−10−08)[2025−05−27]. https://arxiv.org/abs/2410.06380.

[106] XU Runsheng, XIANG Hao, TU Zhengzhong, et al. V2X-ViT: vehicle-to-everything cooperative perception with vision Transformer[C]//Computer Vision – ECCV 2022. Cham: Springer, 2022: 107−124.

[107] HUANG Xun, WANG Jinlong, XIA Qiming, et al. V2X-R: cooperative LiDAR-4D radar fusion with denoising diffusion for 3D object detection[EB/OL]. (2024−11−13)[2025−05−27]. https://arxiv.org/abs/2411.08402.

[108] WANG Tianqi, KIM S, JI Wenxuan, et al. DeepAccident: a motion and accident prediction benchmark for V2X autonomous driving[J]. Proceedings of the AAAI conference on artificial intelligence, 2024, 38(6): 5599−5606.

[109] LI Rongsong, PEI Xin. Multi-V2X: a large scale multi-modal multi-penetration-rate dataset for cooperative perception[EB/OL]. (2024−09−08)[2025−05−27]. https://arxiv.org/abs/2409.04980.

[110] XIANG Hao, ZHENG Zhaoliang, XIA Xin, et al. V2X-real: a largs-scale dataset for vehicle-to-everything cooperative perception[C]//Computer Vision – ECCV 2024. Cham: Springer, 2025: 455−470.

[111] ZHOU Zewei, XIANG Hao, ZHENG Zhaoliang, et al. V2XPnP: vehicle-to-everything spatio-temporal fusion for multi-agent perception and prediction[EB/OL]. (2024−12−02)[2025−05−27]. https://arxiv.org/abs/2412.01812.

[112] FAN Siqi, NIE Zaiqing, RUAN Hongzhi, et al. Learning cooperative trajectory representations for motion forecasting[J]. Advances in neural information processing systems, 2024, 37: 13430−13457.

[113] LUO K Z, DAO M Q, LIU Zhenzhen, et al. Mixed signals: a diverse point cloud dataset for heterogeneous LiDAR V2X collaboration[EB/OL]. (2025−02−19)[2025−05−27]. https://arxiv.org/abs/2502.14156.

[114] HAO Ruiyang, FAN Siqi, DAI Yingru, et al. RCooper: a real-world large-scale dataset for roadside cooperative perception[C]//2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2024: 22347−22357.

[115] 崔明阳, 黄荷叶, 许庆, 等. 智能网联汽车架构、功能与应用关键技术[J]. 清华大学学报 (自然科学版), 2022, 62(3): 493−508.
CUI Mingyang, HUANG Heye, XU Qing, et al. Survey of intelligent and connected vehicle technologies: Architectures, functions and applications[J]. Journal of Tsinghua University(science and technology), 2022, 62(3): 493−508.

[116] 沈甜雨, 陶子锐, 王亚东, 等. 具身智能研究的关键问题: 自主感知、行动与进化[J]. 自动化学报, 2025, 51(1): 43−71.
SHEN Tianyu, TAO Zirui, WANG Yadong, et al. Key problems of embodied intelligence research: Autonomous perception, action, and evolution[J]. Acta automatica sinica, 2025, 51(1): 43−71.

[117] 王文晟, 谭宁, 黄凯, 等. 基于大模型的具身智能系统综述[J]. 自动化学报, 2025, 51(1): 1−19.
WANG Wensheng, TAN Ning, HUANG Kai, et al Embodied intelligence systems based on large models: A survey[J]. Acta automatica sinica, 2025, 51(1): 1−19.

[118] LIN Lei, FU Jiayi, LIU Pengli, et al. Just ask one more time! self-agreement improves reasoning of language models in (almost) all scenarios[C]//Findings of the Association for Computational Linguistics ACL 2024. Stroudsburg: ACL, 2024: 3829−3852.

作者简介:

宫彦, 副研究员, 博士, 主要研究方向为计算机视觉、自动驾驶和多模态信息融合。获国家发明专利授权 6 项, 发表学术论文 20 余篇。E-mail: gongyan2020@foxmail.com。

张新钰, 副研究员, 清华猛狮智能车团队负责人, 主要研究方向为智能驾驶和多模态信息融合。担任国家重点研发计划项目负责人。多次在国内无人驾驶顶级赛事获得冠亚军, 获 2019 年吴文俊人工智能科技进步二等奖。发表学术论文 100 余篇, 入选 ESI(Essential Science Indicators) 高被引论文 1 篇。E-mail: xyzhang@tsinghua.edu.cn。

刘华平, 教授, 博士生导师, 中国人工智能学会理事、中国人工智能学会认知系统与信息处理专业委员会副主任, 主要研究方向为具身感知与学习, 获吴文俊人工智能科学技术奖。主持国家自然科学基金重点项目 2 项, 发表学术论文 100 余篇。E-mail: hpliu@tsinghua.edu.cn。

[责任编辑: 丁钰]