

面向自动驾驶的鸟瞰视图目标检测研究综述

汪凌峰¹ 但启涛¹ 朱耀辉¹ 陈思涵¹ 张瀚清¹ 王坤峰¹

摘 要 鸟瞰视图 (BEV) 表征基于统一的空间坐标系, 能够实现多传感器时空对齐与特征融合, 在自动驾驶领域三维目标检测任务中展现出显著优势. 首先, 阐述 BEV 表征的优势以及常用自动驾驶 3D 目标检测数据集与评价指标. 其次, 不同于现有综述文献的输入模态特性或模态特征融合的分析视角, 基于 BEV 表征, 从其稠密程度切入, 分别从稀疏表示、稠密表示、稀疏-稠密混合表示角度系统梳理现有研究工作, 指出现有方法在时空消耗与表征完备性之间的权衡困境. 最后, 从全天候鲁棒感知、相机参数无关的自适应几何建模、自动驾驶开放世界检测等角度, 指出实现低成本、高精度自动驾驶 BEV 目标检测的研究趋势, 为从事 BEV 感知研究的学者提供启发, 推动自动驾驶技术的发展.

关键词 自动驾驶; 鸟瞰视图; 稠密表示; 稀疏表示; 稀疏-稠密混合表示

引用格式 汪凌峰, 但启涛, 朱耀辉, 陈思涵, 张瀚清, 王坤峰. 面向自动驾驶的鸟瞰视图目标检测研究综述. 自动化学报, 2026, 52(9): 1868–1902

DOI 10.16383/j.aas.c250173 **CSTR** 32138.14.j.aas.c250173

A Survey of Bird's-eye-view Object Detection for Autonomous Driving

WANG Ling-Feng¹ DAN Qi-Tao¹ ZHU Yao-Hui¹ CHEN Si-Han¹ ZHANG Han-Qing¹ WANG Kun-Feng¹

Abstract Bird's-eye-view (BEV) representation, based on a unified spatial coordinate system, enables multi-sensor spatiotemporal alignment and feature fusion, demonstrating significant advantages in 3D object detection tasks for autonomous driving. Firstly, this paper elaborates on the benefits of BEV representation, and introduces commonly used autonomous driving datasets and evaluation metrics for 3D object detection. Differing from existing review literature that focuses on input modality characteristics or modality feature fusion, this study systematically organizes current research works by using BEV representation density as the perspective, categorizing approaches into sparse representations, dense representations and hybrid sparse-dense representations. This study reveals the inherent trade-off dilemma between temporal and spatial overhead and representation completeness in existing methods. Finally, this study outlines future research trends toward low-cost and high-precision BEV object detection in autonomous driving, including but not limited to all-weather robust perception, camera-parameter-independent adaptive geometric modeling, and autonomous driving open-world object detection. These proposed research directions aim to inspire BEV perception researchers and accelerate technological breakthroughs in autonomous driving systems.

Keywords autonomous driving; bird's-eye-view; dense representation; sparse representation; hybrid sparse-dense representation

Citation Wang Ling-Feng, Dan Qi-Tao, Zhu Yao-Hui, Chen Si-Han, Zhang Han-Qing, Wang Kun-Feng. A survey of bird's-eye-view object detection for autonomous driving. *Acta Automatica Sinica*, 2026, 52(9): 1868–1902

自动驾驶系统由车辆定位、环境感知、路径规划与决策控制四大核心模块构成技术闭环. 其中, 环境感知作为实现自动驾驶的基础支撑环节, 其

3D 目标检测精度直接决定着系统对复杂交通场景的解析能力. 为实现高精度 3D 环境感知, 特征表征空间的科学选择成为关键突破口. 鸟瞰视图 (bird's-eye-view, BEV) 在早期主要指代对 3D 场景的俯视图平面紧凑表示, 用于在二维平面上组织点云或图像特征. 随着研究的深入, 基于 BEV 的感知框架凭借其独特的空间表征优势, 成为自动驾驶环境感知领域的重要技术范式. 具体来说, BEV 感知范式构建以自车为中心的全局坐标系, 其技术优势主要体现在以下维度:

1) 全景感知与时空建模能力: BEV 框架建立自车坐标系下 360° 全景感知域, 有效克服传统前

收稿日期 2025-04-22 录用日期 2026-01-29

Manuscript received April 22, 2025; accepted January 29, 2026

国家自然科学基金 (62376017, 62202058), 北京市自然科学基金 (L253019), 中央高校基本科研业务费专项资金 (buctrc202221) 资助

Supported by National Natural Science Foundation of China (62376017, 62202058), Beijing Natural Science Foundation (L253019), and Fundamental Research Funds for the Central Universities (buctrc202221)

本文责任编辑 丛杨

Recommended by Associate Editor CONG Yang

1. 北京化工大学信息科学与技术学院 北京 100029

1. College of Information Science and Technology, Beijing University of Chemical Technology, Beijing 100029

视角的视野局限. 该表征方式不仅具有明确的空间可解释性, 更能通过统一坐标系下的时空建模, 准确捕捉交通参与者的运动轨迹和交互关系, 为决策规划提供完整的场景理解. 2) 几何畸变矫正特性: BEV 空间中的特征通过空间投影变换可消除传统透视投影中的尺度畸变 (透视图像中相同尺寸目标的近大远小现象) 和遮挡干扰. 在透视视图中, 相同尺寸物体因距离差异产生显著尺度变化, 且相互遮挡导致目标特征不完整. BEV 空间通过俯视投影有效解决了这两类感知难题. 3) 降维与计算效率提升: 通过对高维 3D 点云或图像特征进行俯视投影, 可显著减少高度方向的冗余维度, 降低计算与存储开销, 同时保留与下游任务最相关的几何信息, 从而在保持精度的前提下提高推理效率. 4) 多模态融合有效架构: BEV 检测框架通过构建统一的坐标系, 突破点云与图像之间的异构数据鸿沟, 实现点云、图像的高效融合; 同时, 该架构确保检测结果能够直接接入下游的规划控制模块, 形成感知-决策的技术闭环.

然而, 显式 BEV 表征作为数据向统一空间映射的中间状态, 不可避免地存在信息损失与额外计算开销. 随着研究的推进, BEV 不再局限于早期的“显式网格化俯视特征”形式, 而发展为一种更广义的空间表征概念. 近年来的 BEV 目标检测研究开始探索将稠密的显式 BEV 表征稀疏化或隐式化, 在 BEV 空间中建立结构化感知, 在保证精度的同时进一步提升时空效率. BEV 感知范式已在业界主流自动驾驶方案中得到验证, 其空间表征优势显著提升了感知可靠性, 为高等级自动驾驶系统的落地应用提供关键技术支撑.

现有综述研究主要沿三个维度展开: 1) 视图变换策略导向型分类, 如文献 [1] 聚焦纯视觉 BEV 检测的跨视角特征投影机制; 2) 输入模态导向型分类, 典型文献 [2-3] 着重分析不同模态输入情形下的检测技术; 3) 特征融合导向型分类, 如文献 [4] 从特征表示、特征对齐和特征融合三个维度构建多模态融合检测分类框架.

与现有综述不同, 本文的主要贡献在于提出基于 BEV 特征表达稀疏性的新型分析视角, 对自动驾驶领域最新的 BEV 检测方法进行系统分类总结, 为低成本、高精度的 BEV 检测模型提供新的研究路径, 为算法实际落地应用总结可行的优化方向. 根据点云或图像特征在 BEV 空间表达的稀疏程度, 将现有方法分为基于稠密 BEV 表示、基于稀疏 BEV 表示和基于稀疏-稠密混合 BEV 表示三类, 详细分类方法可参见图 1, 主流 BEV 检测研究的时

间线概览如图 2 所示. 具体来说, 本文内容组织如下:

第 1 节总结自动驾驶领域常用数据集, 对四个主流数据集进行详细介绍, 包括数据集特点以及对应的 3D 检测指标. 第 2 节总结基于稠密表示的典型 BEV 检测方法, 其中点云 BEV 方案根据点云组织方式分为点云式、体素式和支柱式; 图像 BEV 方案根据视图转换过程分为基于深度显式转换、基于高度显式转换和基于网络隐式转换; 多模态融合方案根据特征融合阶段、融合方式分为显式 BEV 表示融合和中间级表示融合; 稠密时序融合策略从并行融合与循环融合两个角度进行分析. 第 3 节介绍基于稀疏表示的 BEV 检测方法, 其中点云分支分为基于中心生成和基于特征扩散两类, 图像分支可分为显式 3D 投影采样和隐式 3D 位置编码, 稀疏时序融合同样分为并行融合与循环融合. 此外, 部分多模态融合方法采取稀疏-稠密两分支混合的特征表示方法, 在第 4 节根据候选目标生成方式不同, 从基于感兴趣区域 (region of interests, ROI) 特征融合和基于稀疏查询融合两个方向总结基于稀疏-稠密混合的表示方法. 第 5 节对未来自动驾驶领域 BEV 检测任务发展趋势进行展望.

1 常用数据集及评价指标

在自动驾驶 3D 感知领域, 多传感器融合技术通过整合不同模态数据 (图 3 展示了典型的环视 RGB 图像与 3D 点云数据可视化结果) 的互补特性, 有效克服了单一传感器的感知局限 (如表 1 所示), 成为构建高鲁棒性环境理解系统的关键路径. 这一技术范式已深度融入车路协同自动驾驶^[5] 的分布式感知架构与具身智能^[6] 系统的感知框架, 为复杂动态场景下的可靠感知提供了理论保障. 为支撑多模态融合算法的迭代优化, 学界近年来陆续发布多个具有里程碑意义的自动驾驶数据集. 本节系统梳理当前 3D 感知研究中的基准数据集, 如表 2 所示. 其中重点解析 KITTI^[7]、nuScenes^[15]、Waymo^[16] 等主流激光雷达多模态数据集以及 4D 雷达数据集 K-Radar^[19], 并系统介绍在 BEV 目标检测任务中的常用评价指标.

1.1 主流自动驾驶数据集

KITTI 数据集^[7] 是由德国卡尔斯鲁厄理工学院和丰田美国技术研究院联合建立的多模态数据集, 其数据采集系统配备 2 个采集图像的高分辨率立体摄像头 (灰度和彩色); 1 个每秒产生百万级 3D 点的 Velodyne 64 线激光雷达; 1 个 GPS/IMU 定位系统 OXTS RT 3003. 采集的数据以 10 Hz 的频率

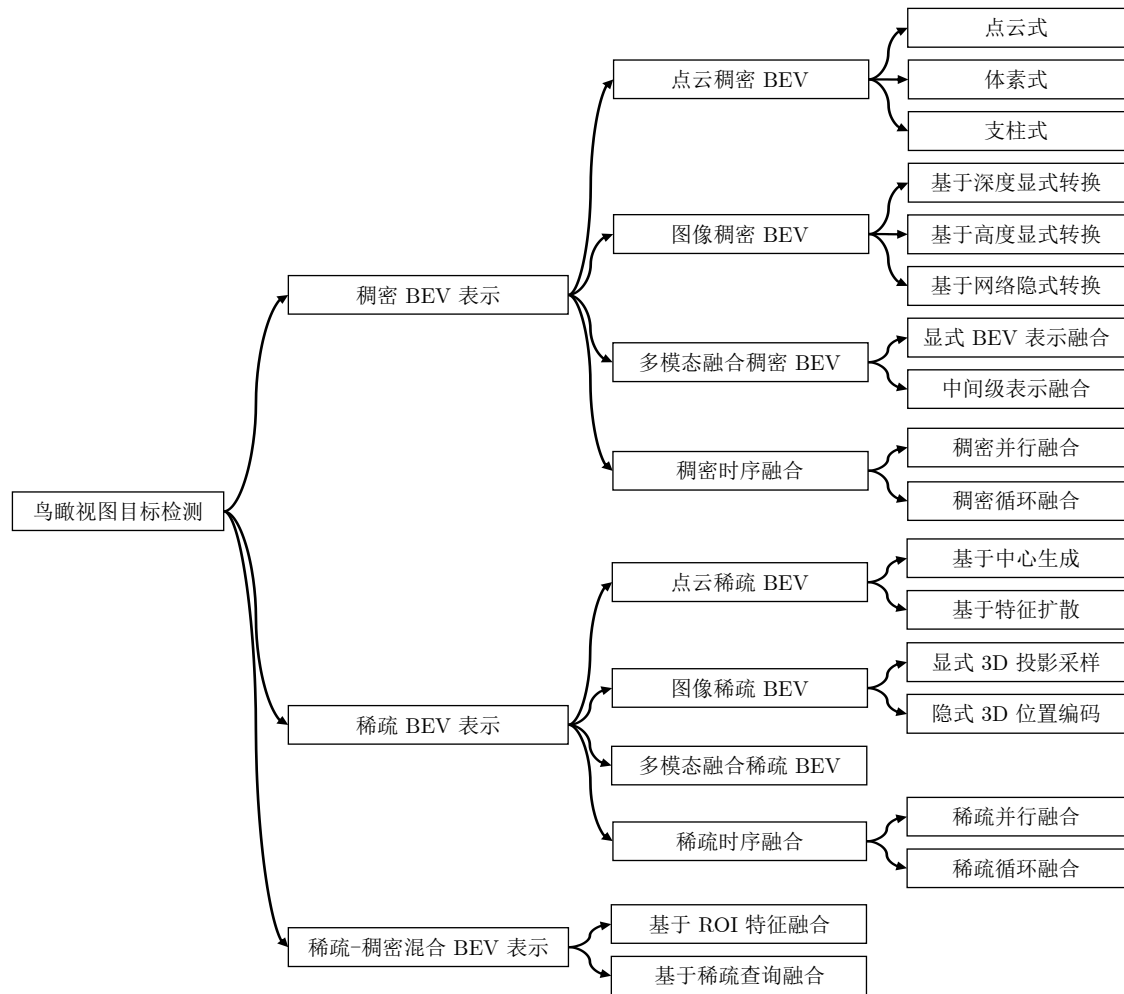


图 1 BEV 目标检测方法分类

Fig.1 Taxonomy of BEV object detection methods

采样及同步, 采集的场景涵盖中等规模城市、乡村以及高速公路. 数据集包含 389 对立体图像和光流图、39.2 km 视觉测距序列以及超过 20 万个 3D 目标框标注, 常被用于立体视觉、光流、视觉里程计以及 3D 目标检测任务. 数据集在 3D 检测和 BEV 检测两个基准测试中包含 7 481 张训练图像和 7 518 张测试图像以及对应的点云数据, 总共约 8 万个标注目标. 数据集检测性能评估采用渐进式难度划分, 基于目标遮挡程度将测试样本划分为简易、中等和困难三个等级.

nuScenes 数据集^[15]是由 Motional 团队推出的一个大规模多模态自动驾驶数据集, 同时开源了数据集处理工具包 nusenes-devkit. 数据采集车辆装备 1 个旋转雷达、5 个远程雷达传感器和 6 个 RGB 相机. 数据集包含在波士顿和新加坡采集的数百小时的高分辨率传感器数据, 约由 1 000 个驾驶场景组成, 覆盖各种复杂的交通环境以及不同气象场景,

包括城市街道、高速公路、乡村道路, 雨天、黄昏、夜晚等驾驶场景. 其中数据集共包含约 140 万个 3D 目标框标注, 常被用于 3D 目标检测、3D 跟踪、运动预测、3D 全景分割等任务.

Waymo 数据集^[16]由谷歌 Waymo 无人驾驶公司于 2020 年发布, 是一个用于 2D、3D 物体检测和跟踪任务的大型多模态自动驾驶数据集. 数据采集设备搭载 5 个 RGB 相机、5 个激光雷达, 其中 RGB 相机在车顶排列, 覆盖角度与 nuScenes 有所区别, 分别拍摄了右侧、右前方、前方、左前方、左侧 5 个方位, 未覆盖车辆后方区域. 数据集由 1 150 个场景组成, 包含在城市与郊区采集的白天、夜晚和黎明三个时段的数据, 每个场景持续 20 s, 其中包含约 1 200 万个 3D 目标框标注. Waymo 在检测基准测试中设置了 LEVEL1 和 LEVEL2 两个难度级别, 同时还根据感知距离设置了三个不同档位, 分别是 0 ~ 30 m、30 ~ 50 m 和 50 m 以上.

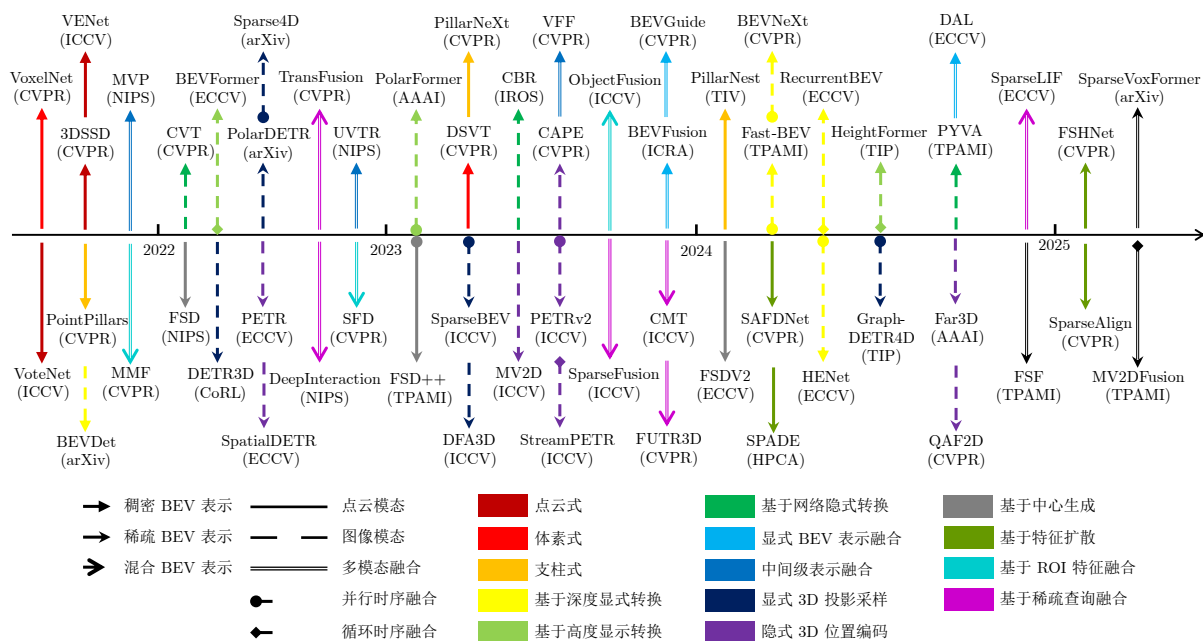


图 2 代表性 BEV 目标检测方法时间线概览

Fig.2 Timeline overview of representative BEV object detection methods

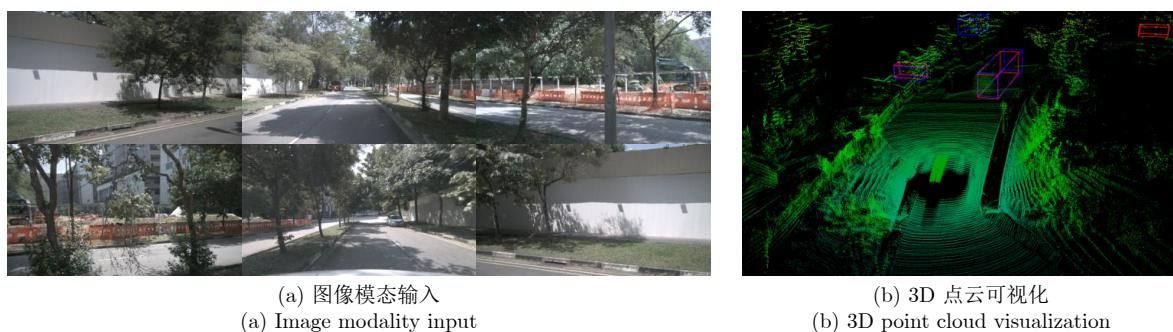


图 3 BEV 目标检测中典型的输入数据可视化示意图

Fig.3 Schematic diagram of visualizing typical input data in BEV object detection

K-Radar 数据集^[19] 由韩国科学技术院发布, 是首个面向自动驾驶场景的中等规模 4D 雷达目标检测数据集. 与以往仅提供 3D 雷达张量的数据集不同, K-Radar 采集了包含多普勒 (Doppler)、距离 (range)、方位角 (azimuth) 及俯仰角 (elevation) 四个维度信息的 4D 雷达张量, 能够完整保留目标的空间结构特征, 支持精确的 3D 场景理解任务. 数据采集平台搭载前向 4D 雷达、64 线长距激光雷达、128 线高分辨率激光雷达、四组环视立体 RGB 相机以及高精度 RTK-GPS 与 IMU 传感器. 数据集共包含约 3.5 万帧多模态同步数据, 覆盖城市道路、郊区道路、高速公路、小巷等多种道路类型, 并涵盖阴天、雾天、雨天、雨夹雪、暴雪等多种气象条件以及昼夜场景, 充分反映了真实复杂交通环境. K-Radar 提供约 9.3 万个 3D 目标框标注、5 种类别, 同时附

带目标跟踪编号, 最大标注距离达到 120 m. 此外, K-Radar 提供了开发工具 Devkits, 数据集可用于 3D 目标检测、目标跟踪、定位与建图、SLAM 以及多模态融合感知研究等任务. 与仅包含激光雷达的数据集相比, K-Radar 更适合于恶劣天气条件下的算法训练, 从而构建全天候工作的自动驾驶感知系统.

1.2 主流评估指标

平均精度: KITTI^[7] 在 $AP|_{R_{11}}$ ^[22] 基础上改用 40 个相等间隔的召回率点计算平均精度 (average precision, AP), 建立更鲁棒的检测器性能评估基准. KITTI 中 $AP|_{R_{40}}$ 的计算如下式所示:

$$AP|_{R_{40}} = \frac{1}{|R_{40}|} \sum_{r \in R_{40}} \rho_{interp}(r) \quad (1)$$

表 1 常用传感器比较
Table 1 Comparison of common sensors

名称	优点	缺点
相机	颜色纹理丰富; 帧率高; 成本较低; 部署简单	不可测量速度; 无深度信息; 受天气因素影响
激光雷达	高精度 3D 位置、分辨率较高	速度测量精度低; 抗干扰能力弱; 对物体材质敏感; 造价高
毫米波雷达	距离、速度测量精度较高; 不受光照和天气影响	分辨率较低; 缺乏高度信息
4D 毫米波雷达	距离、速度测量精度高; 不受光照和恶劣天气影响; 有高度维度	成本高; 分辨率低于激光雷达

表 2 常见多模态自动驾驶数据集
Table 2 Common multi-modal autonomous driving datasets

名称	年份	点云数	图像数	类别数	标注帧数	3D 边界框数	优缺点说明
KITTI ^[7]	2012	1.5×10^4	1.50×10^4	8	1.50×10^4	2.00×10^5	优点: 标定精准, 适合快速开发验证. 缺点: 规模小、场景单一, 仅含前视图图像, 不利于 BEV 任务.
ApolloScape ^[8]	2018	8.0×10^4	1.44×10^5	25	1.44×10^5	7.00×10^4	优点: 像素级语义标注, 车道级检测支持较好. 缺点: 无 360° 传感器覆盖, 3D 标注有限.
KAIST ^[9]	2018	8.9×10^3	8.90×10^3	3	8.90×10^3	0	优点: 包含热成像图像数据, 适配夜间/低光场景任务. 缺点: 类别少、规模小、无 3D 目标框标注, 难以支撑完整检测任务.
A*3D ^[10]	2019	3.9×10^4	3.90×10^4	7	3.90×10^4	2.30×10^5	优点: 包含挑战性场景, 采集车行驶速度快, 适合动态建模. 缺点: 仅含前视视图图像, 标注频率低, 不利于时序建模.
A2D2 ^[11]	2019	—	—	14	1.20×10^4	—	优点: 360° 环视图像数据, 车辆总线信息可辅助 BEV 任务. 缺点: 标注帧少、缺乏点云数据, 场景单一且缺乏多模态信息.
Argoverse ^[12]	2019	4.4×10^4	4.90×10^5	15	2.20×10^4	9.93×10^5	优点: 含高精度地图与大量轨迹数据, 时序建模支持较好. 缺点: 无毫米波雷达数据, 缺乏专用 BEV 检测基准.
H3D ^[13]	2019	2.7×10^4	8.30×10^4	8	2.70×10^4	1.10×10^6	优点: 目标密度高, 标注密集, 适配城市内感知任务. 缺点: 类别少、传感器多样性差, 覆盖范围有限.
Lyft L5 ^[14]	2019	4.6×10^4	3.23×10^5	9	4.60×10^4	1.30×10^6	优点: 规模较大, 传感器多样, 含高清航拍与语义地图, 适配 BEV 建模. 缺点: 类别相对少, 侧重预测/规划任务.
nuScenes ^[15]	2019	4.0×10^5	1.40×10^6	23	4.00×10^4	1.40×10^6	优点: 传感器多样、标注信息丰富, 360° 覆盖, 适配多模态 BEV 任务. 缺点: 类别不平衡严重, 部分驾驶场景挑战性不足.
Waymo ^[16]	2019	2.0×10^5	1.00×10^6	4	2.00×10^5	1.20×10^7	优点: 点云密度高且精度较好, 标注规模大, 360° 覆盖, 广泛用作 BEV 基准. 缺点: 类别少、无语义地图.
KITTI-360 ^[17]	2023	1.0×10^5	3.20×10^5	19	8.00×10^4	6.80×10^4	优点: 相较于 KITTI 扩展鱼眼镜头, 覆盖范围更广, 更适配 BEV 感知. 缺点: 场景数及标注数量较少, 更适合小规模实验.
ONCE ^[18]	2021	1.0×10^6	7.00×10^6	5	1.50×10^4	4.17×10^5	优点: 场景多样、规模大, 相机覆盖范围广, 适配 BEV 任务. 缺点: 类别少、标注帧偏少, 更适用于半/自监督或预训练.
K-Radar ^[19]	2022	3.5×10^4	1.40×10^5	5	3.50×10^4	9.33×10^4	优点: 包含原始 4D 雷达张量, 传感器信息丰富、覆盖角度广, 适合全天候感知. 缺点: 数据总量大、存储成本较高、类别较少.
TJ4DRadSet ^[20]	2022	4.0×10^4	4.00×10^4	8	7.76×10^3	—	优点: 以 4D 雷达为主导, 可用于多模态与全天候感知任务. 缺点: 点云稀疏、类别少、规模有限.
Dual radar ^[21]	2024	5.0×10^4	5.00×10^4	6	1.00×10^4	1.03×10^5	优点: 双 4D 雷达数据, 可对比不同雷达点云性能差异, 场景覆盖全面. 缺点: 类别少、规模有限、传感器覆盖范围较小.

注: “—”表示无统计数据.

$$\rho_{interp}(r) = \max_{r': r' \geq r} P(r') \quad (2)$$

其中, $R_{40} = \{1/40, 2/40, \dots, 40/40\}$; $|R_{40}|$ 表示 R_{40} 的基数, 在 KITTI 中取值为 40; $P(r')$ 表示在实际召回率为 r' 时, 模型对应的原始精确率. 在 nuScenes 基准中, AP 计算与传统计算方法^[7, 22]有所区别, nuScenes 使用目标在地平面上的 2D 投影中心距离 d 代替传统的交并比 (intersection over union, IoU) 作为匹配阈值, 取精度-召回率曲线在召回率和精度超过 0.1 区域下的归一化面积作为 AP. nuScenes 使用 2D 中心距离代替交并比进行目标匹配, 将检测与物体的大小和方向属性解耦, 有效解决了小目标检测中因平移误差导致交并比为 0

的问题, 可以更公平地比较不同方法的性能. 类似地, K-Radar^[19] 也采用该指标测试模型性能.

平均精度均值: 平均精度均值 (mean average precision, mAP) 通过计算所有类别的 AP^[7, 22] 平均值获得. 对于 nuScenes 基准, 2D 投影中心距离 $d \in D$ 且 $D = \{0.5, 1, 2, 4\}$, 单位为 m, 对检测类别集合 C 计算所有阈值 d 的平均值获得 mAP, 如下式所示:

$$\text{mAP} = \frac{1}{|C||D|} \sum_{c \in C} \sum_{d \in D} \text{AP}_{c, d} \quad (3)$$

带朝向的平均精度: Waymo 将目标朝向融入 AP 指标中, 形成了带朝向的平均精度 (average

precision with heading, APH), 计算方法如下式所示:

$$\text{APH} = 100 \int_0^1 \max(h(r') | r' \geq r) dr \quad (4)$$

其中, $h(r)$ 类似于精度-召回率曲线, 但是每个正确正例增加了一个航向准确性权重项: $\min(|\tilde{\theta} - \theta|, 2\pi - |\tilde{\theta} - \theta|)/\pi$, $\tilde{\theta}$ 和 θ 分别是以弧度表示的预测朝向角和真实朝向角。

平均方向相似性: 在 KITTI 中使用平均方向相似性 (average orientation similarity, AOS) 评价预测目标的方向准确性, 具体计算方式如下式所示:

$$\text{AOS} = \frac{1}{11} \sum_{r \in \{0, 0.1, \dots, 1\}} \max_{r': r' \geq r} s(r') \quad (5)$$

$$s(r) = \frac{1}{|D(r)|} \sum_{i \in D(r)} \frac{1 + \cos \Delta_{\theta}^{(i)}}{2} \delta_i \quad (6)$$

其中, $D(r)$ 表示召回率为 r 的检测结果集合; $\Delta_{\theta}^{(i)}$ 表示预测方向与真实方向角度差值; δ_i 是指示函数, 当第 i 个目标为真阳性时取值为 1, 否则为 0。

正确正例: nuScenes 除了使用 mAP 外, 还使用正确正例 (true positive, TP) 描述目标. 对于每个与真实边界框标签匹配的预测框计算 TP, 特别地, 如果某类别召回率低于 0.1, 则将其所有 TP 误差项都设置为 1. 具体来说, TP 包含以下误差项: 1) 平均平移误差 (average translation error, ATE), 表示二维欧几里得中心距离; 2) 平均尺度误差 (average scale error, ASE), $\text{ASE} = 1 - \text{IoU}$, 其中 IoU 代表 3D 交并比, 通过对齐预测物体与真实物体的方向和平移计算 IoU, 再进一步计算 ASE; 3) 平均方向误差 (average orientation error, AOE), 表示预测值与真实值之间最小的偏航角差值; 4) 平均速度误差 (average velocity error, AVE), 表示预测速度与真实速度之间二维速度差的 L2 范数平均值, 用于衡量模型对目标运动速度的估计精度; 5) 平均属性误差 (average attribute error, AAE), 定义为 $\text{AAE} = 1 - \text{acc}$, 其中 acc 为分类准确度. 通过平均所有类别的 TP 值可以获取平均正确正例 (mean true positive, mTP), 即对于任一 TP 误差项, mTP 计算如下式所示:

$$\text{mTP} = \frac{1}{|C|} \sum_{c \in C} \text{TP}_c \quad (7)$$

其中, C 为实际参与评估的目标类别集合; $|C|$ 为类别总数, 在 nuScenes 检测任务中, C 包含 10 个检测类别. 在计算 TP 时, 锥形桶 (cones) 和障碍物 (barriers) 是静止的且未定义属性, 因此不计算这两

个类别的 AVE 以及 AAE; 由于锥形桶没有定义物体朝向, 因此也不计算 AOE.

nuScenes 检测分数: 在 nuScenes 中还提出了一种新的 3D 检测指标——nuScenes 检测分数 (nuScenes detection score, NDS). NDS 结合 AP 和 TP, 是更综合的评估指标, 其计算方法如下式所示:

$$\text{NDS} = \frac{1}{10} \left(5\text{mAP} + \sum_{\text{mTP} \in \text{TP}} (1 - \min(1, \text{mTP})) \right) \quad (8)$$

2 基于稠密 BEV 表示

基于稠密 BEV 表示的检测方法通过建模车辆周围 360° 环境的 3D 空间结构, 成为自动驾驶感知领域的主流研究范式. 其核心在于构建 3D 空间稠密点级表示或将多模态传感器数据 (如图像或点云) 映射至俯视平面, 形成具有空间语义的二维稠密 BEV 特征表示. 其中 BEV 特征在形式上包含长、宽、通道三个维度, 每个网格单元对应实际 3D 空间中的固定物理尺度 (如 200 × 200 的 BEV 特征若采用 0.5 m 每像素的分辨率设置, 可覆盖 100 m × 100 m 的感知范围). 本节将从点云处理、视觉感知以及多模态融合三个维度, 系统梳理基于稠密 BEV 表示的检测方法最新进展.

2.1 点云稠密 BEV

激光雷达点云具有精确的 3D 位置与反射强度信息. 当前主流研究提取点云特征并进行高度维度压缩, 转换为 BEV 表征后实现检测; 早期方法提取原始点级特征表示后, 直接从点级特征表示中识别目标. 根据点云组织形式差异, 现有方法可分为点云式、体素式与支柱式三大技术路线, 其典型流程如图 4 所示.

2.1.1 点云式表示

基于点云表示的方法直接在原始点云域中进行特征建模, 通过点级特征提取与实例查询实现三维目标检测. 虽然未显式投影至 BEV 平面, 但是点云的自然稠密特性 (指代特征占用空间稠密) 使得这类方法无法通过直接构建稀疏 BEV 表征降低特征稠密程度.

该类方法充分保留了点云的空间几何信息, 避免了特征投影造成的空间信息损失, 因此在细粒度几何结构建模方面具有天然优势. 然而, 点云的无序性和稀疏性使得如何高效、稳定地提取结构化特征成为关键挑战. 早期代表性方法 PointNet 系列^[23-24]首次在原始点云空间中引入共享多层感知机、多层次特征提取结构, 为点云特征提取提供了有效思路. 然而, 点级计算在大规模自动驾驶场景中面临计算

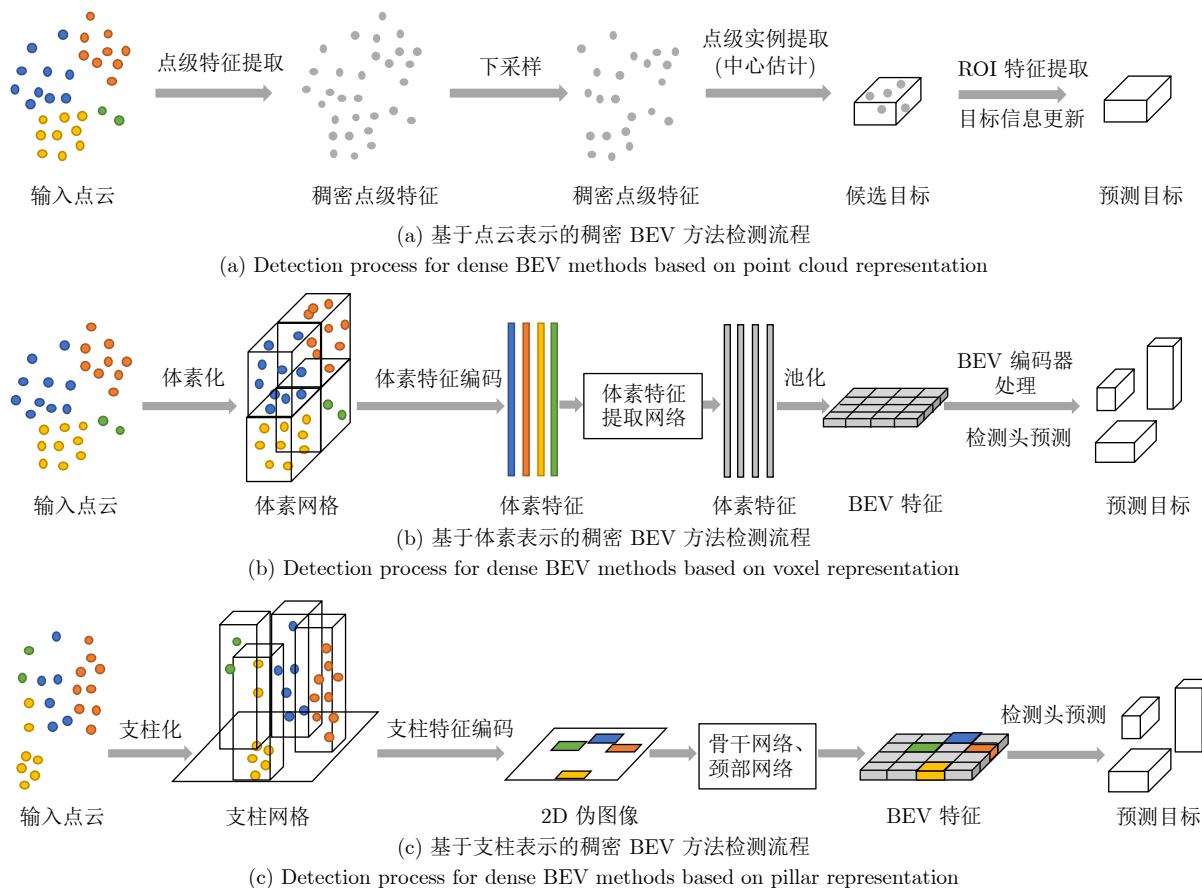


图 4 点云稠密 BEV 检测流程对比

Fig.4 Comparison of dense BEV detection processes based on point clouds

效率瓶颈. 为此, 后续工作针对自动驾驶场景应用进行优化.

一方面, 部分研究在高效特征提取方向持续进行改进. 针对 PointNet^[23-24] 中球状查询与特征传播层计算复杂的问题, Point-GNN^[25] 通过构建动态图网络实现局部邻域特征的消息传递, 避免了大量查询计算; IA-SSD^[26] 提出实例感知的下采样策略, 有效过滤背景点以降低计算与内存消耗; 3DSSD^[27] 进一步摒弃高成本的特征传播层^[24], 通过双重最远点采样兼顾几何分布与语义信息, 提升检测效率.

另一方面, 部分研究围绕中心特征缺失问题优化点云检测方法. 由于激光雷达采样点集中于物体表面, 目标中心区域往往缺乏有效点云信息, 导致模型难以感知中心语义上下文. 针对这一问题, VoteNet^[28] 借鉴霍夫投票思想, 通过投票机制推理目标中心点, 从而实现中心特征聚合, 成为该方向的里程碑工作. 随后, MLCVNet^[29] 在投票阶段引入多层次上下文建模增强特征关联; VENet^[30] 进一步结合注意力机制与加权投票策略, 有效抑制背景噪声干扰^[28].

点云式表示在保持几何精度与细粒度结构建模

方面具有独特优势, 其技术演进体现出从点级特征提取、高效结构建模到上下文感知增强的趋势. 然而点云式表示的较大计算复杂度导致其难以应用于自动驾驶场景下的大规模点云数据, 使其慢慢退出研究者视野, 转而研究更为规整的体素式和支柱式表示检测方法.

2.1.2 体素式表示

体素式方法将点云离散化为规则 3D 网格并提取特征, 通过压缩网格高度维度构建点云 BEV 表示, 随后接入检测头实现目标识别. VoxelNet^[31] 首次提出 3D 体素格式, 使用 3D 卷积提取体素特征后通过池化生成点云 BEV 特征, 通过区域提议网络 (region proposal network, RPN)^[32] 识别目标, 随后 SECOND^[33] 引入稀疏卷积以优化特征提取效率, 成为广泛使用的点云骨干网络. 体素式表示方法在精度与效率之间存在多重权衡, 近年来的研究主要围绕以下几类问题与对应策略展开.

1) 下采样导致的几何信息丢失: 提取特征时, 多次体素下采样操作导致细粒度几何信息损失. 针对该问题主要有两条研究路线. 一是尽量保留或恢

复点级细节, 例如 SA-SSD^[34] 引入点级辅助任务监督, 通过体素到点云的逆向特征映射恢复结构信息; SST^[35] 使用单步长稀疏 Transformer 保持原始体素分辨率以避免信息损失. 二是采用两阶段的候选框细化操作弥补信息损失, 例如 PV-RCNN 系列方法^[36-37] 以及 Voxel R-CNN^[38] 在第一阶段用体素生成粗提议, 第二阶段在感兴趣区域 (region of interests, ROI) 内利用 PointNet++^[24] 提取细粒度点云特征, 或通过体素级池化完成细粒度特征提取, 以提升定位精度.

2) 手工设计导致的过拟合问题: 传统基于锚点和非极大值抑制 (non-maximum suppression, NMS)^[39] 后处理的设计容易导致泛化差、效率低等问题, 因而催生了多种更简洁的检测头设计. 研究重点包括直接预测中心点/关键点/角点以及端到端去除 NMS 的框架以避免手工设计. CenterPoint^[40] 使用中心点检测^[41] 预测目标; CG-SSD^[42] 通过预测目标角点引导生成 3D 边界框; AFDet 系列^[43-44] 通过预测并计算关键点热力图峰值直接识别目标, 避免了 NMS 步骤; Object DGCNN^[45] 通过构建动态图模型实现端到端无 NMS 检测.

3) 构建体素长程依赖: 由于卷积类操作受限于局部感受野, 部分工作引入 Transformer^[46] 扩大上下文感受野. VoTr^[47] 使用局部注意力和膨胀注意力捕捉体素全局空间关系; VoxSeT^[48] 提出体素集合注意力模块, 通过隐式编码处理任意尺度体素簇, 同时显著提高计算效率; DSVT^[49] 提出动态稀疏窗口划分和混合窗口划分方式并行处理稀疏体素, 实现感受野范围扩大; MsSVT++^[50] 采用混合尺度注意力分组处理不同范围特征, 配合棋盘采样降低复杂度; ScatterFormer^[51] 提出散布线性注意力并行计算变长体素序列窗口, 使用分块算法降低计算延迟.

综上所述, 体素式稠密 BEV 检测方法的主要演进体现为对三类基本矛盾的迭代解决: 如何在不过显著增加计算成本的前提下保留点级几何信息, 如何减少手工设计以获得更强泛化能力, 以及如何高效建立跨体素的全局上下文. 未来工作可能会更多地采用卷积层+注意力骨干混合、跨尺度特征融合与自适应体素化策略, 同时强调在实时性约束下的推理效率优化.

2.1.3 支柱式表示

支柱式表示方法将感知空间划分为有限/无限高度的垂直支柱, 通过支柱编码器将点云伪图像化, 进一步提取伪图像特征以获取 BEV 表示. 代表工作 PointPillars^[52] 首次提出支柱表示, 并扩展 2D 检

测头 SSD^[53], 在 BEV 特征上实现 3D 目标检测. 目前支柱式表示方法的研究进展聚焦于解决以下两个问题.

1) 粗粒度划分导致的几何细节丢失: 支柱划分的粗粒度导致几何细节丢失, 尤其是在高度维度上. 由于支柱结构本身舍弃了显式的垂直分辨率, 其表达能力天然受限. 针对这一问题, 研究者提出两类改进方向. 一类方法通过增强高度信息建模来恢复三维几何特征. 例如, 融合柱面视图 (cylindrical view) 的方案^[54] 在 BEV 特征中补充垂直方向信息 (柱面视图保留了高度信息); PillarHist^[55] 通过信息熵引导的高度感知直方图编码提升特征细粒度; 而 PillarNeXt^[56] 则提出渐进式体素到支柱 (Voxel2-Pillar) 编码机制, 使柱状特征逐步吸收三维上下文. 另一类方法则着重于优化特征提取结构本身, 以弥补几何信息缺失. 例如 CAFI-Pillars^[57] 设计语义-几何双分支架构, 通过融合分支特征促进支柱利用几何线索; PillarNet++^[58] 在堆叠多注意力编码模块中联合点级、通道级和支柱级特征; 进一步地, PillarNeXt^[59] 则利用复杂主干设计 (结合 ResNet18^[60] 与 ASPP^[61]) 突破精度瓶颈.

2) 非均匀点云分布与遮挡问题: 点云在空间中分布不均匀, 支柱化方法对稀疏点云区域或遮挡目标的特征建模能力有限, 容易导致特征缺失和漏检. 为此, 研究者从特征增强与上下文补偿两方面提出改进方案. InfoFocus^[62] 引入动态信息聚焦机制, 使模型专注于目标角点、中心点和边缘等关键区域; BSH-Det3D^[63] 通过形状补全模块生成热力图以补偿遮挡区域信息; PSA-Det3D^[64] 结合柱状查询与前景点补偿机制扩大感受野、强化被遮挡小目标特征; Fastpillars^[65] 通过综合点编码、最大池化编码和注意力池化编码以增强细粒度特征表达; 而 PillarNest^[66] 进一步将图像领域预训练骨干 ConvNeXts^[67] 迁移至点云任务, 解决了点云稀疏性问题, 实现了精度提升.

当前支柱式表示方法的研究路径清晰地反映了该类方法的核心诉求: 在体素表示的强几何表达能力与支柱表示的高效性之间寻求折中. 因此未来的发展趋势可能包括: 利用多视图或多尺度编码融合提升垂直方向敏感性、结合注意力机制与稀疏卷积实现可变形支柱建模等. 这预示着支柱式方法不再局限于轻量化, 而是向着兼具效率与几何表达力的中间表示范式演化.

2.1.4 点云稠密 BEV 小结

现有主流点云稠密 BEV 表示通过对原始点云进行规则化建模, 提供了兼具精度与效率的 3D 检

测框架. 体素式与支柱式两类方法分别代表“高保真几何建模”与“高效建模”两种路线, 而点云式方法虽然在空间结构表征能力上表现最佳, 但是因为其高时空复杂度导致这类方法并不适用于大规模点云检测场景.

1) 方法特征与优势. 点云式方法在点云空间直接实现检测, 其 3D 结构信息最小; 体素式方法以细粒度体素划分保留更多空间几何信息, 能够在复杂场景下保持检测精度并加快推理速度; 支柱式方法则以二维卷积结构实现高效特征提取, 具有计算量低、部署友好的特点. 体素式与点云式方法均通过规则化表示将无序点云转化为可卷积结构, 为高性能 BEV 检测奠定了基础.

2) 主要挑战. 点云式方法难以高效提取点级特征是其核心痛点, 且过大的搜索空间也阻碍了目标的识别过程; 体素式方法受制于 3D 卷积计算开销, 在大规模点云场景下可能面临计算瓶颈限制; 支柱式方法因舍弃垂直分辨率导致细粒度几何特征缺失更为严重, 成为精度瓶颈. 现有方法在“精度-效率”之间存在折衷关系, 如何在保持高效性的同时维持几何保真度, 是当前研究的核心矛盾.

3) 发展趋势. 未来的研究可能沿着以下方向演进: 一是在卷积、注意力骨干网络中融合多尺度几何特征; 二是引入自适应体素化与可变形支柱划分方法以实现粒度动态平衡; 三是通过结合多源点云特征与全局上下文建模以提升鲁棒性与泛化性. 总体来看, 点云稠密 BEV 表示正从单一高效/高精度范式迈向兼顾计算效率与几何表达力的统一框架.

2.2 图像稠密 BEV

图像稠密 BEV 方法的核心是基于多视图几何技术或数据驱动的网络学习完成 2D 图像特征到 BEV 图像特征的空间转换过程.

基于透视图像的 BEV 感知本质上是解决单目逆投影的数学病态问题, 即从二维像素空间重构 3D 场景的几何推理过程. 一个简化的针孔相机模型示例如图 5 所示, 为便于描述, 并未画出物体与胶片 (成像平面) 之间的屏障, 并且像素坐标系原点平移至图像中心 (可通过相机内参进行归一化实现). 其中 f 表示已知参数焦距, O 为相机光心, $O-XYZ$ 代表相机坐标系, $O'-UV$ 代表像素坐标系. 对于 3D 空间点 $P = (x, y, z)^T$, 其在像素平面上的对应坐标为 $p = (u, v)^T$. 由相似三角形关系可得:

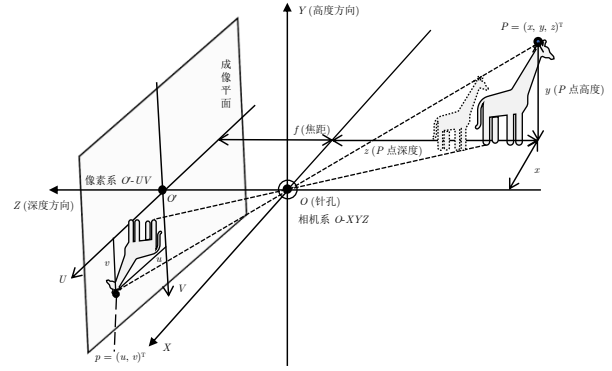


图 5 针孔相机模型

Fig.5 Pinhole camera model

$$\begin{cases} \frac{f}{z} = \frac{v}{y} \\ \frac{f}{z} = \frac{u}{x} \end{cases} \quad (9)$$

通过已知像素坐标 $p = (u, v)^T$ 以及相机参数 f 重构 3D 坐标 $P = (x, y, z)^T$, 但式 (9) 是欠定方程组, 无法直接进行求解, 因此可以对 X, Y, Z 中任一轴增添约束. 在实际应用中, 通常会对深度轴 Z 或高度轴 Y 进行假设. 传统三维重建方法 (如立体视觉和结构光技术) 虽能恢复静态场景的 3D 结构, 但难以满足自动驾驶对动态环境感知的实时性与鲁棒性需求. 当前主流方法通过深度神经网络融合射影几何约束, 构建从透视图像到 BEV 空间的映射关系. 本节从基于深度显式转换、基于高度显式转换以及基于网络隐式转换三大技术路线系统回顾现有工作的演进逻辑.

2.2.1 基于深度显式转换

基于深度显式转换的方法通过显式估计深度信息来恢复三维场景结构, 从而在图像与 BEV 表示之间建立可解释的几何映射. 显式的几何建模思路使得模型不仅具备端到端的可学习性, 也保留了空间结构的可解释性.

这类方法典型流程如图 6 所示: 首先逐像素地预测深度分布; 然后利用深度图与图像特征生成虚拟点云, 并通过相机参数矩阵转换至自车坐标系; 最后通过 BEV 池化操作生成 BEV 特征用作检测头输入. LSS^[68] 作为典型代表工作, 首次提出“Lift-Splat”视图转换器 (view Transformer), 通过预测像素深度分布并结合相机参数完成从透视图像到 BEV 的结构映射. 这缓解了像素与相机视锥空间 (view frustum) 射线上 3D 点之间非双射关系带来的深度模糊性问题, 实现自底向上构建 BEV 特征. 后续 BEVDet^[69] 将视图转换器拓展至 BEV 目标检测任务中, 并通过 scale-NMS 优化 BEV 视图下的

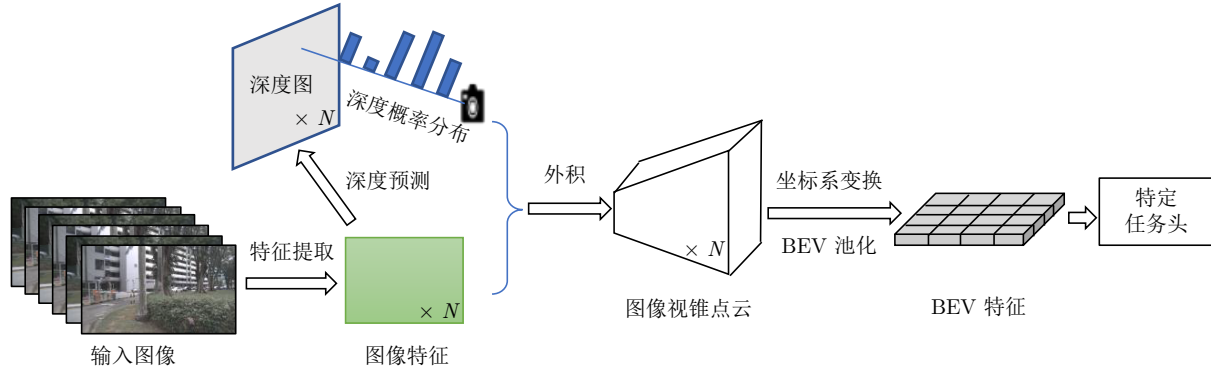


图6 图像稠密 BEV 方法中基于深度显式转换的检测框架

Fig.6 Detection framework of depth explicit transformation-based dense BEV methods in image modality

重复小目标过滤. 此后, 大量研究围绕视图转换过程优化与几何-语义的协同蒸馏两大方向展开, 逐步构建兼顾性能与可解释性的新架构.

1) 视图转换优化: LSS^[68] 提出的视图转换模块存在两类问题, 一是大规模虚拟点云带来巨大的计算与存储负担, 二是深度估计不精确导致模型精度下降. 为此, 研究者从两条路径进行优化.

一方面, 结构层面的加速与内存优化成为主流方向. M²BEV^[70] 通过将深度分布建模为均匀分布, 显著减少了可学习参数量, 以扩展到更大骨干网络与高分辨率; Fast-BEV 系列^[71-72] 与 BEVPoolv2^[73] 构建离线索引实现实时推理; BEVNet^[74] 利用查找表和全卷积结构构建轻量级视图变换结构, 更具部署友好性.

另一方面, 深度预测的准确性直接影响 BEV 特征的几何质量. BEVDepth^[75] 通过点云投影生成显式深度监督信号, 提高深度分布估计精度; BEVStereo 系列^[76-77] 利用多帧与立体几何约束融合单目与双目深度; BEVNeXt^[78] 引入条件随机场 (conditional random field, CRF), 通过颜色平滑先验约束深度预测的一致性. 这一系列方法从减少冗余计算与增强几何精度两个角度出发, 使显式视图转换演化为实用的 BEV 特征生成机制.

2) 知识蒸馏增强: 单目深度估计病态问题与图像传感器的固有缺陷导致模型检测精度低且对光照敏感. 为此, 研究者尝试引入 LiDAR 模型作为教师, 通过跨模态知识蒸馏将空间几何知识传递给视觉模型, 实现几何引导语义-语义增强几何的协同优化. 例如, DualCross^[79] 采用点云到图像的跨模态蒸馏结合对抗训练, 提升模型在域偏移场景下的稳健性; BEVDistill^[80] 在特征与目标两层同时进行蒸馏, 实现多粒度优化; UniDistill^[81] 仅对前景区域进行关系蒸馏与响应蒸馏, 强化关键对象建模; DistillBEV^[82] 引入区域分解、自适应缩放和空间注意

力策略, 使学生模型专注于关键几何线索; SimDistill^[83] 以多模态融合教师为指导, 通过几何补偿模块实现像素、BEV 空间双域学习; 最新的 FSD-BEV^[84] 则提出前景自蒸馏策略, 摆脱对教师模型的依赖, 简化训练流程并减轻蒸馏目标分布差异. 跨模态知识蒸馏使视觉 BEV 模型具备点云模型甚至多模态融合模型的鲁棒感知能力, 同时避免了额外的推理开销, 具有较强的应用前景.

综上, 基于深度显式转换的 BEV 检测方法通过视图转换优化和跨模态知识蒸馏增强, 克服了早期模型计算冗余、深度失真与视觉模态鲁棒性不足等核心难题. 未来, 该方向仍将持续向“高效、鲁棒”的目标深度演进: 一是轻量化几何建模^[85], 通过稀疏化视图变换减少内存占用; 二是自适应深度建模, 融合时序、上下文约束, 动态调整深度分布预测; 三是探索跨模态协同训练^[86], 进一步结合优势模态蒸馏信号实现视觉模型几何能力增强. 可以预见, 显式深度转换会是连接透视视觉感知与 3D 几何建模的关键桥梁, 为 BEV 表征带来更高的几何一致性与鲁棒性.

2.2.2 基于高度显式转换

基于高度显式转换的方法通过假设场景中目标高度已知, 将 2D 图像特征显式映射至 3D 空间, 再通过高度方向压缩形成 BEV 表征. 不同于依赖深度预测的策略, 这类方法以几何投影为核心, 直接利用透视关系将图像特征“绑定”到预定义或可学习的 3D 参考点上, 从而避免深度模糊带来的误差. 其检测流程如图 7 所示, 通常包括: 预定义或初始化参考点、透视投影聚合与 BEV 池化. 相较于深度显式转换, 该方法摆脱了深度预测流程, 但也面临高度假设失准与相机标定失效的挑战.

早期的研究 [87-88] 多依赖平面假设, 通过计算相机平面与地平面的单应性矩阵 (homography matrix) 实现逆透视映射 (inverse perspective map-

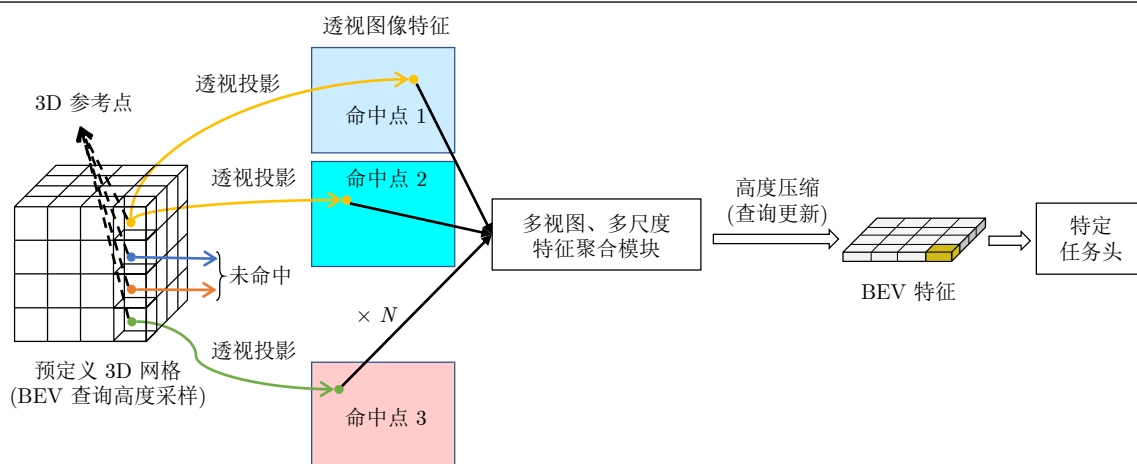


图 7 图像稠密 BEV 方法中基于高度显式转换的检测框架

Fig.7 Detection framework of height explicit transformation-based dense BEV methods in image modality

ping, IPM), 从透视图像转换至俯视图视角. 然而, 这种方法仅在地面近似平坦时有效, 文献 [89] 引入卷积网络来矫正路面不平坦造成的影响, 但仍然难以处理真实世界复杂的道路场景以及相机标定失效的问题. 近年来, 研究重心逐渐从“基于静态假设的映射”转向“基于 3D 参考点的显式投影”, 形成笛卡尔坐标系与极坐标系两类技术路线.

1) 笛卡尔坐标系: 这类方法以笛卡尔坐标系空间中规则网格为基础, 通过显式透视投影将图像特征聚合到 BEV 网格. 代表工作 OFT^[90] 首次通过固定 3D 网格建立图像特征与 BEV 的几何对应, 但是固定 3D 网格缺乏灵活性, 难以应对相机标定误差. 此后的研究者引入可学习参考点与注意力机制, 使 BEV 查询能够根据图像上下文自适应调整采样位置. 例如 BEVFormer 系列^[91-92] 在 BEV 网格查询高度方向均匀采样 3D 参考点, 通过透视投影和可变形注意力^[93] 聚合多尺度图像特征以生成 BEV 特征. GKT^[94] 进一步为 BEV 查询设置统一高度, 利用查找表存储映射关系, 降低了推理开销. HeightFormer^[95] 动态预测 BEV 网格最优高度采样位置. 类似地, 在多模态融合 BEV 检测中也存在大量方法^[96-101] 通过共享 BEV 查询建立多传感器特征的统一 BEV 表示.

2) 极坐标系: 相机成像的视锥空间在俯视图下呈现楔形, 如图 8 所示. 由于这种形状在极坐标系下更易表达且可解释性强, 并且可以避免在笛卡尔坐标系下视锥边界模糊的问题, 因此部分研究转向在极坐标系下构建 BEV 特征. 文献 [102] 在极坐标系下对齐图像特征扫描线和 BEV 极射线 (polar ray), 通过注意力机制建立上下文依赖并生成 BEV 极向特征. PolarFormer^[103] 设计交叉注意力驱动的极坐标系 BEV 查询与检测头, 解决极坐标系

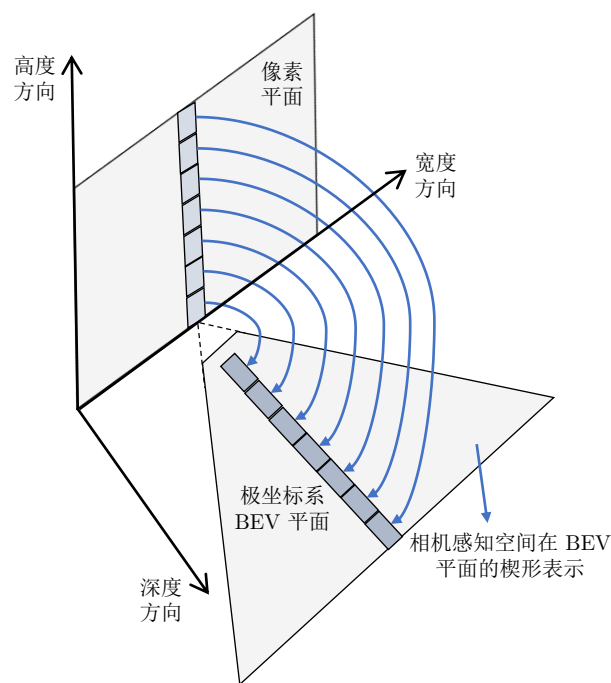


图 8 图像特征列在极坐标系下的表示

Fig.8 Representation of image feature columns in polar coordinate system

下网格不规则的难题. 文献 [104] 则利用极坐标 BEV 查询生成特征后再映射回笛卡尔空间进行检测, 兼顾下游任务兼容性. FocusBEV^[105] 进一步引入自校准循环视图变换模块, 通过交叉注意力在透视空间与极坐标 BEV 空间之间反复校正, 有效抑制无关特征干扰. 与笛卡尔坐标系相比, 极坐标系在表示相机感知区域边界时展现出更强的几何合理性, 但其坐标变换和坐标系转换插值策略仍需优化以保证效率与稳定性.

基于高度显式转换的方法提供了不同于深度已

知几何假设的一种替代路径. 该范式同样以相机成像几何原理为核心, 通过可学习的 3D 参考点实现像素空间与 BEV 空间的精确对齐. 未来研究可从以下方向深化: 第一, 可利用动态采样或引入历史预测优化参考点分布, 减少对固定高度假设的依赖; 第二, 融合笛卡尔与极坐标表示优势, 实现兼具几何合理性与灵活性的混合空间建模方法. 总体来看, 基于高度显式转换和基于深度显式转换都为纯视觉 BEV 检测提供了大量端到端的、兼具效率和可解释性的解决方案.

2.2.3 基于网络隐式转换

基于网络隐式转换的方法不显式建立像素深度或 3D 参考点的几何映射, 而是通过多层感知机 (multilayer perceptron, MLP) 或 Transformer 等网络直接学习透视特征到 BEV 特征的隐式映射函数, 其检测流程如图 9 所示, 包括图像多尺度特征提取、隐式映射网络完成透视特征空间变换、检测头进行任务预测等核心步骤. 从网络结构来看, 当前方法主要分为基于 MLP 隐式建模和基于 Transformer 隐式建模两类.

1) 基于 MLP 隐式建模: 这类方法以透视图像特征作为输入, 利用 MLP 的非线性映射能力, 融入几何约束或自监督信号, 建模不同视图间的复杂几何关系. 代表方法 VPN^[106]、HDMaNet^[107] 以多层 MLP 为核心建模 2D 特征与 BEV 特征之间的像素级全局关系. PON^[108] 进一步结合图像高度方向与极坐标 BEV 的几何约束 (如图 8 所示), 沿极坐标系深度轴生成 BEV 特征. NEAT^[109] 通过 MLP 学习 BEV 网格对图像特征的动态注意力权重, 实现自适应特征聚合. PYVA^[110-111] 最小化透视特征与

BEV 特征 (基于 MLP 生成) 反投影生成的透视特征差异, 通过循环一致性损失约束 BEV 特征的信息完整性. CBR^[112] 通过跨视角相似度匹配模块融合前视图和 BEV 特征, 在无标定场景下保持跨视角一致性.

整体来看, 基于 MLP 的隐式建模方案实现了从几何驱动向数据驱动的转变, 为 BEV 感知提供了高度灵活的结构基础, 但同时也因缺乏几何支撑面临多视角间特征几何一致性弱的难题.

2) 基于 Transformer 隐式建模: Transformer 凭借全局依赖建模能力, 为隐式空间转换提供新的思路. 与基于 MLP 的点对点映射不同, Transformer 通过注意力机制动态学习透视空间与 BEV 空间之间的全局关联, 在隐式建模中兼顾空间一致性与语义聚合. 代表工作 CVT^[113] 首次将像素坐标逆投影到 3D 空间以生成 3D 位置编码, 通过全局交叉注意力实现多视图特征与 BEV 查询的隐式融合. LaRa^[114] 则创新提出视锥射线位置编码, 实现 BEV 查询与图像的高效交互. BEVSegFormer^[115] 直接使用 BEV 查询预测图像 2D 参考点, 在标定失效或无标定情形下实现端到端映射. 相较于基于 MLP 的局部映射, 基于 Transformer 的隐式转换方法在几何一致性建模与多视图融合方面表现更优.

基于网络隐式转换的方法以数据驱动的方式实现从图像空间到 BEV 空间的几何推理, 相比于显式转换策略, 此类方法对相机标定参数的依赖程度更低, 部分方法甚至无需标定参数介入即可自动捕捉视角间的几何关系. 因此该范式具有结构灵活、标定鲁棒性强的优点, 但同时也面临可解释性较弱、几何约束不足等挑战. 未来的研究也可以新颖的方

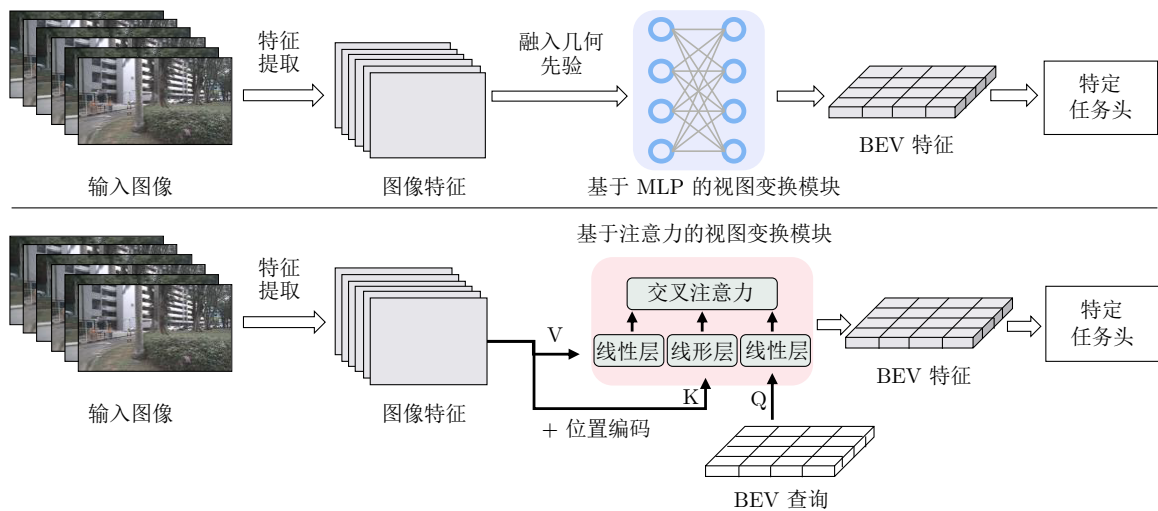


图 9 图像稠密 BEV 方法中基于网络隐式转换的检测框架

Fig. 9 Detection framework of network-based implicit transformation in image dense BEV methods

式融入相同的多视图几何约束,以提高标定误差容忍度,同时保持不同特征表征的几何、语义一致性,避免陷入纯数据驱动的低效学习困境。

2.2.4 图像稠密 BEV 小结

1) 方法特征与优势. 图像稠密 BEV 方法利用相机的高分辨率与低成本优势,通过将多视角透视图像特征统一映射至 BEV 平面,实现对场景的结构化建模. 根据几何建模方式的不同,可分为基于深度显式转换、基于高度显式转换与基于网络隐式转换三类. 前两类方法通过几何约束建立可解释的从透视空间到 BEV 空间的投影关系,具有较好的几何一致性、可控性与高精度特性;隐式转换方法则依托神经网络的强表达能力,通过融入少量几何先验直接学习特征映射关系,能够在缺乏精确标定的场景下仍然保持较强的鲁棒性. 总体而言,现有图像稠密 BEV 方法兼顾几何可解释性与特征表达灵活性,相机的易部署特性使其成为自动驾驶 BEV 感知的重要研究方向。

2) 主要挑战. 当前的图像稠密 BEV 建模方法仍存在下述核心挑战: 首先是基于深度显式建模依赖深度估计与相机标定精度,易受光照变化与遮挡影响,几何误差累积与深度不确定性直接导致模型精度受限;其次,基于高度假设的显式建模仍然依赖相机参数,对标定误差的敏感性导致其在真实世界应用场景受限;最后,基于网络的隐式建模往往因为缺乏先验知识或是先验知识利用不充分,导致模型对复杂的空间变换的学习低效,几何一致性难以保证,需依赖大规模数据补偿几何约束缺失带来的建模偏差,同时黑盒式的建模过程也阻碍了模型优化进程。

3) 发展趋势. 未来研究将沿几何显式建模与神经隐式学习的融合方向发展,重点探索几何先验引导的神经网络隐式转换模式、跨视图几何蒸馏与自监督特征对齐等策略,以在保持几何一致性的同时提升特征鲁棒性与推理效率. 随着实时推理框架的发展,图像稠密 BEV 有望成为端到端自动驾驶视觉感知的核心支撑范式。

2.3 多模态融合稠密 BEV

多模态融合稠密 BEV 目标检测方法基于不同的策略融合多个模态特征信息,以生成稠密的多模态 BEV 特征表示,为下游任务提供丰富的多模态几何、语义信息。

多模态融合感知的核心挑战在于建立异构数据的统一表征空间,以突破传感器间的语义-几何双重鸿沟. 当前主流的多模态融合稠密 BEV 检测方

法根据融合空间可分为两类: 一类方法在显式 BEV 表示下实现特征的对齐与融合; 另一类方法在更原始的 3D 空间(点云/体素空间)实现融合,进一步生成稠密 BEV 特征进行检测. 两类检测流程如图 10 所示。

2.3.1 显式 BEV 表示融合

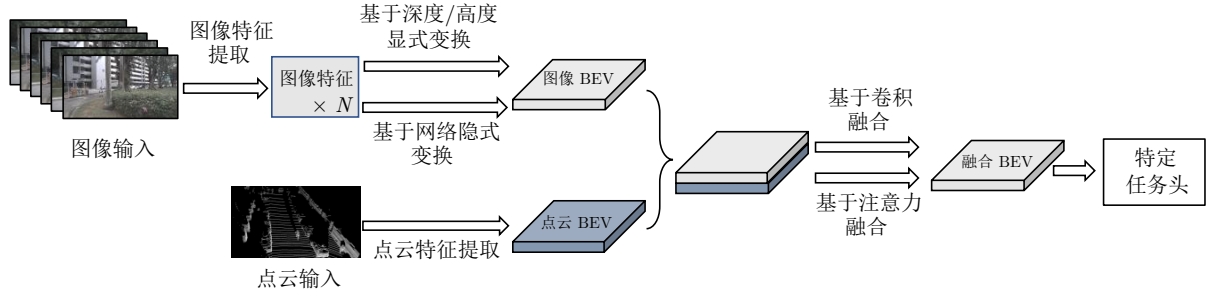
显式 BEV 表示融合方法以几何一致的多模态 BEV 特征对齐与融合为核心思路,其总体流程是: 首先分别从各模态中提取 BEV 表示特征,再将这些特征进行空间对齐与语义融合,最后输入至检测头完成下游任务. 由于各模态特征在 BEV 空间中具有统一的几何参考系,显式融合框架能够在保持结构一致性的同时,实现跨模态信息的高效整合。

在实现范式上,不同方法主要在“如何对齐”与“如何融合”两方面展开研究. 前者关注模态间几何差异的建模与补偿,后者聚焦于特征交互机制的高效与自适应性. 基于此,主流方法大体可分为两类: 基于轻量化卷积的高效融合与注意力驱动的自适应融合。

1) 基于轻量化卷积: 此类方法以几何一致性为核心进行特征整合. 从形式上看,不同模态的 BEV 特征有相同的通道维度和形状,因此一种自然的思路是通过加权或拼接的方式融合不同来源的 BEV 特征,并利用轻量化卷积操作实现多模态特征的语义对齐. 这类方法的目标在于在保持计算效率的前提下实现语义统一与几何对齐. 典型代表方法如 BEVFusion^[116] 和 DAL^[117] 等通过拼接点云与图像的 BEV 特征,并利用轻量化卷积模块实现模态融合,显著提升了目标检测性能。

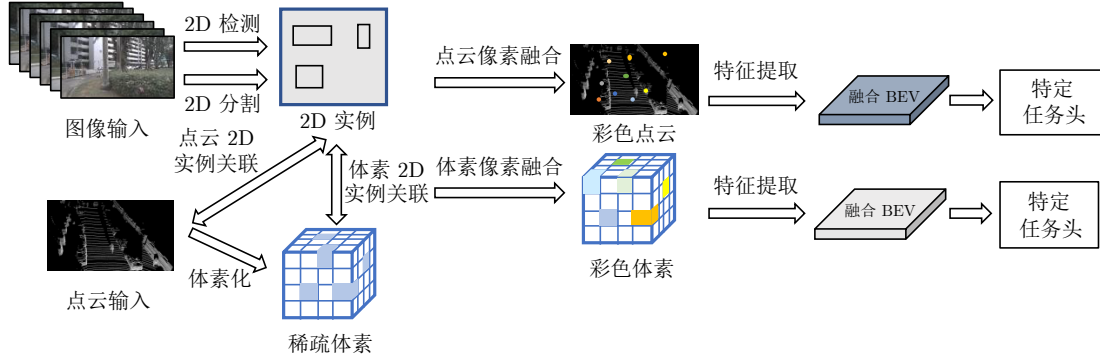
简单的卷积结构虽然可以用于语义融合,但很难校正多模态 BEV 特征间的几何差异,有高空间一致性的多模态 BEV 特征具备更紧密的几何关联性,可以减轻融合模块学习负担,因此部分研究通过增强 BEV 特征之间的空间一致性促进特征融合. 例如 ECFusion^[118] 借鉴光流^[119-120] 思想,应用流变换模块提高多模态 BEV 特征一致性; EA-LSS^[121] 改进目标边缘区域深度估计, GraphBEV^[122] 引入图匹配机制获取邻域深度信息进行辅助,均减小了图像 BEV 几何误差,提高了空间一致性; BEVFusion4D^[96]、LAS^[99] 则是通过注意力完成透视图像特征到 BEV 特征的变换,降低图像分支累计误差,提升多模态一致性. 总体而言,卷积融合方法以结构简洁、效率高为特征,适合实时自动驾驶感知场景,但需要不同的策略缓解多模态 BEV 之间的不一致性(尤其是图像分支),促进融合效果。

2) 注意力驱动的自适应融合: 尽管卷积操作在



(a) 稠密范式下基于显式 BEV 表示融合检测流程

(a) Detection process of fusion based on explicit BEV representation under dense paradigm



(b) 稠密范式下基于中间级表示融合检测流程

(b) Detection process of fusion based on intermediate-level representation under dense paradigm

图 10 稠密范式下两类多模态融合 BEV 检测流程对比

Fig.10 Comparison of two types of multi-modal fusion BEV detection processes under dense paradigm

效率上具备优势,但是这种策略很难弥补模态特征之间过大的语义差距,忽视了多模态信息交互和全局空间相关性.注意力机制的引入,使得模型能够在融合过程中自适应选择关键模态特征并捕获长程依赖,从而提升全局空间一致性与语义互补性.例如,在 BEV 融合阶段 BEVFusion^[123] 引入通道注意力以选择关键特征融合.

另一类融合设计并不局限于仅在多模态 BEV 融合阶段引入注意力,而是选择更为灵活的融合方式.例如 LIFT^[124] 结合点级注意力和 BEV 窗口自注意力进行层次化融合; GAFusion^[125] 额外引入 LiDAR 引导的图像 BEV 生成模块; UniBEV^[100] 基于 BEV 查询向量从图像和点云特征中自适应采样融合; BEVGuide^[97] 借助位置编码与模态无注意力模块,从全局视角隐式捕获多模态互补信息以生成融合 BEV; IS-FUSION^[101] 提出实例引导融合模块联合捕捉实例级和场景级上下文,以提高模型前景敏感性.这类注意力驱动的融合策略展现出更强的语义自适应性与模态鲁棒性,但是也需要注意额外的注意力计算带来的时空消耗.

当前显式 BEV 表示融合方法通过统一空间参考系实现多模态信息的结构化融合.其中,卷积式方法以高效、轻量为核心,适合实时检测与部署;而

注意力驱动方法则侧重于跨模态交互与语义一致性建模,适合高精度场景.未来研究趋势主要包括:一是引入耦合度低的可学习几何补偿模块,增强模态间 BEV 一致性的同时可以在融合阶段根据需要选择不同融合机制;二是引入层次化融合机制设计^[126],融合局部几何与全局语义信息,实现多尺度协同.此类方法由于有较高的精度优势,正推动 BEV 感知由单模态几何建模迈向多模态协同空间理解,为自动驾驶系统构建更稳健、更一致的环境感知基础.

2.3.2 中间级表示融合

中间级表示融合方法在更原始的三维点云或体素空间中完成模态对齐与融合,再进一步压缩为稠密的 BEV 特征.

与显式 BEV 表示融合不同,这类方法直接在统一的三维空间内进行特征对齐与交互,从而在融合时最大程度保留模态的原始几何(主要是高度方向)与语义信息.通过在点云或体素层面进行融合,模型能够充分挖掘多模态间的互补性与几何一致性,提升空间建模精度.

1) 点云空间融合:虽然点云的稀疏性和无序性导致其难以处理,但其原始形式天然保留了最完整的三维几何结构,因此成为一种高保真度的融合空间.典型方法通过将图像的语义或实例特征映射至

点云空间,实现多模态特征的显式对齐与融合.例如 PointPainting^[127]利用 2D 分割模型获取图像的像素级语义,并将点云投影至图像以融合像素语义特征,随后可接入成熟的点云检测框架使用.类似地, MVP^[128]将 2D 图像前景特征逆投影,与点云进行融合; FusionPainting^[129]则是结合 2D、3D 分割结果,基于注意力自适应地融合前景区域特征.这些方法都获得了语义增强的点云输入,可接入不同类型的点云处理网络,可拓展性较强.

2) 体素空间融合:理论上点云表示较好地留存了空间结构,但在实际应用时仍需应对模态融合不充分、计算复杂度高的挑战.相比之下,体素表示将点云规则化,在保持空间拓扑的同时处理更高效,因此成为中间级表示融合的主流方案.根据构建方式不同,现有方法主要包含以下三类融合策略.

第一,基于体素投影的 3D-2D 关联融合策略.此类方法通过建立体素与图像像素之间的几何对应关系,将二维语义特征注入三维体素结构中. VP-FNet^[130]将 3D 目标提议离散化后生成虚拟点,通过投影采样对应的图像特征与邻域体素特征进行融合; DVF^[131]将体素提议中心点投影到图像进行关联,使用 2D 检测器预测的置信度作为权重,与体素特征进行加权融合,类似思路也被应用于文献 [132–133] 等工作中.

第二,基于图像逆投影的特征提升融合策略.这类方法从相反方向出发,将图像特征沿深度轴逆投影至体素空间,与体素特征进行联合建模. VFF^[134]在视锥空间沿像素射线采样 3D 点,使用 MLP 生成的 3D 位置感知的图像体素与点云体素进行动态融合; UVTR^[135]进行深度预测,将图像映射到 3D 体素,逐体素特征相加完成融合; MSMDFusion^[136]采用多深度逆投影策略生成图像体素,通过门控模态感知卷积实现多模态体素局部邻域的自适应融合.

第三,基于注意力机制的体素融合.这类方法利用注意力机制灵活地捕获跨模态特征之间的依赖关系. DeepFusion^[137]通过交叉注意力在点云体素与图像特征之间直接建立联系,其图像特征的空间升维、模态对齐融合的过程都由注意力隐式完成; AutoAlign^[138–139]进一步引入实例级相似性损失度量优化融合结果; HMF1^[140]设计基于查询的多模态体素融合机制; UniTR^[141]构建动态稀疏体素 Transformer^[49]结构,实现体素、像素空间双向跨模态融合.

总结而言,中间级表示融合以三维空间为统一对齐域,有更好的几何保真度与语义完整性.通过在点云或体素层级实现跨模态交互,这类方法显著提升了多模态融合的鲁棒性与表达能力,为多模态 BEV 特征生成提供了更具结构约束的基础.未来该

方向的研究趋势可能会围绕模态融合效率提升(例如构建可控分辨率空间进行特征融合)、自适应几何对齐(例如可学习的隐式空间对齐,摆脱标定依赖)、轻量化注意力机制探索(例如针对体素序列较长的特性,设计局部-全局的多级注意力建模机制)等方向发展.

2.3.3 多模态融合稠密 BEV 小结

1) 方法特征与优势.多模态融合稠密 BEV 方法通过在统一空间内整合来自相机、激光雷达等异构传感器的信息,实现了几何、语义互补.其核心特征在于建立跨模态的几何一致表征,使得视觉模态提供的高分辨率语义信息与点云模态的精确几何结构得以联合建模.显式 BEV 融合在共享的俯视参考系下进行特征对齐,具备结构清晰、可扩展性强的优势;中间级表示融合则直接在点云或体素空间完成对齐与交互,具有更高的几何保真度与语义完整性,累积误差少.总体上,此类方法在保持几何一致性的同时增强了场景理解的鲁棒性,为复杂环境下的高精度感知奠定了基础,也是当前自动驾驶多模态 BEV 感知的主流范式.

2) 主要挑战.尽管多模态融合稠密 BEV 取得显著进展,但仍面临以下关键挑战:第一,几何对齐精度不足.显式 BEV 在特征投影过程中压缩高度信息,导致不同空间位置特征聚合到同一网格,引入语义歧义;传感器标定误差与边缘区域深度模糊等物理因素进一步造成视图转换误差,并在融合阶段累积放大.第二,模态互补性利用受限.中间级表示融合虽能保持原始几何特性,但受限于点云稀疏性与图像密度差异,难以充分提取语义互补信息;同时,依赖特定模态检测或分割结果的设计易受单模态误检与漏检影响,降低融合稳定性.第三,效率与对齐精度矛盾.基于注意力的动态融合在对齐精度上表现优越,但引入的大量计算与显存开销使实时性受限;而采用轻量化卷积结构虽能提升效率,却难以实现严格的跨模态对齐.

3) 发展趋势.未来多模态稠密 BEV 的研究将主要沿以下方向发展:一是可学习的几何对齐与误差自校正机制,利用可学习的几何补偿或隐式坐标映射实现跨模态一致性,无需精确标定;二是模态自适应融合,利用语义与不确定性引导机制,动态评估各模态贡献,实现几何与语义的分层协同;三是高效融合架构,发展例如模态共享骨干与稀疏注意力等策略,在保持融合精度的同时降低计算负担;四是弱监督或自监督学习,利用模态间几何一致性构建自监督信号,减少对精确标注与完备模态数据的依赖,提升泛化与鲁棒性.这些方向将为自动驾驶感知系统构建更全面、更稳健的环境理解能力.

2.4 稠密时序融合

稠密时序融合方法将来自多帧 (历史/未来) 传感器数据的 BEV 网格特征在时域上对齐并聚合, 从而生成几何、运动一致性的 BEV 特征, 以提升检测性能。

目标间相互遮挡是自动驾驶场景的核心挑战, 当目标被遮挡时其对应的图像和点云信息将部分或完全丢失。由于物体间的相对运动, 被遮挡目标通常会在相邻帧中重新显现。因此, 通过整合历史帧信息实现两大核心功能: 1) 增强遮挡场景下的感知鲁棒性; 2) 增强目标运动建模以提升速度估计等动态感知能力。当前稠密时序融合主要包含并行融合与循环融合两种技术路线, 两类时序融合方法的流程如图 11 所示。

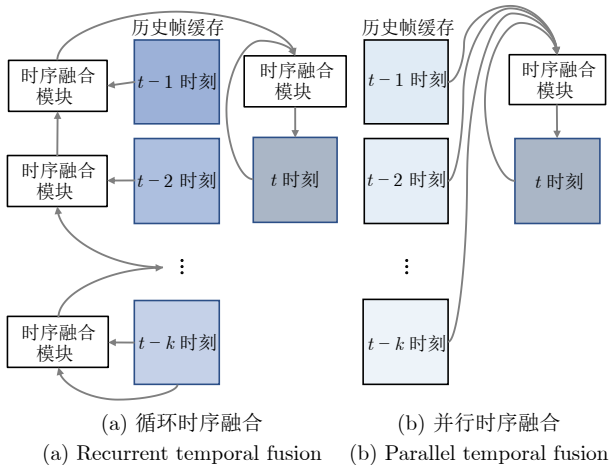


图 11 两类稠密时序融合技术框架对比

Fig. 11 Comparison of two types of dense temporal fusion technology frameworks

2.4.1 稠密 BEV 并行融合

此类方法的核心挑战在于在车辆坐标系下有效对齐并融合不同时刻的 BEV 网格特征, 以避免静态目标的伪影干扰, 同时充分挖掘时序上下文信息。

1) 基础问题与初步方案。在自车坐标系中, 场景中的静止目标会由于自车运动产生虚假位移, 因此直接拼接或叠加 BEV 特征会产生“伪影问题”, 干扰网络学习。为此, 部分研究基于位姿信息进行运动补偿对齐相邻 BEV 帧; BEVDet4D^[142] 利用自车运动信息对齐多帧 BEV 特征, 沿通道维度拼接后通过 BEV 编码器进行时序信息融合, 相比文献 [69] 显著改善了速度估计误差; 该思想后续被广泛应用于文献 [71–72, 74, 103, 105, 125] 等工作中。另一些方法则是进一步从时序特征融合方面进行改进, 如 BEVerse 等方法^[143–145] 对齐多帧 BEV 特征

坐标系后引入时间维度, 通过 3D 卷积进行时序特征建模, 实现空间-时序信息的高效并行建模; 而 BEVFormerV2^[92] 则结合运动补偿与残差降维结构, 提升了特征聚合的稳定性与计算效率。

2) 向长时序融合的演进。随着基础时序对齐方案的成熟, 研究焦点逐渐转向长时序特征建模以获取更丰富的历史运动与语义线索。由于直接对所有历史帧进行高分辨率处理代价过高, 因此出现了长短期结合的分层融合结构。SOLOFusion^[146] 利用低分辨率图像构建 BEV 特征进行长时序融合, 并结合 MVS^[147] 从短时序高分辨率图像中提取深度信息实现时空互补; HOP^[148] 则在辅助任务中利用注意力机制联合建模长短期特征: 长时序分支捕捉物体运动趋势, 短时序分支保留空间细节。进一步地, HENet^[149] 提出新颖的后向-前向结合的迭代时序融合策略, 以复杂结构建模短期精细 BEV 特征, 简单结构建模长期低精度 BEV 特征。

3) 高效时序融合趋势。随着时间推移, 时序 BEV 特征之间的重叠区域会逐渐缩减, 基于拼接的时序融合方式无法充分利用历史信息, 因此 HENet^[149] 使用交叉注意力自适应融合相邻帧 BEV, 实现更灵活的时序建模方式; UniFusion^[150] 则将历史帧图像变换为当前帧的虚拟视图, BEV 查询通过时空交叉注意力统一融合时空特征, 充分利用场景历史信息。另一方面, 长时序融合增加了历史帧到当前帧的对齐操作, 引入大量计算, 因此 BEVNeXT^[78] 提出分组时序融合策略: 组内卷积建立局部交互, 组间卷积扩大时序感受野, 显著减少对齐开销并保留长时序特征建模能力。

2.4.2 稠密 BEV 循环融合

稠密循环时序融合方法通过递归或记忆结构在时间步上逐步累积并更新 BEV 特征。每帧将新观测的状态与历史隐状态融合得到当前场景下的 BEV 表示。

此类方法旨在通过递归结构对 BEV 特征进行持续更新, 使模型具备长期时序记忆与增量式推理能力。与并行式融合获取长时序信息的策略不同, 循环融合方法在当前帧融入前序帧信息, 并在完成检测任务后传递至下一帧, 从而实现在 BEV 特征中累积时序信息。循环融合处理多帧 BEV 的思想与循环神经网络^[151] 类似, 通过共享的隐藏状态, 将不同时间步的 BEV 特征进行传递, 使得历史信息能够得到有效利用, 同时减少了对存储和计算资源的需求。

1) 基于显式对齐的循环更新。稠密 BEV 循环时序融合采取了与并行式融合类似的思路, 先通过

运动补偿减少时序伪影,然后基于卷积或注意力结构将稠密历史 BEV 特征融入当前帧中。代表方法 BEVFormer^[91]、HeightFormer^[95] 均对齐前序帧与当前帧的稠密 BEV 特征,再使用自注意力建模跨帧目标关联,实现时序的逐帧推理与传递过程,为后续循环时序感知奠定基础。

2) 轻量融合与显式反向传播机制。在效率和稳定性方面,后续研究进一步改进了特征更新机制。其中,VideoBEV^[152] 拼接对齐的时序 BEV 特征,使用 1×1 卷积实现高效特征压缩与融合;FusionFormer^[153] 则去除了显式运动补偿步骤,直接利用交叉注意力在相邻帧 BEV 间实现隐式对齐,从而减少基于位姿空间变换带来的累积误差。此外,为探索循环长时序融合,RecurrentBEV^[154] 借鉴 RNN 的反向传播机制,使得所有历史帧特征直接向融合模块反向传播梯度。这种思路等效于并行融合所有时间帧的梯度贡献机制,而且不增加推理开销,使得循环式长时序建模成为可能。

2.4.3 稠密时序融合小结

1) 方法特征与优势。稠密时序融合方法通过时间维度上对齐并聚合多帧 BEV 网格特征,实现了对动态场景的连续建模。其核心优势在于可综合利用历史帧信息以补偿当前帧中被遮挡或信息缺失的区域,同时显著提升目标的运动感知与速度估计能力。并行式融合通过对齐多帧 BEV 特征实现全局时序聚合,具备较强的长时序上下文整合能力;循环式融合则以递归方式逐帧更新 BEV 特征,通过共享隐状态实现历史信息连续传递,在保证时序记忆能力的同时大幅降低了显式存储与计算开销。总体而言,稠密时序融合在提升检测稳健性方面展现出显著优势,已成为动态场景感知的重要技术路线。

2) 主要挑战。当前稠密时序融合仍面临多重瓶颈。首先,精确的时序对齐依赖高质量的位姿估计,运动补偿误差会在多帧融合过程中被放大,导致静态目标伪影或动态目标拖影。其次,并行融合虽能同时聚合多帧信息,但特征冗余和时序噪声易造成信息混淆;而循环融合虽然高效,却易受梯度衰减、误差累积与长时序遗忘的影响。最后,长时序建模引入了显著的计算与存储开销,如何在高分辨率 BEV 网格中实现高效的跨帧特征交互仍是难点。

3) 发展趋势。未来的稠密时序融合研究将朝着统一、高效与自适应方向演进。一方面,可结合显式运动建模与隐式时序对齐,以在复杂场景中实现稳健的跨帧一致性;另一方面,长短期时序特征的分层融合与分组融合机制有望进一步平衡感知精度与

实时性能。同时,轻量化注意力的引入^[155] 将实现高效跨帧建模能力,突破当前效率瓶颈。

2.5 本节小结

本节系统回顾了基于稠密 BEV 表示在点云、图像、多模态感知以及稠密时序融合中的主要技术路线与代表性工作。点云分支主流方法以体素式与支柱式为主,现有方法在高保真几何建模与高效推理之间做出权衡,而较早期的研究在原始点级表示提取特征,较难应用于大规模点云场景;图像分支则围绕基于深度/高度的显式投影与基于网络的隐式转换三类方法展开,主流工作目标为平衡几何可解释性与标定误差鲁棒性;多模态融合方向则区分了显式 BEV 表示融合与三维中间表征层面的融合,分别侧重于多模态结构一致性与几何保真度;稠密时序融合以并行融合和循环融合两种技术维度,进一步突破了感知精度瓶颈,分别在长时序建模精度和效率方面各有优势。总体来看,稠密 BEV 表示方法的显著特征是显式构建 BEV 特征并服务于下游任务,该表征方式同时存在优势与不足。

稠密 BEV 表示的优势在于完整保留了目标对象的特征信息,还能对自行车周边环境要素(如车道线、路缘等)进行细粒度建模。对于依赖场景上下文的感知任务而言,这类空间信息可有效提升目标定位精度与分类可靠性。此外,稠密 BEV 特征天然适配时序特征融合机制,通过多帧信息聚合显著增强了感知系统的鲁棒性。

稠密 BEV 表示的固有局限包括高度信息压缩导致的语义歧义、对深度/标定误差敏感以及随分辨率增长而急剧上升的计算与内存成本。当前研究主要通过多尺度建模、注意力融合、自适应体素/支柱划分和跨模态蒸馏等手段进行精度补偿。然而 BEV 网格分辨率随着感知范围的线性提升将导致计算量和内存消耗呈平方级增长,这对远距离(大于 200 m)、高精度感知需求形成制约,更高的硬件门槛也限制了其在边缘计算设备的部署可行性。第 3 节将转向基于稀疏 BEV 表示的方法,探讨如何通过稀疏化 BEV 建模,在接近稠密 BEV 表示精度的同时降低开销。

3 基于稀疏 BEV 表示

稀疏 BEV 表征范式突破了显式构建 BEV 网格特征的传统框架,其核心在于对场景关键区域的定向建模:针对图像模态,通过可学习查询向量动态捕获前景目标的语义线索;对于点云数据,则充分利用其物理采集机制形成的固有稀疏特性,直接

在物体表面稀疏体素单元中提取稀疏 BEV 特征 (通常是目标特征). 当前研究致力于突破稀疏框架下的感知精度与感知范围限制. 本节系统梳理基于稀疏 BEV 表征的检测技术进展.

3.1 点云稀疏 BEV

点云稀疏 BEV 方法直接利用体素特征的稀疏特性, 从点云提取稀疏体素特征后, 使用初始化的目标查询或直接从未预测的目标, 聚合目标范围内的体素特征, 从而优化预测结果, 其检测流程如图 12 所示.

LiDAR 的工作机制决定了点云主要分布在物体表面, 因此场景中点云分布较为稀疏, 尤其在远距离场景中更加显著. 此前部分研究将点云投影到规则的 BEV 网格或范围视图 (range view), 但这类投影引入了稠密特征, 限制了模型感知范围与计算效率; 而点云稀疏 BEV 方法避免了对无用区域的显式建模, 从而可以突破感知范围限制. 代表方法 VoxelNeXt^[156] 通过合并高度相同的体素以生成稀疏 BEV 特征, 在稀疏检测头中完成体素选择与边界框回归, 保持了稀疏检测流程. 针对基于稀疏体素检测存在的目标中心特征缺失的关键问题^[157], 当前的稀疏点云 BEV 建模方法主要可分为基于中心生成与基于特征扩散的两类稀疏体素建模方案. 前者通过显式生成目标中心实现实例聚合, 后者则通过特征传播或上下文建模隐式恢复目标中心特征.

3.1.1 基于中心生成

基于中心生成的方法通过显式确定目标中心或实例边界来组织稀疏体素特征, 其核心思想是利用点云在物体表面的分布特性, 通过几何线索与体素分割结果生成潜在的目标中心, 从而实现稀疏目标

级特征聚合.

代表方法 FSD 系列^[157-158] 基于体素分割结果投票^[28] 生成候选中心点, 以候选中心为基准通过连通分量标记法聚类目标体素, 实现中心特征的补全. FSDV2^[159] 则摒弃了显式的投票与聚类过程, 提出虚拟体素模块直接预测目标中心位置, 提高了简洁性、泛化性和通用性.

这类方法针对性地指出了点云稀疏 BEV 检测相对于点云稠密 BEV 检测的劣势所在, 为后来的点云稀疏 BEV 研究方法奠定了基础. 其通过中心提议机制获得了稳定的几何先验, 提升了目标定位精度. 然而, 该类方法依赖中心生成过程, 往往在复杂场景或极端稀疏区域存在中心不稳定、聚类敏感等问题, 限制了模型的鲁棒性. 未来的研究可从以下方向进一步拓展: 首先, 可通过引入自适应中心生成机制来提升中心提议的鲁棒性与灵活性, 例如结合不确定性估计与置信度建模, 使得中心预测能够根据点云密度或场景复杂度动态调整生成策略, 从而在极端稀疏或遮挡场景下仍保持稳定性; 其次, 结合时序信息进行中心预测, 有望缓解因遮挡或复杂场景导致的中心预测失效, 从而提升稀疏 BEV 检测的稳健性.

3.1.2 基于特征扩散

基于特征扩散的方法通过特征扩散、上下文注意力或全局聚合机制来隐式恢复中心特征, 避免了显式中心生成的复杂依赖, 实现更高的灵活性与泛化性.

代表性方法 SAFDNet^[160] 提出动态特征扩散机制, 根据目标尺寸自适应地分组扩散特征, 以补齐稀疏体素中物体中心缺失的语义信息; SparseAlign^[161] 引入可扩展坐标的稀疏卷积层, 扩大了稀疏卷积感

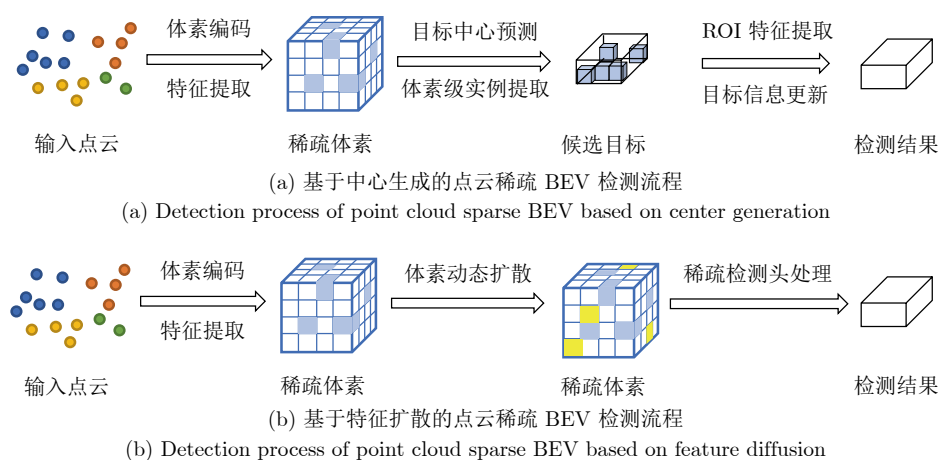


图 12 点云稀疏 BEV 方法检测流程示意图

Fig. 12 Schematic diagram of detection process for sparse BEV methods based on point clouds

受野.

另一部分研究通过稀疏 Transformer 显著提升了特征扩散效率与长程依赖建模能力. 例如, SWFormer^[162] 采用基于窗口的稀疏注意力结构, 通过体素扩散模块将前景特征传播至邻域, 实现连续空间表征; SparseDet^[163] 则将稀疏体素视为查询节点, 聚合局部多尺度特征与全局上下文构建目标、场景级语义; FSHNet^[164] 提出新颖的槽分区机制, 结合线性注意力实现稀疏体素间的全局长距离交互. 此外, SPADE^[165] 基于稠密范式下的支柱特征提取策略, 提出动态向量剪枝算法, 动态剔除不重要的背景支柱特征, 规避体素特征提取过程的中心缺失问题, 同时保持了模型的高稀疏性.

基于特征扩散的稀疏体素建模方法以隐式构建中心特征作为核心思想, 通过动态扩散或注意力机制, 在不依赖显式中心生成的前提下, 显著提升了稀疏表示的语义完整性与特征表达能力. 然而, 当前特征扩散策略仍面临扩散范围与特征噪声间的权衡, 以及 Transformer 结构带来的计算瓶颈等问题. 未来研究可在轻量化稀疏注意力建模、自适应扩散半径学习以及跨模态稀疏特征统一表示等方向展开探索, 从而进一步推动稀疏体素建模在自动驾驶感知中的精度与时空效率双重突破.

3.1.3 点云稀疏 BEV 小结

1) 方法特征与优势. 点云稀疏 BEV 方法充分利用体素表示的稀疏特性, 通过避免对空白区域的冗余建模, 构建了更轻量化的点云 BEV 检测模型, 有更高的计算效率与更广的感知范围. 与稠密 BEV 不同, 稀疏 BEV 直接在有效点云区域进行特征提取与聚合, 从而保持了稀疏空间原始几何结构. 其中基于中心生成的方法通过显式预测目标中心实现实例级体素特征聚合, 强化定位精度; 基于特征扩散的方法依托特征扩散或上下文注意力机制隐式恢复中心特征, 提升目标语义完整性与泛化能力. 总体而言, 点云稀疏 BEV 表征方法检测精度已逐步接近点云稠密 BEV 表征方法, 最新的研究在精度和速度上超越了部分点云稠密 BEV 检测工作, 这说明点云稀疏 BEV 在保持高效性的同时兼顾了空间结构约束, 为远距离与复杂场景下的稀疏感知提供了新范式.

2) 主要挑战. 尽管点云稀疏 BEV 方法在模型轻量化与感知范围方面具有显著优势, 但仍面临若干核心挑战: 第一, 尽管现有方法通过特征扩散和中心点特征聚合的方法弥补缺失的中心特征, 但是局部的邻域特征的近似无法真正替代缺失的中心特征, 传感器特性导致的问题需要精心的结构设计才

能缓解; 第二, 特征扩散方法虽能隐式补全中心特征, 但扩散范围过大会引入噪声, 过小则无法覆盖完整目标语义, 影响检测; 第三, Transformer 等结构显著增强了长程依赖建模能力, 但高计算成本限制了其实时性; 第四, 对不同物体尺度和点云密度使用统一的体素分辨率与聚合策略, 难以兼顾多场景鲁棒性.

3) 发展趋势. 未来点云稀疏 BEV 的发展主要聚焦于三方面: 一是自适应中心生成与动态聚合机制, 通过引入不确定性建模、置信度估计及时序一致性约束, 提升中心预测在复杂场景下的稳定性; 二是轻量化特征扩散与注意力结构, 利用计算稀疏的注意力或特殊体素分块机制, 在保持全局上下文建模能力的同时显著降低计算成本; 三是时空稀疏表示融合, 结合时序点云特征, 可捕捉时间维度的邻域特征以弥补单帧中心特征缺失的缺陷, 使精度向点云稠密 BEV 靠齐, 同时能维持稀疏特性带来的轻量化特性. 总体来看, 点云稀疏 BEV 仍然围绕补齐中心特征的核心进行持续发展, 其未来趋势是实现“结构稀疏、特征充分、推理高效”的感知范式.

3.2 图像稀疏 BEV

图像稀疏 BEV 方法仅使用定制的 3D 检测头直接从 2D 图像特征中提取稀疏的前景 BEV 特征用于检测; 基于投影几何从多尺度图像特征中提取与 3D 目标提议相关的部分, 用于物体的分类与回归任务, 使得图像稀疏 BEV 方法同样突破了由于 BEV 分辨率导致的感知范围限制, 这类方法检测流程如图 13 所示.

早期方法 FCOS3D^[166] 基于 FCOS^[167] 框架解耦 3D 边界框属性, 通过修改检测头直接在 2D 特征上实现 3D 检测, 但由于使用了缺乏几何约束的卷积结构, 该方法难以有效建模图像到 3D 空间的转换关系. 当前主流方法主要采用 Transformer 解码器架构, 通过初始化 3D 空间目标分布的可学习查询向量, 利用注意力机制迭代更新查询状态, 最终由前馈网络解码目标位置、类别及速度等信息. 这类方法继承端到端 2D 检测模型 DETR^[168] 的集合到集合匹配思想, 将 3D 查询作为目标表征, 通过聚合多视图透视特征实现 3D 实例识别, 同时避免了稠密 BEV 特征的构造, 包含显式 3D 投影采样和隐式 3D 位置编码两大技术路线.

3.2.1 显式 3D 投影采样

显式 3D 投影采样方法通过将空间中的 3D 参考点 (即目标锚点) 显式地投影到图像平面, 在对应的 2D 特征图上执行跨视图特征采样, 从而在前景

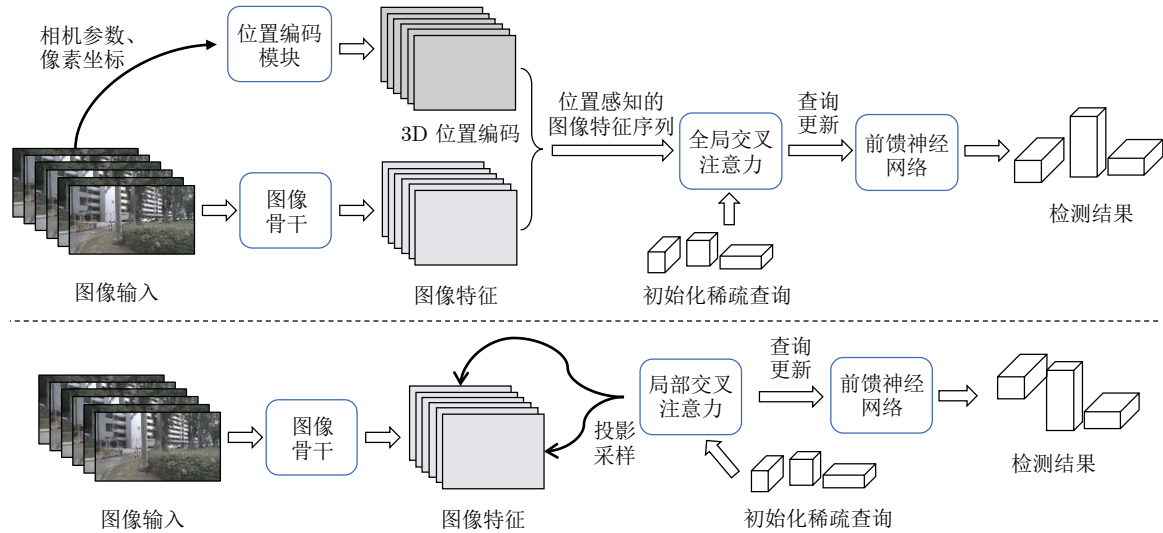


图 13 两类图像稀疏 BEV 方法检测流程示意图

Fig. 13 Schematic diagram of detection process for two types of image-based sparse BEV methods

区域构建稀疏 3D 表征. 为便于描述, 下文简称该方法为“3D-2D 采样”. 其核心思想在于以少量参考点为桥梁, 建立 3D 目标与多视图图像之间的高效特征交互.

代表性方法 DETR3D^[169] 将 DETR^[168] 的端到端 2D 检测框架扩展到 3D 空间, 通过查询预测的 3D 参考点进行跨视图特征采样, 经多级解码器迭代优化检测结果; 随后, PolarDETR^[170] 采用类似架构, 在极坐标系下改进参考点参数化方式. 两类方法共同奠定了 3D-2D 采样检测范式的基础. 随着该方向的发展, 研究的重点逐渐转向采样特征的上下文增强与模型训练效率优化.

在特征采样方面, 由于 DETR3D^[169] 仅在单点上采样特征, 难以充分利用局部上下文信息, 后续工作引入多点采样与结构约束以提升空间一致性. Sparse4D 系列^[171-172] 提出复合参考点机制, 结合几何中心、面中心及查询语义驱动的可学习关键点, 实现多层次特征聚合; Graph-DETR3D^[173-174] 通过构建 3D 图结构并基于图亲和矩阵加权聚合采样特征, 以动态图传播增强采样点局部关联; Sparse-BEV^[175] 引入距离感知自注意力, 通过欧氏距离约束构建多尺度感受野, 增强查询局部多尺度聚合能力. 此外, 针对透视投影下的深度模糊问题, DFA3D^[176] 设计深度感知的 3D 可变形注意力机制, 有效缓解了不同参考点采样到相同特征所带来的耦合误检.

在模型效率与训练稳定性方面, 由于基于 DETR 架构的检测方法通常需要大量随机初始化查询 (600 ~ 900 个) 以保证召回率, 导致计算开销大、收敛缓慢等问题. Sparse4DV3^[177] 通过引入查询去噪

与质量预测辅助任务, 显著加速了收敛过程; GPQ^[178] 则提出逐步修剪策略, 根据分类置信度动态移除冗余查询, 在维持精度的同时降低计算成本, 展现出显式采样在高效稀疏检测中的潜力.

总结而言, 显式 3D 投影采样方法以多视图几何一致性为核心, 构建了图像特征与 3D 空间的直接映射通道, 在统一的 Transformer 框架下实现了端到端的稀疏目标建模. 该方向已逐步从单点采样向多点融合演变, 并基于一系列动态查询优化机制, 使得模型在特征表达与训练效率之间取得更优平衡. 未来的研究可进一步探索基于深度感知的几何一致性建模以及基于少量查询的稀疏检测框架的轻量化设计, 以提升显式 3D-2D 采样方法在复杂动态场景下的鲁棒性与实时性.

3.2.2 隐式 3D 位置编码

基于隐式 3D 位置编码的方法通过引入 3D 位置嵌入 (如正/余弦编码、射线编码等), 在 Transformer 结构中以全局注意力实现多视角图像特征的聚合, 无需显式执行投影与采样操作. 该类方法的核心思想是利用隐式几何约束构建图像与三维空间之间的潜在映射, 从而在统一的端到端框架中实现跨视角特征关联.

然而, 不合适的隐式位置编码难以直接建立像素特征与 BEV 空间的几何对应关系, 模型面临几何关联不稳定、深度模糊歧义以及查询初始化缺乏先验等挑战. 因此, 现有工作主要围绕上述难点展开研究.

1) 位置编码设计优化: 位置编码的改进可以增强 Transformer 表达能力^[179], 是提升基于隐式 3D 位置编码方法的一种关键策略. PETR 系列^[180-183]

通过预定义离散深度值反投影像素坐标生成 3D 位置编码, 突破传统二维编码的几何表征局限性; VEDet^[184] 进一步摒弃离散深度先验, 结合相机位姿与单位逆透视射线生成连续几何编码, 实现更高精度的空间表示; SpatialDETR^[185] 构建基于内参的归一化方向向量编码, 促进跨视图学习目标及其上下文特征; CAPE^[186] 通过相机内参生成相机系下的局部视角位置编码, 解耦内外参学习, 降低模型学习视角变换的难度。

2) 引入深度特征改善深度模糊歧义: 像素深度模糊问题导致基于像素射线的 3D 位置编码阻碍了解码器的学习。为解决该问题, 部分研究显式引入深度估计或深度先验以辅助隐式空间映射。3DPPE^[187] 通过混合深度模块预测像素级深度并映射至精确 3D 坐标进行位置编码; MonoDETR^[188] 利用前景深度图与 Transformer 生成的场景级编码联合建模, 构建深度与视觉双重编码; CrossDTR^[189] 在解码阶段将预测的深度特征作为编码向量执行交叉注意力, 有效缓解了深度模糊带来的小目标误检问题。

3) 查询初始化优化: 为解决 DETR 检测范式共有的收敛性问题, 部分研究尝试引入目标先验或 2D 检测结果以提高模型收敛效率与稳定性。PARQ^[190] 结合 3D 参考点位置与对应像素外观特征进行初始化, 使查询分布更具鲁棒性; MV2D^[191] 利用 2D 检测器生成图像语义丰富的 3D 物体查询, 通过 3D 查询与 2D 目标特征的交叉注意力机制引导模型专注于特定物体区域; Far3D^[192] 基于 2D 目标先验生成自适应 3D 查询, 通过多视角聚合方法捕捉不同尺度、视角的重要特征; QAF2D^[193] 利用投影几何约束筛选高质量锚点, 避免查询初始化依赖于不稳定的深度估计。

隐式 3D 位置编码方法以几何位置嵌入与全局注意机制为核心, 实现了多视角信息的隐式几何融合, 展现出端到端建模和结构简洁的优势。随着位置编码设计、深度特征引入以及查询初始化等方向的不断完善, 该类方法在提升三维几何表征精度与模型收敛稳定性方面取得了显著进展。然而, 隐式位置编码仍面临空间几何解释性不足、深度估计噪声敏感以及全局注意计算代价较高等挑战。未来研究可进一步探索自适应几何感知的位置编码机制、深度与特征协同建模策略以及更高效的轻量级全局注意结构, 以进一步提升模型的几何准确性与部署效率。

3.2.3 图像稀疏 BEV 小结

1) 方法特征与优势。图像稀疏 BEV 方法跳过稠密 BEV 特征的构建环节, 直接在 2D 图像特征中

提取与 3D 目标相关的稀疏表示, 从而显著降低了空间复杂度并扩大了有效感知范围。其核心思想是利用显式或隐式几何建模在 Transformer 框架下建立多视图图像与 3D 目标之间的映射关系。其中, 显式 3D 投影采样基于投影操作显式捕获与前景相关的图像特征, 具备较强的几何可解释性, 局部注意力计算高效也使得该方法可以扩展至高分辨率图像输入场景; 隐式 3D 位置编码则以全局注意力实现跨视角信息融合, 避免投影采样带来的几何误差与计算开销, 同时全局注意力强大的建模能力使其精度表现优异。两类方法共同推动了图像 BEV 检测由稠密向稀疏、高效端到端方向的转变。

2) 主要挑战。尽管图像稀疏 BEV 方法在建模效率与几何一致性方面取得显著进展, 但仍存在若干关键问题: 第一, 显式 3D 采样方法依赖参考点投影, 相机标定误差可能导致跨视图采样特征的不对齐, 导致模型收敛困难; 第二, 显式 3D 采样方法基于局部注意力采样特征, 感受野受限可能导致丢失关键特征; 第三, 深度模糊特性可能导致多个 3D 点采样相同像素点, 从而出现“伪影”检测; 第四, Transformer 解码过程中需维持大量查询以保障召回率, 导致计算负担较重; 第五, 隐式位置编码虽然简化了采样流程, 但由于几何约束不显式, 其空间解释性较弱, 需要合理的位置编码设计保证收敛性; 第六, 查询初始化仍是影响模型收敛性与检测稳定性的关键因素, 如何在稀疏空间中高效引导查询仍具挑战。

3) 发展趋势。未来, 可以通过引入更鲁棒的几何约束和优化方法来减少投影偏差, 探索深度感知驱动的自适应采样机制, 加速模型训练过程; 探索结合全局和局部注意力机制或设计具有更大感受野的网络结构以降低计算成本, 使图像稀疏 BEV 检测能更好地应用于资源受限及实时性要求高的平台。此外, 可以引入 2D/3D 目标联合先验知识, 通过语义引导优化查询初始化, 以获得更好的初始查询空间分布情况, 从而减少冗余查询数量, 为端到端图像稀疏 BEV 检测提供更高效的解决方案。

3.3 多模态融合稀疏 BEV

多模态融合稀疏 BEV 方法以稀疏实例或 3D 查询为中心进行跨模态特征交互, 不构建任一模态的稠密 BEV 特征图, 从而实现高效目标特征建模与目标检测任务。

这类方法是兼顾精度与效率的新兴方法, 与传统基于稠密 BEV 表示的多模态融合方法相比, 其核心思想是仅聚焦于前景或潜在目标区域的稀疏特

征建模与多模态交互, 从而在保持检测性能的同时显著降低计算与存储开销, 尤其适用于远距离感知场景。

该方向的研究大体经历了从稠密特征融合向全稀疏特征融合的演进过程。早期方法如 UVTR^[135] 通过生成图像虚拟体素与点云体素实现多模态融合, 并利用注意力机制自适应选择融合多模态体素以更新目标查询。然而, 其图像分支仍依赖于类似 LSS^[68] 的稠密虚拟点云构造, 导致内存和计算开销较大。与之相反, HS-Fusion^[194] 基于深度图生成稀疏图像虚拟体素, 使用前景感知稀疏卷积提取体素特征以保持稀疏性, 最后基于注意力自适应融合图像体素特征。这种设计的好处是可以在稀疏体素空间融合点云与图像特征, 然后无缝衔接到最先进的体素检测器中, 可扩展性强。随后的研究逐步提出了以实例或查询为中心的全稀疏融合范式。

例如, FSF^[195] 将图像中的 2D 实例逆投影生成稀疏虚拟点云, 基于稀疏实例识别模块^[157] 分别提取多模态 3D 实例特征, 并借助自注意力实现跨模态实例级特征融合。FSF 检测流程如图 14 所示, 作为实例级多模态融合稀疏 BEV 检测的先驱性工作, FSF 在保持稀疏表示的同时获得了优于多种多模态融合稠密 BEV 方法的精度与速度, 使得高精度稀疏 BEV 检测方法成为可能。

基于注意力的建模是稀疏 BEV 范式高精度检测的重要方法。例如, MV2DFusion^[196] 继承“查询即目标”的研究思路, 以 2D/3D 检测结果初始化查询向量, 再通过查询与图像特征、点云稀疏体素特征的交互实现多模态融合, 精度登顶 nuScenes^[15] 检测榜单。SparseFusion^[197] 则是对典型稠密 BEV 方法 BEVFusion^[116] 的改进, 其图像分支仅对前景区域进行视图变换, 以生成稀疏图像 BEV 特征, 从而基

于稀疏卷积结构融合多模态稀疏 BEV 信息, 实现了更高效的远距离检测; 类似地, SparseVoxFormer^[198] 基于全局注意力, 隐式选择图像、稀疏体素混合特征序列中的关键特征用于更新查询, 在保持检测精度的同时显著降低了显式 BEV 特征带来的空间开销。

1) 方法特征与优势。多模态融合稀疏 BEV 方法的核心特征在于构建稀疏多模态 BEV 特征进行融合与检测, 通常以稀疏实例或查询为中心进行跨模态特征交互。该策略通过稀疏卷积与注意力机制, 仅聚焦于前景或潜在目标区域的特征融合, 从而显著降低计算与存储成本, 同时在长距离与复杂场景下保持检测精度。相比传统稠密方法, 此类方法在时空效率、可扩展性等方面具有明显优势, 是多模态感知走向实时化与工程化的重要方向。

2) 主要挑战。尽管稀疏融合范式提升了空间效率, 但在跨模态融合方面仍面临精度与效率权衡的挑战。实例级融合策略 (例如 FSF^[195]) 在时空效率上取得显著优势, 但融合效果受限于单模态检测器能力, 导致模态信息利用不充分, 因此在精度上难与基于稠密 BEV 表征的方法抗衡; 而基于 Transformer 的稀疏 BEV 方案理论上存在内存优势, 并且对应工作已取得精度优势, 但大量的注意力计算和庞大的网络参数使这类模型变得臃肿, 从而失去稀疏范式在时空效率上的优势。因此, 当前该类方法的研究存在明显的精度与效率的权衡挑战。此外, 稀疏表示本身在背景区域的信息缺失也可能影响全局语义理解, 而基于查询的隐式交互机制在无几何先验约束下容易导致特征匹配不稳定, 这些问题也限制了模型的鲁棒性与泛化性能。

3) 发展趋势。未来的研究将重点聚焦于“可学习几何先验引导的自适应融合”与“轻量化跨模态注

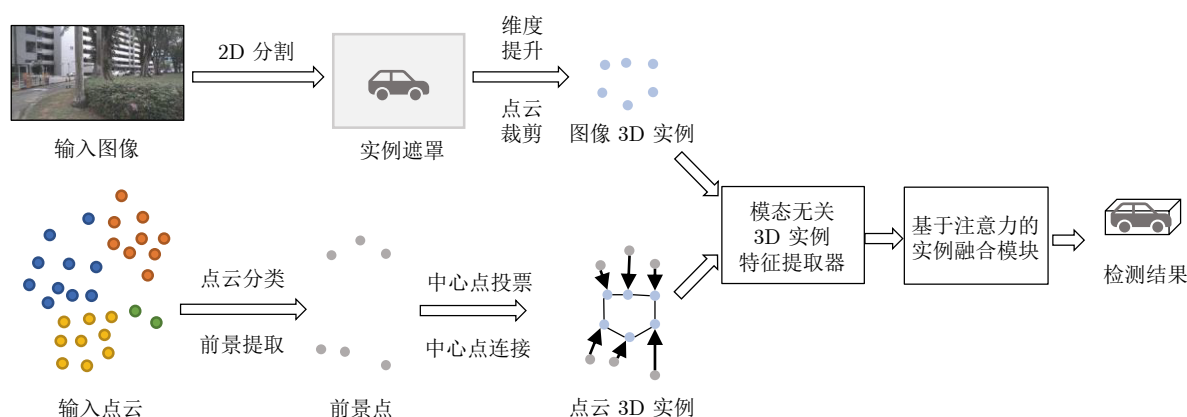


图 14 典型多模态融合稀疏 BEV 检测方法 FSF

Fig. 14 Typical multi-modal fusion sparse BEV detection method: FSF

意力结构”的设计,以进一步增强多模态稀疏 BEV 特征间的空间一致性与融合效率. 总体来看,多模态稀疏 BEV 融合范式在计算效率、长距离感知能力以及实际部署可行性方面均展现出显著潜力,将成为未来多模态 3D 感知系统的重要发展方向.

3.4 稀疏时序融合

稀疏时序融合通过对齐多帧数据坐标系,或直接建立多帧目标级特征关联以利用时序信息,增强模型的动态感知能力. 稀疏范式下的时序融合侧重于在只维护少量目标级查询或关键前景特征的前提下利用多帧信息,因此这类方法在时序融合结构方面存在特点:强调目标级的跨帧关联,追求低存储开销与高效的时空交互机制,同时需解决跨帧目标区分与对齐问题.

3.4.1 稀疏 BEV 并行融合

稀疏范式下的并行时序融合通常将多帧数据作为输入,同时从多帧数据中提取构建目标的关键信息以实现检测. 当前研究主要聚焦于如何在稀疏表示中识别多帧目标关联性,并构建轻量化的时空交互机制.

代表方法 FSD++^[158] 在 FSD^[157] 的基础上把历史前景点并入点云输入以增强对动态部分的敏感性,在输入端显式保留时序线索. 类似地,在纯视觉稀疏 BEV 检测中也存储了多帧图像以实现按时序信息的利用,其核心是显式地聚合多帧图像中关于前景目标部分的信息. Sparse4D^[171]、Graph-DETR4D^[174]、SparseBEV^[175] 以目标查询为媒介,基于局部注意力执行 4D 跨时空图像特征采样以融合时序信息;有研究进一步尝试在查询空间中显式编码时序几何信息,以隐式实现跨帧特征的对齐能力, PETRv2^[182] 为历史帧图像生成当前帧坐标系的 3D 位置编码,通过全局注意力融合多帧图像特征以隐式利用时序.

稀疏 BEV 并行时序融合方法保持了稀疏推理范式的高效性,输入级时序增强思路为后续的稀疏时空建模提供了实践基础. 然而,目前此方向仍处于初步探索阶段,由于缺乏稠密 BEV 表示,当前方法需要存储大量的历史特征,过大的存储开销导致难以支撑长时序融合,因此促进了稀疏范式下的循环时序融合策略.

3.4.2 稀疏 BEV 循环融合

稀疏 BEV 范式仅维护少量目标级特征,因此多帧目标关联与实例对齐成为影响性能的关键因素. 当前的研究主要围绕如何在保持稀疏结构高效性的前提下,实现历史记忆的传递过程. Stream-

PETR^[183] 首次在稀疏 BEV 检测框架中引入流式循环融合机制,其通过运动感知层归一化隐式编码自行车与目标的动态变化,并利用混合自注意力实现时序查询的连续更新. 这种设计在无需显式运动补偿的情况下完成跨帧信息传递,使查询能够在时间维度上保持一致的语义表示,有效缓解了多帧目标失配问题,为基于稀疏查询的检测范式提供了一种新的长时序融合思路,被广泛用在基于查询^[172, 177, 196, 199] 的 BEV 检测工作中.

总体来看,稀疏 BEV 循环融合充分保留了稀疏结构的高效性,同时在避免大量内存开销的同时兼顾长时序的上下文表达能力. 然而目前相关研究较少,亟待开发新的融合策略,以应对长时序知识遗忘、稀疏目标在复杂交通场景下的跨帧一致性等挑战.

3.4.3 稀疏时序融合小结

1) 方法特征与优势. 稀疏时序融合以目标级或前景查询为核心,在保持稀疏推理高效性的前提下实现多帧信息聚合. 其基本思想是在稀疏特征空间内对历史帧与当前帧进行目标级对齐与信息传递,以捕获动态目标的连续运动特征. 并行式融合通常缓存多帧历史传感器特征,检测时同步整合多帧特征;循环式融合则通过时序查询的递归更新隐式传递历史记忆,在不依赖显式运动补偿的情况下实现时序一致的特征演化. 整体而言,稀疏时序融合以轻量、高效为核心优势,在维持推理速度的同时显著提升了跨帧检测稳定性.

2) 主要挑战. 当前稀疏时序融合仍存在多方面瓶颈. 首先,稀疏结构对历史信息的表达能力有限,难以完整保留目标在时间维度的细粒度变化;其次,跨帧目标的精确匹配与对齐仍依赖准确的运动建模,在复杂场景中可能受到检测抖动影响导致时序信息污染. 并行式方法高存储开销限制了其在长时序下的扩展性. 循环式方法易出现特征遗忘与误差累积问题;同时,稀疏查询在时序交互中缺乏显式几何约束,可能会造成训练不稳定.

3) 发展趋势. 未来的稀疏时序融合可朝如下方向进一步优化. 第一,可结合并行式与循环式融合的优势,引入动态识别关键帧机制,采取关键帧并行融合、非关键帧循环融合的混合策略,避免大量冗余帧平均化关键区域特征. 第二,探索可学习时间编码方式,利用 Transformer 结构隐式地融合多帧目标信息,实现对长时序信息的可控保留与动态过滤,但需要注意避免纯数据驱动的学习方式;探索并行-循环混合融合范式或突破现有循环融合范式局限,是构建高效、高精度感知算法的可行路径.

3.5 本节小结

本节系统梳理了基于稀疏 BEV 表示的最新进展, 涵盖不同模态稀疏 BEV 生成以及多模态稀疏 BEV 融合三条主线。点云分支充分利用点云天然稀疏性, 围绕基于中心生成与基于特征扩散的稀疏体素建模展开; 图像分支则以显式 3D 投影采样与隐式 3D 位置编码为两大技术路线, 借助 Transformer 查询机制实现端到端稀疏目标建模; 多模态稀疏融合方法以稀疏实例或查询为中心, 致力于在前景区域进行高效跨模态交互, 从而避免构建任一模态的稠密 BEV 特征。稀疏时序融合遵循类似的并行式与循环式融合范式, 但在稀疏 BEV 表征下面临新的挑战。稀疏 BEV 范式通过只对关键区域建模, 在保持或接近稠密方法精度的同时显著降低了计算与存储开销, 并拓展了远距离与大范围感知的工程可行性。

稀疏 BEV 的主要优势包括以下两点: 首先是高效性, 通过忽略空白背景区域或仅对前景/候选位置建模, 显著减少了计算与内存消耗, 便于在资源受限或实时性要求高的平台部署; 其次是可扩展的感知范围, 稀疏 BEV 策略不依赖于 BEV 网格分辨率, 不必担心远距离感知场景 BEV 网格分辨率过高带来的额外计算资源消耗, 能更容易扩展到更大覆盖范围而不成倍增加成本。

稀疏 BEV 同样也存在显著挑战: 一是特征完整性不足, 跳过稠密 BEV 的构建过程容易导致目标中心特征缺失或背景上下文信息不足, 影响小目标与遮挡目标的检测; 二是全局注意力建模开销, 以目标为核心的稀疏 BEV 构建适配 Transformer 解码架构, 凭借全局注意力的长程建模能力提高了当前方法的精度, 因此稀疏 BEV 范式的高精度方法大都依赖 DETR 检测框架, 然而全局注意力的计算复杂度削弱了稀疏 BEV 范式的计算效率优势, 稠密范式则与之相反, 可以通过更轻量化的特征提

取结构和检测头达到较高精度。第 4 节会讨论另一类 BEV 检测范式, 这类方法基于图像稀疏、点云稠密的混合表达模式, 规避了稀疏体素中较难解决的中心特征缺失问题, 是稠密范式向稀疏范式过渡过程中不可或缺的研究路径。

4 基于稀疏-稠密混合 BEV 表示

稀疏-稠密混合 BEV 表征通过协同利用稠密点云 BEV 特征与原始图像特征交互构建 3D 检测框架, 具有点云分支稠密、图像分支稀疏的特点。其核心特征包括: 1) 点云分支保留稠密 BEV 特征以捕捉精细几何细节; 2) 图像分支通过稀疏查询机制或 ROI 池化操作规避全局 BEV 生成的计算负载; 3) 跨模态特征交互采用注意力驱动或 ROI 特征对齐策略。这类方法包括基于 ROI 特征融合与稀疏查询融合两大技术路线。

4.1 基于 ROI 特征融合

基于 ROI 特征融合的方法通过 3D RPN 网络对目标进行粗略估计, 然后使用 ROI/体素池化操作提取多模态特征进行融合并精细化目标信息, 其检测流程如图 15 所示。此类方法在融合阶段的策略呈现简单拼接、细粒度融合、实例级语义融合的演进过程。

简单拼接融合: 这类方法的核心在于提取同一目标的多模态 ROI 特征, 对齐几何和语义一致性后通过拼接进行融合。MV3D^[200] 用点云 BEV 特征预测 3D 提议, 并将提议投影至图像和点云前视图中, 通过 ROI 池化提取多模态特征; AVOD^[201] 基于预定义 3D 锚点框, 分别在点云 BEV 特征和图像特征上进行 ROI 池化以提取多模态特征; ObjectFusion^[202] 利用点云热力图生成目标提议, 并通过体素池化和 ROI 池化提取每种模态的物体中心特征, 进一步采用单层 Transformer 编码器对 ROI 特征进行上下

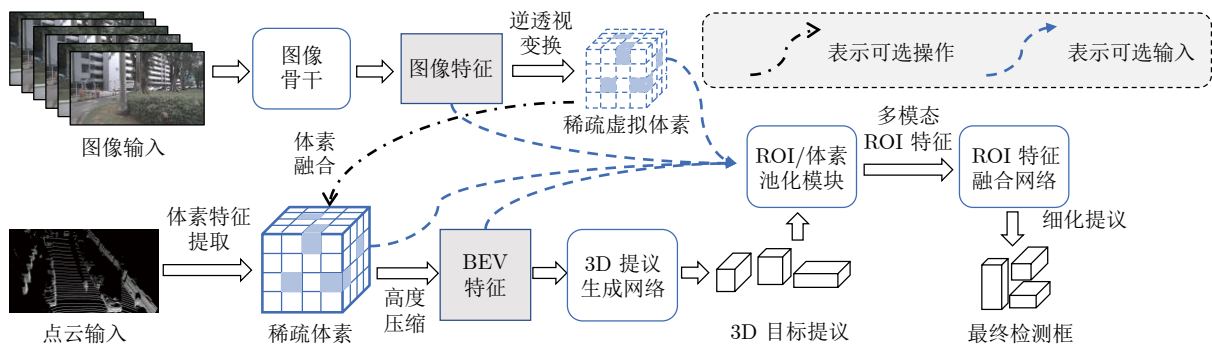


图 15 稀疏-稠密混合 BEV 表示范式中基于 ROI 特征融合的检测框架

Fig. 15 Detection framework based on ROI feature fusion under the sparse-dense hybrid BEV representation paradigm

文建模, 增强表达能力.

细粒度特征增强融合: 由于图像 ROI 区域往往同时包含前景和背景特征, 直接拼接或相加多模态 ROI 特征容易导致模型混淆融合区域中的目标和背景. 针对这个问题, MMF^[203] 利用稀疏深度图生成图像虚拟点云以增强点云 BEV 特征, 再进行 ROI 层级的模态融合, 通过多层次融合策略生成精确的目标表示; SFD^[204] 生成了图像虚拟点云, 通过网格级注意力融合图像体素 ROI 和 LiDAR 体素 ROI 特征, 以避免背景模糊. 由于这类方法未使用离散深度分布, 因此生成的虚拟点云远比 LSS^[68] 稀疏, 很好地保持了图像分支特征的稀疏性.

实例级几何语义融合: 还有少部分方法直接融合 2D、3D 实例, 这类方法可以灵活选择点云检测网络和 2D 检测模型, 以便从最新的检测方法中获益. CLOCs^[205] 将具有几何一致性的目标提议通过卷积网络进行融合; Fast-CLOCs^[206] 将 3D 检测结果投影到图像进行 2D 提议的初始化, 辅助预测 2D 目标的边界框和类别, 促进多模态信息利用.

1) 方法特征与优势. 基于 ROI 特征融合的方法通过在 ROI 层级进行特征对齐与交互, 有效实现了多模态几何-语义特征的互补融合, 提升了目标表征的完备性. 此类方法聚焦于候选区域的细粒度特征提取和融合策略, 避免了全局特征融合的冗余计算, 保证了模型计算效率. 此外, 模块化的设计兼容各类单模态检测框架, 支持灵活组合最新的 2D/3D 检测模型以实现性能持续进化.

2) 主要挑战. 当前基于 ROI 特征融合的方法主要面临三个核心挑战: 第一是模态特征语义混淆, 点云的稀疏几何特征与图像的稠密语义特征在 ROI 层级进行融合, 易引发多模态表征失配, 从而导致融合后的特征产生语义混淆; 第二是累积误差问题, 一阶段提议生成的质量缺陷 (如锚框偏移、误检漏检) 会在后续 ROI 池化中产生几何错位, 导致误差累积到检测头; 第三是标定依赖问题, 图像 ROI 特征提取通常涉及投影操作, 标定误差会造成提议与

ROI 特征无法对应, 错误的前景特征导致模型学习困难.

3) 发展趋势. 未来的研究将围绕几何一致性建模、动态融合机制等方向进一步推动基于 ROI 特征融合方法的演进. 在几何一致性建模方面, 可通过显式引入深度先验、可学习外参校正或不确定性建模机制, 提升 ROI 级特征对齐的精度与鲁棒性, 减轻标定误差与提议偏移带来的累积误差. 在动态融合机制方面, 研究趋势正从拼接或简单加权策略转向基于注意力的自适应融合, 可通过模态相关性评估实现对前景区域的精细建模与多模态间信息的高效融合.

4.2 基于稀疏查询融合

基于稀疏查询的融合方法设置初始查询向量表示目标, 通过与多模态特征进行交叉注意力来实现动态融合, 最终利用前馈网络完成位置、类别以及速度等信息的预测, 其检测流程如图 16 所示. 此类方法延续了稀疏范式下查询即目标的思想, 主要包括全局注意力、局部注意力和实例级融合等策略.

基于全局注意力融合: 部分研究引入全局交叉注意力结构并计算多模态特征序列的注意力权重, 为目标查询自适应筛选并融合多模态特征. TransFusion^[207] 利用点云 BEV 特征生成的热力图引导查询初始化, 并在图像空间构建高斯掩码以约束注意力范围, 实现对目标区域的自适应聚焦. 随后, CMT^[208] 在此基础上引入基于深度的 3D 位置编码以增强几何关联, 使查询能够在全局空间中融合不同模态的语义特征. DeepInteraction 系列工作^[209-210] 进一步提出双分支同步交互结构, 在编码阶段联合学习透视与点云 BEV 特征, 在解码阶段通过交替关注实现跨模态信息双向流动, 从而显著增强了特征一致性与语义互补性.

基于局部注意力融合: 针对全局注意力的计算冗余问题, DETR3D^[169] 所使用的局部注意力建模方式被应用于多模态融合方案中. FUTR3D^[211] 通

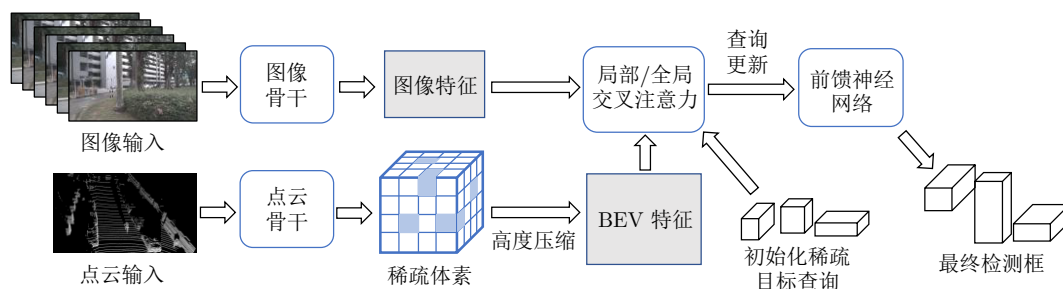


图 16 稀疏-稠密混合 BEV 表示范式中基于稀疏查询融合的检测框架

Fig. 16 Detection framework based on sparse query fusion under the sparse-dense hybrid BEV representation paradigm

过模态无关采样模块将查询映射至不同传感器坐标系中以采样局部特征, 实现基于目标邻域的融合策略。然而基于局部注意力虽然降低了注意力计算量, 但受限于投影精度, 对传感器噪声敏感。SparseLIF^[212]进一步引入透视感知查询生成、ROI 感知采样以及不确定性感知融合, 弥合与稠密 BEV 检测器的性能差距并提升对传感器噪声的鲁棒性。

实例级动态融合: 随着查询机制的发展, 部分研究尝试突破特征级融合的限制, 从实例层面进行多模态语义整合。SparseFusion^[213]基于 TransFusion^[207]与 Deformable-DETR^[93]的结构框架, 生成多模态候选查询, 通过坐标对齐与自注意力交互实现实例级动态融合, 从而更好地保持了语义一致性与几何可解释性。

1) 方法特征与优势. 本类方法通过注意力机制实现跨模态特征的软对齐与自适应聚合, 有效缓解了传感器标定误差导致的几何错位问题。相较于稠密 BEV 融合方法, 稀疏查询机制通过目标导向的特征交互减少了计算冗余。在泛化性方面, 使用模态无关的查询交互框架可兼容激光雷达、相机等多源异构传感器融合, 与基于 ROI 特征融合的方式相比有更好的泛化性。

2) 主要挑战. 现有方法仍面临四大核心挑战: 一是过度依赖点云热力图或随机初始化, 易导致查询空间分布偏离真实目标, 特别是在传感器退化场景下可能引发级联错误; 二是显式特征采样需要密集的坐标投影计算, 而纯注意力驱动的隐式学习方法需要同时建立多视图几何关联和多模态语义关联, 学习难度较大; 三是低质量传感器数据 (如相机过曝、点云缺失等) 可能会通过交叉注意力机制污染全局查询, 影响更新过程; 四是点云模态的稠密 BEV 特征限制了模型进一步轻量化与远距离感知能力。

3) 发展趋势. 基于稀疏查询融合的方法在检测架构上与多模态融合稀疏 BEV 方法高度类似, 均以稀疏的目标查询为多模态融合桥梁, 核心区别在于本类方法中点云模态选择了更规整的稠密 BEV 特征。因此该类方法未来同样可以从优化查询生成机制、探索高效注意力机制等方面入手, 采取与多模态融合稀疏 BEV 方法同样的演进方向, 进一步强化“全稀疏”理念, 构建精度-效率平衡的新检测范式, 推进算法落地应用。

4.3 本节小结

本节系统回顾了基于稀疏-稠密混合 BEV 表示的主要研究路线与代表性方法。该类方法以“稠

密点云提供几何精度、稀疏图像补充语义信息”为核心理念, 主要分为基于 ROI 特征融合与基于稀疏查询融合两大方向。前者通过候选区域级的特征对齐与多模态互补, 提升候选目标的语义完备性与局部精度; 后者则以查询为中心, 通过跨模态注意力机制实现动态融合, 进一步增强了对标定误差的鲁棒性与端到端优化能力。两类方法共同构成了从显式几何对齐到隐式语义关联的技术演进脉络, 展示了混合表示 BEV 范式在精度与效率兼顾下的多模态融合潜力。

稀疏-稠密混合 BEV 表示的主要优势体现在以下几个方面: 一是几何-语义互补性强, 稠密点云 BEV 分支保证了目标特征区域的信息完整性, 图像分支提供了细粒度的语义特征, 使模型能在多传感器感知中实现高精度目标建模; 二是融合粒度具备灵活性, ROI 级或查询级融合机制可在保证目标局部几何-语义一致性的同时自适应调节特征融合范围, 实现对前景目标的精细建模; 三是计算效率较优, 相较于全稠密多模态 BEV, 图像稀疏化处理显著降低了稠密 BEV 特征生成的计算成本, 兼顾精度与实时性。

当前该检测框架仍存在若干挑战。一是模态间几何对齐误差, 由于传感器标定存在一定误差, ROI 级或查询级的特征对齐可能产生偏移, 影响融合效果。二是误差累积与传播, 候选框生成、投影与特征采样阶段的误差会层层传递, 削弱模型的整体鲁棒性。三是注意力建模开销, 跨模态注意力计算在多层交互结构中引入额外计算与显存负担, 限制了实时部署。四是点云模态存在感知范围瓶颈, 虽然避免了图像稠密 BEV 计算过程, 但是点云稠密 BEV 特征仍需要占用较大的显存空间, 因此面临与稠密 BEV 检测相同的感知范围瓶颈问题。总体来看, 稀疏-稠密混合 BEV 表示是自动驾驶 BEV 检测从稠密范式向稀疏范式过渡的承接性工作, 虽然存在时空效率限制, 但是这类方法为高精度全稀疏 BEV 表示的多模态方法提供了大量可借鉴思想, 为下一代多模态感知系统奠定了基础。

5 未来展望

本文从特征表达稀疏性视角系统回顾了当前主流的 BEV 检测方法, 重点剖析了基于稠密 BEV 表示、稀疏 BEV 表示及稀疏-稠密混合 BEV 表示的三类典型范式。在自动驾驶系统架构中, 前端感知模块的检测精度直接影响下游规划决策的可靠性。除了检测精度外, 系统的低延迟特性同样具有关键意义, 尤其在高速行驶场景下, 感知模块的毫秒级

响应能力可为规划控制模块预留充足的决策时间窗口。尽管当前主流 BEV 检测算法在公开数据集上的性能持续突破,但在小尺度目标检测、远距离感知、实时高精度感知以及复杂动态场景适应性等方面仍面临严峻挑战。本节结合最新研究进展分析未来自动驾驶 BEV 检测发展趋势。

5.1 面向全天候工作的多模态鲁棒感知

目前自动驾驶领域的主流多模态感知算法数据集中,主要包括激光雷达和毫米波雷达两类点云传感器。其中激光雷达可生成厘米级精度的 3D 点云,但易受降水、雾霾等复杂气象条件散射干扰,导致点云质量退化;毫米波雷达在雨雪、沙尘等恶劣环境下仍保持稳定探测能力,但缺乏垂直维度信息(或垂直维度分辨率极低),难以满足高精度 3D 感知需求。在真实道路环境中,有光照、雨、雾霾、雪等多种极端工况,但是当前的多模态感知方法受限于传感器劣势,在极端气候条件下表现不佳。近年来兴起的 4D 毫米波雷达技术在保留传统毫米波雷达全天候工作优势的基础上,实现了对高度维信息的有效感知(角度分辨率可达到 $0.5^\circ \sim 1^\circ$)。这种兼具环境适应性与空间感知能力的技术特性,与自动驾驶系统对全天候、高鲁棒性感知的核心需求高度契合。因此,在融合模态中引入 4D 毫米波雷达点云,将会显著提升系统的可用性与安全性。

然而,现有公开数据集(例如 K-Radar^[19])虽然同时涵盖了激光雷达、4D 毫米波雷达以及图像等多种数据源,但仍缺乏系统性的大规模多模态标注数据,严重制约了相关感知算法的迭代优化。此外,基于 4D 毫米波雷达的感知工作仍处于初步探索阶段,亟待开发新的检测框架以利用 4D 毫米波雷达点云的技术特性。因此,未来将进一步构建以 4D 毫米波雷达为主、包含多气候条件及复杂交通场景的大规模基准数据集与多模态融合框架,可以进一步推动面向全天候工作的自动驾驶多模态鲁棒感知系统的发展进程。

5.2 相机参数无关的自适应几何建模框架

传统多相机 BEV 检测方法普遍依赖精确的内外参数标定,以确保像素到 BEV 空间的几何一致性,但在实际自动驾驶系统中,相机姿态扰动、温度变化、机械误差及设备老化等各类因素往往会导致标定参数失效,从而引起特征对齐误差和性能退化。因此,未来的研究还可探索不依赖于相机内外参数的自适应几何感知框架,以新的几何先验引导模型的几何建模与特征融合,兼具几何可解释性与标定噪声鲁棒性,在动态场景中维持感知系统长期稳定

性能,同时还能够降低相机标定与维护成本,从而推动 BEV 感知迈向更灵活、更通用的自适应感知范式。

5.3 高效长时序融合机制探索

时序融合是提升自动驾驶感知系统精度与鲁棒性的核心途径,其目标在于充分利用历史观测实现对动态场景的连续理解。短时序聚合显著提升瞬时精度,而长期历史信息能提供运动趋势、轨迹先验和遮挡补偿(尤其对远距/小目标有关键作用)。然而,现有 BEV 感知模型多聚焦于短时序特征聚合,难以有效利用长期历史信息。在复杂动态环境中,这类方法普遍面临计算与存储开销大、时序信息丢失、长时序跨帧特征错位等瓶颈问题。

未来的研究应从新的角度重新审视时序建模:一方面,可以引入时序稀疏化采样策略,通过关键帧选择机制(如目标遮挡恢复关键帧)降低计算复杂度;另一方面可以从运动建模优化入手,显式分离场景动/静态目标,并进行针对性运动建模,避免传统匀速模型导致的特征空间错位,以强化模型对复杂动态目标的连续感知与预测能力。此外,结合时序置信度评估、记忆压缩等机制,将会推动 BEV 感知从短时局部融合向具备全局时序理解与自主记忆管理能力的方向演进,从而为长时预测与规划提供更稳健的感知支撑。

5.4 稀疏感知算法与边缘硬件协同优化

在实际的自动驾驶感知系统中,不仅需要高精度的检测结果,还需要毫秒级响应能力以及长远距离感知域,以支撑复杂驾驶场景下的安全决策需求。然而,现有 BEV 感知研究多依赖显式 BEV 特征构建或 Transformer 长程建模能力以追求检测精度,导致模型计算复杂度与存储开销居高不下,难以满足远距离、高精度的实时感知要求。在资源受限的车载边缘平台中,这一矛盾尤为突出,形成了算法性能与部署可行性之间的核心瓶颈。

因此,未来研究需从算法-硬件协同的系统层面重构感知架构。一方面,应构建以稀疏前景特征为核心、以轻量化高效特征提取网络为基础的端侧感知新范式,通过精准捕捉关键场景信息降低冗余计算;另一方面,需实现算法设计、编译优化、硬件适配的协同贯通,定制化算子、模型量化与编译优化等软硬件协同技术可深度适配边缘硬件架构特性,进一步提升模型推理效率。该研究方向不仅有助于突破算法性能与车载部署之间的落地鸿沟,更将推动 BEV 感知从单纯的算法性能迭代迈向算法、编译、硬件深度协同的系统级优化,最终实现高精

度、低延迟感知算法在车载边缘平台的工程化落地。

5.5 面向自动驾驶场景的开放世界检测

现有 BEV 感知方法通常假设目标类别封闭、标注完备, 同时现有标注训练数据的场景多样性较为有限。真实交通场景复杂多变且存在大量未知或罕见目标, 例如异常车辆、道路施工设施以及非标准交通参与者, 这对模型的泛化性和安全性提出更高要求。然而, 实际应用中无法通过覆盖所有驾驶场景和类别标注实现泛化能力提升。因此, 随着自动驾驶感知系统逐步应用于开放道路环境, 面向开放世界的 BEV 感知框架已成为研究核心需求, 但当前相关研究仍然处于缺位状态。

开放世界检测旨在使感知系统具备“识别已知类别、发现未知类别、自适应增量学习”的综合能力, 从而在遇到分布外目标时仍能保持稳定决策。未来可结合 BEV 的空间一致性优势, 设计基于不确定性估计、能量建模或对比学习的开放集识别机制, 并探索在线增量学习, 实现对新类别的持续适应与知识更新。此类研究将持续推动 BEV 感知系统迈向真正具备环境理解与自进化能力的开放世界感知范式。

参考文献

- Ma Y X, Wang T, Bai X Y, Yang H T, Hou Y N, Wang Y M, et al. Vision-centric BEV perception: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024, **46**(12): 10978–10997
- Qian R, Lai X, Li X R. 3D object detection for autonomous driving: A survey. *Pattern Recognition*, 2022, **130**: Article No. 108796
- Mao J G, Shi S S, Wang X G, Li H S. 3D object detection for autonomous driving: A comprehensive survey. *International Journal of Computer Vision*, 2023, **131**(8): 1909–1963
- Wang L, Zhang X Y, Song Z Y, Bi J F, Zhang G X, Wei H Y, et al. Multi-modal 3D object detection in autonomous driving: A survey and taxonomy. *IEEE Transactions on Intelligent Vehicles*, 2023, **8**(7): 3781–3798
- Zhang Xin-Yu, Lu Yi-Guo, Gao Xin, Huang Yu-Ning, Liu Hua-Ping, Wang Yun-Peng, et al. Vehicle-road collaborative perception technology and development trend for intelligent connected vehicles. *Acta Automatica Sinica*, 2025, **51**(2): 233–248 (张新钰, 卢毅果, 高鑫, 黄雨宁, 刘华平, 王云鹏, 等. 面向智能网联汽车的车路协同感知技术及发展趋势. *自动化学报*, 2025, **51**(2): 233–248)
- Shen Tian-Yu, Tao Zi-Rui, Wang Ya-Dong, Zhang Ting-Zhen, Liu Yu-Hang, Wang Xing-Xia, et al. Key problems of embodied intelligence research: Autonomous perception, action, and evolution. *Acta Automatica Sinica*, 2025, **51**(1): 43–71 (沈甜雨, 陶子锐, 王亚东, 张庭祯, 刘宇航, 王兴霞, 等. 具身智能研究的关键问题: 自主感知、行动与进化. *自动化学报*, 2025, **51**(1): 43–71)
- Geiger A, Lenz P, Urtasun R. Are we ready for autonomous driving? The KITTI vision benchmark suite. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Providence, USA: IEEE, 2012. 3354–3361
- Huang X Y, Cheng X J, Geng Q C, Cao B B, Zhou D F, Wang P, et al. The ApolloScope dataset for autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Salt Lake City, USA: IEEE, 2018. 954–960
- Choi Y, Kim N, Hwang S, Park K, Yoon J S, An K, et al. KAIST multi-spectral day/night data set for autonomous and assisted driving. *IEEE Transactions on Intelligent Transportation Systems*, 2018, **19**(3): 934–948
- Pham Q H, Sevestre P, Pahwa R S, Zhan H J, Pang C H, Chen Y D, et al. A*3D dataset: Towards autonomous driving in challenging environments. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA). Paris, France: IEEE, 2020. 2267–2273
- Geyer J, Kassahun Y, Mahmudi M, Ricou X, Durgesh R, Chung A S, et al. A2D2: Audi autonomous driving dataset. arXiv preprint arXiv: 2004.06320, 2020.
- Chang M F, Lambert J, Sangkloy P, Singh J, Bak S, Hartnett A, et al. Argoverse: 3D tracking and forecasting with rich maps. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA: IEEE, 2019. 8748–8757
- Patil A, Malla S, Gang H M, Chen Y T. The H3D dataset for full-surround 3D multi-object detection and tracking in crowded urban scenes. In: Proceedings of the International Conference on Robotics and Automation (ICRA). Montreal, Canada: IEEE, 2019. 9552–9557
- Houston J, Zuidhof G, Bergamini L, Ye Y W, Chen L, Jain A, et al. One thousand and one hours: Self-driving motion prediction dataset. In: Proceedings of the 4th Conference on Robot Learning. Cambridge, USA: PMLR, 2020. 409–418
- Caesar H, Bankiti V, Lang A H, Vora S, Liong V E, Xu Q, et al. nuScenes: A multimodal dataset for autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, 2020. 11621–11631
- Sun P, Kretschmar H, Dotiwalla X, Chouard A, Patnaik V, Tsui P, et al. Scalability in perception for autonomous driving: Waymo open dataset. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, 2020. 2446–2454
- Liao Y Y, Xie J, Geiger A. KITTI-360: A novel dataset and benchmarks for urban scene understanding in 2D and 3D. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023, **45**(3): 3292–3310
- Mao J G, Niu M Z, Jiang C H, Liang H X, Chen J H, Liang X D, et al. One million scenes for autonomous driving: ONCE dataset. In: Proceedings of the 1st Neural Information Processing Systems Track on Datasets and Benchmarks. Virtual Event: NeurIPS, 2021.
- Paek D H, Kong S H, Wijaya K T. K-Radar: 4D radar object detection for autonomous driving in various weather conditions. In: Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates, 2022. Article No. 276
- Zheng L Q, Ma Z X, Zhu X C, Tan B, Li S, Long K, et al. TJ4DRadSet: A 4D radar dataset for autonomous driving. In: Proceedings of the 25th International Conference on Intelligent Transportation Systems (ITSC). Macao, China: IEEE, 2022. 493–498
- Zhang X Y, Wang L, Chen J, Fang C, Yang G Q, Wang Y C, et al. Dual radar: A multi-modal dataset with dual 4D radar for autonomous driving. *Scientific Data*, 2025, **12**(1): Article No. 439
- Everingham M, van Gool L, Williams C K I, Winn J, Zisserman A. The Pascal visual object classes (VOC) challenge. *International Journal of Computer Vision*, 2010, **88**(2): 303–338
- Qi C R, Su H, Mo K C, Guibas L J. PointNet: Deep learning on point sets for 3D classification and segmentation. In: Pro-

- ceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA: IEEE, 2017. 652–660
- 24 Qi C R, Yi L, Su H, Guibas L J. PointNet++: Deep hierarchical feature learning on point sets in a metric space. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates, 2017. 5105–5114
 - 25 Shi W J, Rajkumar R. Point-GNN: Graph neural network for 3D object detection in a point cloud. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, 2020. 1711–1719
 - 26 Zhang Y F, Hu Q Y, Xu G Q, Ma Y X, Wan J W, Guo Y L. Not all points are equal: Learning highly efficient point-based detectors for 3D LiDAR point clouds. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE, 2022. 18953–18962
 - 27 Yang Z T, Sun Y N, Liu S, Jia J Y. 3DSSD: Point-based 3D single stage object detector. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, 2020. 11040–11048
 - 28 Qi C R, Litany O, He K M, Guibas L J. Deep Hough voting for 3D object detection in point clouds. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, South Korea: IEEE, 2019. 9277–9286
 - 29 Xie Q, Lai Y K, Wu J, Wang Z T, Zhang Y M, Xu K, et al. MLCVNet: Multi-level context VoteNet for 3D object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, 2020. 10447–10456
 - 30 Xie Q, Lai Y K, Wu J, Wang Z T, Lu D N, Wei M Q, et al. VENet: Voting enhancement network for 3D object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, Canada: IEEE, 2021. 3712–3721
 - 31 Zhou Y, Tuzel O. VoxelNet: End-to-end learning for point cloud based 3D object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018. 4490–4499
 - 32 Ren S Q, He K M, Girshick R, Sun J. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017, **39**(6): 1137–1149
 - 33 Yan Y, Mao Y X, Li B. SECOND: Sparsely embedded convolutional detection. *Sensors*, 2018, **18**(10): Article No. 3337
 - 34 He C H, Zeng H, Huang J Q, Hua X S, Zhang L. Structure aware single-stage 3D object detection from point cloud. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, 2020. 11873–11882
 - 35 Fan L, Pang Z Q, Zhang T Y, Wang Y X, Zhao H, Wang F, et al. Embracing single stride 3D object detector with sparse Transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE, 2022. 8458–8468
 - 36 Shi S S, Guo C X, Jiang L, Wang Z, Shi J P, Wang X G, et al. PV-RCNN: Point-voxel feature set abstraction for 3D object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, 2020. 10529–10538
 - 37 Shi S S, Jiang L, Deng J J, Wang Z, Guo C X, Shi J P, et al. PV-RCNN++: Point-voxel feature set abstraction with local vector representation for 3D object detection. *International Journal of Computer Vision*, 2023, **131**(2): 531–551
 - 38 Deng J J, Shi S S, Li P W, Zhou W G, Zhang Y Y, Li H Q. Voxel R-CNN: Towards high performance voxel-based 3D object detection. In: Proceedings of the 35th AAAI Conference on Artificial Intelligence. Virtual Event: AAAI, 2021. 1201–1209
 - 39 Neubeck A, van Gool L. Efficient non-maximum suppression. In: Proceedings of the 18th International Conference on Pattern Recognition. Hong Kong, China: IEEE, 2006. 850–855
 - 40 Yin T W, Zhou X Y, Krähenbühl P. Center-based 3D object detection and tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE, 2021. 11784–11793
 - 41 Zhou X Y, Wang D Q, Krähenbühl P. Objects as points. arXiv preprint arXiv: 1904.07850, 2019.
 - 42 Ma R Q, Chen C, Yang B S, Li D R, Wang H P, Cong Y Z, et al. CG-SSD: Corner guided single stage 3D object detection from LiDAR point cloud. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2022, **191**: 33–48
 - 43 Ge R Z, Ding Z Z, Hu Y H, Wang Y, Chen S J, Huang L, et al. AFDet: Anchor free one stage 3D object detection. arXiv preprint arXiv: 2006.12671, 2020.
 - 44 Hu Y H, Ding Z Z, Ge R Z, Shao W X, Huang L, Li K, et al. AFDetV2: Rethinking the necessity of the second stage for object detection from point clouds. In: Proceedings of the 36th AAAI Conference on Artificial Intelligence. Virtual Event: AAAI, 2022. 969–979
 - 45 Wang Y, Solomon J M. Object DGCNN: 3D object detection using dynamic graphs. In: Proceedings of the 35th International Conference on Neural Information Processing Systems. Virtual Event: Curran Associates, 2021. Article No. 1587
 - 46 Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, et al. Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates, 2017. 6000–6010
 - 47 Mao J G, Xue Y J, Niu M Z, Bai H Y, Feng J S, Liang X D, et al. Voxel Transformer for 3D object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, Canada: IEEE, 2021. 3164–3173
 - 48 He C H, Li R H, Li S, Zhang L. Voxel set Transformer: A set-to-set approach to 3D object detection from point clouds. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE, 2022. 8417–8427
 - 49 Wang H Y, Shi C, Shi S S, Lei M, Wang S, He D, et al. DSVT: Dynamic sparse voxel Transformer with rotated sets. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE, 2023. 13520–13529
 - 50 Li J N, Dong S C, Ding L H, Xu T F. MsSVT++: Mixed-scale sparse voxel Transformer with center voting for 3D object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024, **46**(5): 3736–3752
 - 51 He C H, Li R H, Zhang G W, Zhang L. ScatterFormer: Efficient voxel Transformer with scattered linear attention. In: Proceedings of the 18th European Conference on Computer Vision. Milan, Italy: Springer, 2025. 74–92
 - 52 Lang A H, Vora S, Caesar H, Zhou L B, Yang J, Beijbom O. PointPillars: Fast encoders for object detection from point clouds. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA: IEEE, 2019. 12697–12705
 - 53 Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu C Y, et al. SSD: Single shot MultiBox detector. In: Proceedings of the 14th European Conference on Computer Vision. Amsterdam, The Netherlands: Springer, 2016. 21–37
 - 54 Wang Y, Fathi A, Kundu A, Ross D A, Pantofaru C, Funkhouser T, et al. Pillar-based object detection for autonomous driving. In: Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer, 2020. 18–34
 - 55 Zhou S F, Yuan Z H, Yang D W, Zhao Z Y, Hu X, Shi Y G, et al.

- Information entropy guided height-aware histogram for quantization-friendly pillar feature encoder. arXiv preprint arXiv: 2405.18734, 2024.
- 56 Li X S, Wang C L, Wang S M, Zeng Z, Liu J. PillarNeXt: Improving the 3D detector by introducing Voxel2Pillar feature encoding and extracting multi-scale features. arXiv preprint arXiv: 2405.09828, 2024.
 - 57 Wang C, Liu Z W. CAFI-pillars: Infusing geometry priors for pillar-based 3D detectors through centroid-aware feature interaction. *IEEE Transactions on Intelligent Vehicles*, 2024, 9(1): 2399–2408
 - 58 Guo D B, Yang G H, Wang C H. PillarNet++: Pillar-based 3D object detection with multiattention. *IEEE Sensors Journal*, 2023, 23(22): 27733–27743
 - 59 Li J Y, Luo C X, Yang X D. PillarNeXt: Rethinking network designs for 3D object detection in LiDAR point clouds. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE, 2023. 17567–17576
 - 60 He K M, Zhang X Y, Ren S Q, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA: IEEE, 2016. 770–778
 - 61 Chen L C, Papandreou G, Kokkinos I, Murphy K, Yuille A L. DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018, 40(4): 834–848
 - 62 Wang J, Lan S Y, Gao M F, Davis L S. InfoFocus: 3D object detection for autonomous driving with dynamic information modeling. In: Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer, 2020. 405–420
 - 63 Shen Y, Zhang Y Z, Wu Y M, Wang Z Y, Yang L H, Coleman S, et al. BSH-Det3D: Improving 3D object detection with BEV shape heatmap. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Detroit, USA: IEEE, 2023. 5730–5737
 - 64 Huang Z C, Zheng Z J, Zhao J W, Hu H F, Wang Z X, Chen D H. PSA-Det3D: Pillar set abstraction for 3D object detection. *Pattern Recognition Letters*, 2023, 168: 138–145
 - 65 Zhou S F, Zhang X Y, Chu X X, Zhang B, Zhao Z Y, Lu X B, et al. FastPillars: A deployment-friendly pillar-based 3D detector. *IEEE Transactions on Circuits and Systems for Video Technology*, 2026, 36(4): 4999–5010
 - 66 Mao W X, Wang T C, Zhang D K, Yan J J, Yoshie O. PillarNeSt: Embracing backbone scaling and pretraining for pillar-based 3D object detection. *IEEE Transactions on Intelligent Vehicles*, 2025, 10(4): 2905–2914
 - 67 Liu Z, Mao H Z, Wu C Y, Feichtenhofer C, Darrell T, Xie S N. A ConvNet for the 2020s. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE, 2022. 11976–11986
 - 68 Philion J, Fidler S. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3D. In: Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer, 2020. 194–210
 - 69 Huang J J, Huang G, Zhu Z, Ye Y, Du D L. BEVDet: High-performance multi-camera 3D object detection in bird-eye-view. arXiv preprint arXiv: 2112.11790, 2021.
 - 70 Xie E Z, Yu Z D, Zhou D Q, Philion J, Anandkumar A, Fidler S, et al. M²BEV: Multi-camera joint 3D detection and segmentation with unified birds-eye view representation. arXiv preprint arXiv: 2204.05088, 2022.
 - 71 Huang B, Li Y G, Xie E Z, Liang F, Wang L Y, Shen M Z, et al. Fast-BEV: Towards real-time on-vehicle bird's-eye view perception. arXiv preprint arXiv: 2301.07870, 2023.
 - 72 Li Y G, Huang B, Chen Z R, Cui Y F, Liang F, Shen M Z, et al. Fast-BEV: A fast and strong bird's-eye view perception baseline. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024, 46(12): 8665–8679
 - 73 Huang J J, Huang G. BEVPoolv2: A cutting-edge implementation of BEVDet toward deployment. arXiv preprint arXiv: 2211.17111, 2022.
 - 74 Li Y X, Han Q, Yu M Y, Jiang Y X, Yeo C K, Li Y H, et al. Towards efficient 3D object detection in bird's-eye-space for autonomous driving: A convolutional-only approach. In: Proceedings of the IEEE 26th International Conference on Intelligent Transportation Systems. Bilbao, Spain: IEEE, 2023. 2170–2177
 - 75 Li Y H, Ge Z, Yu G Y, Yang J R, Wang Z R, Shi Y K, et al. BEVDepth: Acquisition of reliable depth for multi-view 3D object detection. In: Proceedings of the 37th AAAI Conference on Artificial Intelligence. Washington, USA: AAAI, 2023. 1477–1485
 - 76 Li Y H, Bao H, Ge Z, Yang J R, Sun J J, Li Z M. BEVStereo: Enhancing depth estimation in multi-view 3D object detection with temporal stereo. In: Proceedings of the 37th AAAI Conference on Artificial Intelligence. Washington, USA: AAAI, 2023. 1486–1494
 - 77 Li Y H, Yang J R, Sun J J, Bao H, Ge Z, Xiao L. BEVStereo++: Accurate depth estimation in multi-view 3D object detection via dynamic temporal stereo. arXiv preprint arXiv: 2304.04185, 2023.
 - 78 Li Z X, Lan S Y, Alvarez J M, Wu Z X. BEVNeXt: Reviving dense BEV frameworks for 3D object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, 2024. 20113–20123
 - 79 Man Y Z, Gui L Y, Wang Y X. DualCross: Cross-modality cross-domain adaptation for monocular BEV perception. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Detroit, USA: IEEE, 2023. 10910–10917
 - 80 Chen Z H, Li Z Y, Zhang S Q, Fang L J, Jiang Q H, Zhao F. BEVDistill: Cross-modal BEV distillation for multi-view 3D object detection. In: Proceedings of the 11th International Conference on Learning Representations. Kigali, Rwanda: OpenReview.net, 2023.
 - 81 Zhou S C, Liu W Z, Hu C, Zhou S C, Ma C. UniDistill: A universal cross-modality knowledge distillation framework for 3D object detection in bird's-eye view. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE, 2023. 5116–5125
 - 82 Wang Z Y, Li D W, Luo C X, Xie C H, Yang X D. DistillBEV: Boosting multi-camera 3D object detection with cross-modal knowledge distillation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France: IEEE, 2023. 8637–8646
 - 83 Zhao H M, Zhang Q M, Zhao S S, Chen Z, Zhang J, Tao D C. SimDistill: Simulated multi-modal distillation for BEV 3D object detection. In: Proceedings of the 38th AAAI Conference on Artificial Intelligence. Vancouver, Canada: AAAI, 2024. 7460–7468
 - 84 Jiang Z, Zhang J Q, Zhang Y N, Liu Q J, Hu Z H, Wang B H, et al. FSD-BEV: Foreground self-distillation for multi-view 3D object detection. In: Proceedings of the 18th European Conference on Computer Vision. Milan, Italy: Springer, 2025. 110–126
 - 85 Zhang J Q, Zhang Y N, Qi Y L, Fu Z H, Liu Q J, Wang Y H. GeoBEV: Learning geometric BEV representation for multi-view 3D object detection. In: Proceedings of the 39th AAAI Conference on Artificial Intelligence. Philadelphia, USA: AAAI, 2025. 9960–9968
 - 86 Xu S Q, Li F, Huang P X, Song Z Y, Yang Z X. TiGDistill-BEV: Multi-view BEV 3D object detection via target inner-geometry learning distillation. *IEEE Transactions on Circuits*

- and Systems for Video Technology*, 2026, **36**(1): 846–860
- 87 Alhamwi A, Vandeportael B, Piat J. Real time vision system for obstacle detection and localization on FPGA. In: Proceedings of the 10th International Conference on Computer Vision Systems. Copenhagen, Denmark: Springer, 2015. 80–90
 - 88 Kim Y, Kum D. Deep learning based vehicle position and orientation estimation via inverse perspective mapping image. In: Proceedings of the IEEE Intelligent Vehicles Symposium (IV). Paris, France: IEEE, 2019. 317–323
 - 89 Reiher L, Lampe B, Eckstein L. A Sim2Real deep learning approach for the transformation of images from multiple vehicle-mounted cameras to a semantically segmented image in bird's eye view. In: Proceedings of the IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC). Rhodes, Greece: IEEE, 2020. 1–7
 - 90 Roddick T, Kendall A, Cipolla R. Orthographic feature transform for monocular 3D object detection. In: Proceedings of the 30th British Machine Vision Conference. Cardiff, UK: BMVA Press, 2019.
 - 91 Li Z Q, Wang W H, Li H Y, Xie E Z, Sima C H, Lu T, et al. BEVFormer: Learning bird's-eye-view representation from LiDAR-camera via spatiotemporal Transformers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025, **47**(3): 2020–2036
 - 92 Yang C Y, Chen Y T, Tian H, Tao C X, Zhu X Z, Zhang Z X, et al. BEVFormer v2: Adapting modern image backbones to bird's-eye-view recognition via perspective supervision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE, 2023. 17830–17839
 - 93 Zhu X Z, Su W J, Lu L W, Li B, Wang X G, Dai J F. Deformable DETR: Deformable Transformers for end-to-end object detection. In: Proceedings of the 9th International Conference on Learning Representations. Virtual Event: OpenReview.net, 2021.
 - 94 Chen S Y, Cheng T H, Wang X G, Meng W M, Zhang Q, Liu W Y. Efficient and robust 2D-to-BEV representation learning via geometry-guided kernel Transformer. arXiv preprint arXiv: 2206.04584, 2022.
 - 95 Wu Y M, Li R X, Qin Z Q, Zhao X H, Li X. HeightFormer: Explicit height modeling without extra data for camera-only 3D object detection in bird's eye view. *IEEE Transactions on Image Processing*, 2025, **34**: 689–700
 - 96 Cai H X, Zhang Z Y, Zhou Z Y, Li Z Y, Ding W B, Zhao J H. BEVFusion4D: Learning LiDAR-camera fusion under bird's-eye-view via cross-modality guidance and temporal aggregation. arXiv preprint arXiv: 2303.17099, 2023.
 - 97 Man Y Z, Gui L Y, Wang Y X. BEV-guided multi-modality fusion for driving perception. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE, 2023. 21960–21969
 - 98 Chen Y L, Yu Z D, Chen Y K, Lan S Y, Anandkumar A, Jia J Y, et al. FocalFormer3D: Focusing on hard instance for 3D object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France: IEEE, 2023. 8394–8405
 - 99 Gunn J, Lenyk Z, Sharma A, Donati A, Buburuzan A, Redford J, et al. Lift-attend-splat: Bird's-eye-view camera-LiDAR fusion using Transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Seattle, USA: IEEE, 2024. 4526–4536
 - 100 Wang S M, Caesar H, Nan L L, Kooij J F P. UniBEV: Multi-modal 3D object detection with uniform BEV encoders for robustness against missing sensor modalities. In: Proceedings of the IEEE Intelligent Vehicles Symposium (IV). Jeju Island, South Korea: IEEE, 2024. 2776–2783
 - 101 Yin J B, Shen J B, Chen R N, Li W, Yang R G, Frossard P, et al. IS-fusion: Instance-scene collaborative fusion for multimodal 3D object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, 2024. 14905–14915
 - 102 Saha A, Mendez O, Russell C, Bowden R. Translating images into maps. In: Proceedings of the International Conference on Robotics and Automation (ICRA). Philadelphia, USA: IEEE, 2022. 9200–9206
 - 103 Jiang Y Q, Zhang L, Miao Z W, Zhu X T, Gao J, Hu W M, et al. PolarFormer: Multi-camera 3D object detection with polar Transformer. In: Proceedings of the 37th AAAI Conference on Artificial Intelligence. Washington, USA: AAAI, 2023. 1042–1050
 - 104 Yang H T, Bai X Y, Zhu X G, Ma Y X. One training for multiple deployments: Polar-based adaptive BEV perception for autonomous driving. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA). London, UK: IEEE, 2023. 5602–5609
 - 105 Zhao J W, Jiang Q X, Li X D, Luo J F. Focus on BEV: Self-calibrated cycle view transformation for monocular birds-eye-view segmentation. arXiv preprint arXiv: 2410.15932, 2024.
 - 106 Pan B W, Sun J K, Leung H Y T, Andonian A, Zhou B L. Cross-view semantic segmentation for sensing surroundings. *IEEE Robotics and Automation Letters*, 2020, **5**(3): 4867–4873
 - 107 Li Q, Wang Y, Wang Y L, Zhao H. HDMapNet: An online HD map construction and evaluation framework. In: Proceedings of the International Conference on Robotics and Automation (ICRA). Philadelphia, USA: IEEE, 2022. 4628–4634
 - 108 Roddick T, Cipolla R. Predicting semantic map representations from images using pyramid occupancy networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, 2020. 11138–11147
 - 109 Chitta K, Prakash A, Geiger A. NEAT: Neural attention fields for end-to-end autonomous driving. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, Canada: IEEE, 2021. 15793–15803
 - 110 Yang W X, Li Q, Liu W X, Yu Y L, Ma Y X, He S F, et al. Projecting your view attentively: Monocular road scene layout estimation via cross-view transformation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE, 2021. 15536–15545
 - 111 Liu W X, Li Q, Yang W X, Cai J X, Yu Y L, Ma Y X, et al. Monocular BEV perception of road scenes via front-to-top view projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024, **46**(9): 6109–6125
 - 112 Fan S Q, Wang Z, Huo X L, Wang Y, Liu J J. Calibration-free BEV representation for infrastructure perception. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Detroit, USA: IEEE, 2023. 9008–9013
 - 113 Zhou B, Krähenbühl P. Cross-view Transformers for real-time map-view semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE, 2022. 13760–13769
 - 114 Bartoccioni F, Zablocki E, Bursuc A, Pérez P, Cord M, Alahari K. LaRa: Latents and rays for multi-camera bird's-eye-view semantic segmentation. In: Proceedings of the 6th Conference on Robot Learning. Auckland, New Zealand: PMLR, 2023. 1663–1672
 - 115 Peng L, Chen Z R, Fu Z J, Liang P P, Cheng E K. BEVsegFormer: Bird's eye view semantic segmentation from arbitrary camera rigs. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). Waikoloa, USA: IEEE, 2023. 5935–5943
 - 116 Liu Z J, Tang H T, Amini A, Yang X Y, Mao H Z, Rus D L, et al. BEVFusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. In: Proceedings of the IEEE In-

- international Conference on Robotics and Automation (ICRA). London, UK: IEEE, 2023. 2774–2781
- 117 Huang J J, Ye Y, Liang Z J, Shan Y, Du D L. Detecting as labeling: Rethinking LiDAR-camera fusion in 3D object detection. In: Proceedings of the 18th European Conference on Computer Vision. Milan, Italy: Springer, 2025. 439–455
- 118 Fu J H, Gao C, Wang Z T, Yang L R, Wang X F, Mu B P, et al. Eliminating cross-modal conflicts in BEV space for LiDAR-camera 3D object detection. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA). Yokohama, Japan: IEEE, 2024. 16381–16387
- 119 Zhu X Z, Xiong Y W, Dai J F, Yuan L, Wei Y C. Deep feature flow for video recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA: IEEE, 2017. 2349–2358
- 120 Li X T, You A S, Zhu Z, Zhao H L, Yang M K, Yang K Y, et al. Semantic flow for fast and accurate scene parsing. In: Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer, 2020. 775–793
- 121 Hu H T, Wang F Y, Su J W, Wang Y N, Hu L F, Fang W Y, et al. EA-LSS: Edge-aware lift-splat-shot framework for 3D BEV object detection. arXiv preprint arXiv: 2303.17895, 2023.
- 122 Song Z Y, Yang L, Xu S Q, Liu L, Xu D Y, Jia C Y, et al. GraphBEV: Towards robust BEV feature alignment for multi-modal 3D object detection. In: Proceedings of the 18th European Conference on Computer Vision. Milan, Italy: Springer, 2025. 347–366
- 123 Liang T T, Xie H W, Yu K C, Xia Z X, Lin Z W, Wang Y T, et al. BEVFusion: A simple and robust LiDAR-camera fusion framework. In: Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates, 2022. Article No. 757
- 124 Zeng Y H, Zhang D, Wang C W, Miao Z W, Liu T, Zhan X, et al. LIFT: Learning 4D LiDAR image fusion Transformer for 3D object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE, 2022. 17172–17181
- 125 Li X T, Fan B J, Tian J D, Fan H J. GAFusion: Adaptive fusing LiDAR and camera with multiple guidance for 3D object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, 2024. 21209–21218
- 126 Chen M, Yang M C, Zhang Y, Han T, Li X C, Zhao H C, et al. Multi-modal BEV enhancement fusion for 3D object detection in autonomous driving. *IEEE Transactions on Intelligent Transportation Systems*, 2025, **26**(11): 18447–18459
- 127 Vora S, Lang A H, Helou B, Beijbom O. PointPainting: Sequential fusion for 3D object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, 2020. 4604–4612
- 128 Yin T W, Zhou X Y, Krähenbühl P. Multimodal virtual point 3D detection. In: Proceedings of the 35th International Conference on Neural Information Processing Systems. Virtual Event: Curran Associates, 2021. Article No. 1261
- 129 Xu S Q, Zhou D F, Fang J, Yin J B, Bin Z, Zhang L J. FusionPainting: Multimodal fusion with adaptive attention for 3D object detection. In: Proceedings of the IEEE International Intelligent Transportation Systems Conference (ITSC). Indianapolis, USA: IEEE, 2021. 3047–3054
- 130 Zhu H Q, Deng J J, Zhang Y, Ji J M, Mao Q Y, Li H Q, et al. VPFNet: Improving 3D object detection with virtual point based LiDAR and stereo data fusion. *IEEE Transactions on Multimedia*, 2023, **25**: 5291–5304
- 131 Mahmoud A, Hu J S K, Waslander S L. Dense voxel fusion for 3D object detection. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). Waikoloa, USA: IEEE, 2023. 663–672
- 132 Chen Y K, Li Y W, Zhang X Y, Sun J, Jia J Y. Focal sparse convolutional networks for 3D object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE, 2022. 5428–5437
- 133 Chen Y K, Liu J H, Zhang X Y, Qi X J, Jia J Y. LargeKernel3D: Scaling up kernels in 3D sparse CNNs. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE, 2023. 13488–13498
- 134 Li Y W, Qi X J, Chen Y K, Wang L W, Li Z M, Sun J, et al. Voxel field fusion for 3D object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE, 2022. 1120–1129
- 135 Li Y W, Chen Y L, Qi X J, Li Z M, Sun J, Jia J Y. Unifying voxel-based representation with Transformer for 3D object detection. In: Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates, 2022. Article No. 1340
- 136 Jiao Y, Jie Z Q, Chen S X, Chen J J, Ma L, Jiang Y G. MSMDFusion: Fusing LiDAR and camera at multiple scales with multi-depth seeds for 3D object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE, 2023. 21643–21652
- 137 Li Y W, Yu A W, Meng T J, Caine B, Ngiam J, Peng D Y, et al. DeepFusion: LiDAR-camera deep fusion for multi-modal 3D object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE, 2022. 17182–17191
- 138 Chen Z H, Li Z Y, Zhang S Q, Fang L J, Jiang Q H, Zhao F, et al. AutoAlign: Pixel-instance feature aggregation for multi-modal 3D object detection. In: Proceedings of the 31st International Joint Conference on Artificial Intelligence. Vienna, Austria: ijcai.org, 2022.
- 139 Chen Z H, Li Z Y, Zhang S Q, Fang L J, Jiang Q H, Zhao F. Deformable feature aggregation for dynamic multi-modal 3D object detection. In: Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer, 2022. 628–644
- 140 Li X, Shi B T, Hou Y N, Wu X J, Ma T L, Li Y K, et al. Homogeneous multi-modal feature fusion and interaction for 3D object detection. In: Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer, 2022. 691–707
- 141 Wang H Y, Tang H, Shi S S, Li A X, Li Z G, Schiele B, et al. UniTR: A unified and efficient multi-modal Transformer for bird's-eye-view representation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France: IEEE, 2023. 6792–6802
- 142 Huang J J, Huang G. BEVDet4D: Exploit temporal cues in multi-camera 3D object detection. arXiv preprint arXiv: 2203.17054, 2022.
- 143 Zhang Y P, Zhu Z, Zheng W Z, Huang J J, Huang G, Zhou J, et al. BEVerse: Unified perception and prediction in birds-eye-view for vision-centric autonomous driving. arXiv preprint arXiv: 2205.09743, 2022.
- 144 Hu A, Murez Z, Mohan N, Dudas S, Hawke J, Badrinarayanan V, et al. FIERY: Future instance prediction in bird's-eye view from surround monocular cameras. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, Canada: IEEE, 2021. 15273–15282
- 145 Hu S C, Chen L, Wu P H, Li H Y, Yan J C, Tao D C. ST-P3: End-to-end vision-based autonomous driving via spatial-temporal feature learning. In: Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer, 2022. 533–549

- 146 Park J, Xu C F, Yang S J, Keutzer K, Kitani K M, Tomizuka M, et al. Time will tell: New outlooks and a baseline for temporal multi-view 3D object detection. In: Proceedings of the 11th International Conference on Learning Representations. Kigali, Rwanda: OpenReview.net, 2023.
- 147 Yao Y, Luo Z X, Li S W, Fang T, Quan L. MVSNet: Depth inference for unstructured multi-view stereo. In: Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer, 2018. 767–783
- 148 Zong Z F, Jiang D Z, Song G L, Xue Z Y, Su J Y, Li H S, et al. Temporal enhanced training of multi-view 3D object detector via historical object prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France: IEEE, 2023. 3781–3790
- 149 Xia Z Y, Lin Z W, Wang X H, Wang Y T, Xing Y, Qi S X, et al. HENet: Hybrid encoding for end-to-end multi-task 3D perception from multi-view cameras. In: Proceedings of the 18th European Conference on Computer Vision. Milan, Italy: Springer, 2025. 376–392
- 150 Qin Z Q, Chen J Y, Chen C, Chen X Z, Li X. UniFusion: Unified multi-view fusion Transformer for spatial-temporal representation in bird's-eye-view. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France: IEEE, 2023. 8690–8699
- 151 Zaremba W, Sutskever I, Vinyals O. Recurrent neural network regularization. arXiv preprint arXiv: 1409.2329, 2014.
- 152 Han C R, Yang J R, Sun J J, Ge Z, Dong R P, Zhou H Y, et al. Exploring recurrent long-term temporal fusion for multi-view 3D perception. *IEEE Robotics and Automation Letters*, 2024, **9**(7): 6544–6551
- 153 Hu C Y, Zheng H, Li K, Xu J Y, Mao W B, Luo M C, et al. FusionFormer: A multi-sensory fusion in bird's-eye-view and temporal consistent Transformer for 3D object detection. arXiv preprint arXiv: 2309.05257, 2023.
- 154 Chang M, Zhang X S, Zhang R, Zhao Z P, He G H, Liu S L. RecurrentBEV: A long-term temporal fusion framework for multi-view 3D detection. In: Proceedings of the 18th European Conference on Computer Vision. Milan, Italy: Springer, 2025. 131–147
- 155 Saini L, Meuter M, Tercan H, Meisen T. AttentiveGRU: Recurrent spatio-temporal modeling for advanced radar-based BEV object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Nashville, USA: IEEE, 2025. 2415–2424
- 156 Chen Y K, Liu J H, Zhang X Y, Qi X J, Jia J Y. VoxelNeXt: Fully sparse VoxelNet for 3D object detection and tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE, 2023. 21674–21683
- 157 Fan L, Wang F, Wang N Y, Zhang Z X. Fully sparse 3D object detection. In: Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates, 2022. Article No. 26
- 158 Fan L, Yang Y X, Wang F, Wang N Y, Zhang Z X. Super sparse 3D object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023, **45**(10): 12490–12505
- 159 Fan L, Wang F, Wang N Y, Zhang Z X. FSD V2: Improving fully sparse 3D object detection with virtual voxels. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025, **47**(2): 1279–1292
- 160 Zhang G, Chen J N, Gao G H, Li J M, Liu S, Hu X L. SAFD-Net: A simple and effective network for fully sparse 3D object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, 2024. 14477–14486
- 161 Yuan Y S, Xia Y, Cremers D, Sester M. SparseAlign: A fully sparse framework for cooperative object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE, 2025. 22296–22305
- 162 Sun P, Tan M X, Wang W Y, Liu C X, Xia F, Leng Z Q, et al. SWFormer: Sparse window Transformer for 3D object detection in point clouds. In: Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer, 2022. 426–442
- 163 Liu L, Song Z Y, Xia Q M, Jia F Y, Jia C Y, Yang L, et al. SparseDet: A simple and effective framework for fully sparse LiDAR-based 3-D object detection. *IEEE Transactions on Geoscience and Remote Sensing*, 2024, **62**: Article No. 5707114
- 164 Liu S, Cui M Y, Li B Y, Liang Q M, Hong T H, Huang K, et al. FSHNet: Fully sparse hybrid network for 3D object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE, 2025. 8900–8909
- 165 Lee M, Park S, Kim H, Yoon M, Lee J, Choi J W, et al. SPADE: Sparse pillar-based 3D object detection accelerator for autonomous driving. In: Proceedings of the IEEE International Symposium on High-performance Computer Architecture (HPCA). Edinburgh, UK: IEEE, 2024. 454–467
- 166 Wang T, Zhu X G, Pang J M, Lin D H. FCOS3D: Fully convolutional one-stage monocular 3D object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW). Montreal, Canada: IEEE, 2021. 913–922
- 167 Tian Z, Shen C H, Chen H, He T. FCOS: A simple and strong anchor-free object detector. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022, **44**(4): 1922–1933
- 168 Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A, Zagoruyko S. End-to-end object detection with Transformers. In: Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer, 2020. 213–229
- 169 Wang Y, Guizilini V, Zhang T Y, Wang Y L, Zhao H, Solomon J. DETR3D: 3D object detection from multi-view images via 3D-to-2D queries. In: Proceedings of the 5th Conference on Robot Learning. London, UK: PMLR, 2021. 180–191
- 170 Chen S Y, Wang X G, Cheng T H, Zhang Q, Huang C, Liu W Y. PolarDETR: Polar parametrization for vision-based surround-view 3D detection. *Image and Vision Computing*, 2025, **156**: Article No. 105438
- 171 Lin X W, Lin T W, Pei Z X, Huang L C, Su Z Z. Sparse4D: Multi-view 3D object detection with sparse spatial-temporal fusion. arXiv preprint arXiv: 2211.10581, 2022.
- 172 Lin X W, Lin T W, Pei Z X, Huang L C, Su Z Z. Sparse4D v2: Recurrent temporal fusion with sparse model. arXiv preprint arXiv: 2305.14018, 2023.
- 173 Chen Z H, Li Z Y, Zhang S Q, Fang L J, Jiang Q H, Zhao F. Graph-DETR3D: Rethinking overlapping regions for multi-view 3D object detection. In: Proceedings of the 30th ACM International Conference on Multimedia. Lisbon, Portugal: ACM, 2022. 5999–6008
- 174 Chen Z H, Chen Z, Li Z Y, Zhang S Q, Fang L J, Jiang Q H, et al. Graph-DETR4D: Spatio-temporal graph modeling for multi-view 3D object detection. *IEEE Transactions on Image Processing*, 2024, **33**: 4488–4500
- 175 Liu H S, Teng Y, Lu T, Wang H G, Wang L M. SparseBEV: High-performance sparse 3D object detection from multi-camera videos. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France: IEEE, 2023. 18580–18590
- 176 Li H Y, Zhang H, Zeng Z Y, Liu S L, Li F, Ren T H, et al. DFA3D: 3D deformable attention for 2D-to-3D feature lifting. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France: IEEE, 2023. 6684–6693
- 177 Lin X W, Pei Z X, Lin T W, Huang L C, Su Z Z. Sparse4D v3: Advancing end-to-end 3D detection and tracking. arXiv preprint arXiv: 2311.11722, 2023.
- 178 Xu L Z, Wu Z H, Qiu W Z, Pang S M, Bai X X, Mei K Z, et al. Redundant queries in DETR-based 3D detection methods: Unnecessary and prunable. In: Proceedings of the 40th AAAI

- Conference on Artificial Intelligence. Singapore: AAAI, 2026.
- 179 Tian Yong-Lin, Wang Yu-Tong, Wang Jian-Gong, Wang Xiao, Wang Fei-Yue. Key problems and progress of vision Transformers: The state of the art and prospects. *Acta Automatica Sinica*, 2022, **48**(4): 957–979
(田永林, 王雨桐, 王建功, 王晓, 王飞跃. 视觉 Transformer 研究的关键问题: 现状及展望. *自动化学报*, 2022, **48**(4): 957–979)
 - 180 Liu Y F, Wang T C, Zhang X Y, Sun J. PETR: Position embedding transformation for multi-view 3D object detection. In: Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer, 2022. 531–548
 - 181 Wang S H, Jiang X H, Li Y. Focal-PETR: Embracing foreground for efficient multi-camera 3D object detection. *IEEE Transactions on Intelligent Vehicles*, 2024, **9**(1): 1481–1489
 - 182 Liu Y F, Yan J J, Jia F, Li S L, Gao A, Wang T C, et al. PETRv2: A unified framework for 3D perception from multi-camera images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France: IEEE, 2023. 3262–3272
 - 183 Wang S H, Liu Y F, Wang T C, Li Y, Zhang X Y. Exploring object-centric temporal modeling for efficient multi-view 3D object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France: IEEE, 2023. 3621–3631
 - 184 Chen D, Li J, Guizilini V, Ambruş R, Gaidon A. Viewpoint equivariance for multi-view 3D object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE, 2023. 9213–9222
 - 185 Doll S, Schulz R, Schneider L, Benzin V, Enzweiler M, Lensch H P A. SpatialDETR: Robust scalable Transformer-based 3D object detection from multi-view camera images with global cross-sensor attention. In: Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer, 2022. 230–245
 - 186 Xiong K X, Gong S, Ye X Q, Tan X, Wan J, Ding E R, et al. CAPE: Camera view position embedding for multi-view 3D object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE, 2023. 21570–21579
 - 187 Shu C Y, Deng J J, Yu F, Liu Y F. 3DPPE: 3D point positional encoding for Transformer-based multi-camera 3D object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France: IEEE, 2023. 3580–3589
 - 188 Zhang R R, Qiu H, Wang T, Guo Z Y, Cui Z T, Qiao Y, et al. MonoDETR: Depth-guided Transformer for monocular 3D object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France: IEEE, 2023. 9155–9166
 - 189 Tseng C Y, Chen Y R, Lee H Y, Wu T H, Chen W C, Hsu W H. CrossDTR: Cross-view and depth-guided Transformers for 3D object detection. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA). London, UK: IEEE, 2023. 4850–4857
 - 190 Xie Y M, Jiang H Z, Gkioxari G, Straub J. Pixel-aligned recurrent queries for multi-view 3D object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France: IEEE, 2023. 18370–18380
 - 191 Wang Z T, Huang Z H, Fu J H, Wang N Y, Liu S. Object as query: Lifting any 2D object detector to 3D detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France: IEEE, 2023. 3791–3800
 - 192 Jiang X H, Li S L, Liu Y F, Wang S H, Jia F, Wang T C, et al. Far3D: Expanding the horizon for surround-view 3D object detection. In: Proceedings of the 38th AAAI Conference on Artificial Intelligence. Vancouver, Canada: AAAI, 2024. 2561–2569
 - 193 Ji H X Y, Liang P P, Cheng E K. Enhancing 3D object detection with 2D detection-guided query anchors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, 2024. 21178–21187
 - 194 Tang Y J, He H W, Wang Y, Wu J D. Towards efficient multi-modal 3D object detection: Homogeneous sparse fuse network. *Expert Systems With Applications*, 2024, **256**: Article No. 124945
 - 195 Li Y Y, Fan L, Liu Y, Huang Z H, Chen Y T, Wang N Y, et al. Fully sparse fusion for 3D object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024, **46**(11): 7217–7231
 - 196 Wang Z T, Huang Z H, Gao Y L, Wang N Y, Liu S. MV2DFusion: Leveraging modality-specific object semantics for multi-modal 3D detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026, **48**(1): 609–623
 - 197 Li Y H, Li H Y, Huang Z H, Chang H, Wang N Y. SparseFusion: Efficient sparse multi-modal fusion framework for long-range 3D perception. arXiv preprint arXiv: 2403.10036, 2024.
 - 198 Son H, He J, Park S I, Min Y, Zhang Y H, Yoo B. SparseVoxFormer: Sparse voxel-based Transformer for multi-modal 3D object detection. arXiv preprint arXiv: 2503.08092, 2025.
 - 199 Ji H, Ni T, Huang X F, Shi Z, Luo T, Zhan X, et al. RoPETR: Improving temporal camera-only 3D detection by integrating enhanced rotary position embedding. arXiv preprint arXiv: 2504.12643, 2025.
 - 200 Chen X Z, Ma H M, Wan J, Li B, Xia T. Multi-view 3D object detection network for autonomous driving. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA: IEEE, 2017. 1907–1915
 - 201 Ku J, Mozifian M, Lee J, Harakeh A, Waslander S L. Joint 3D proposal generation and object detection from view aggregation. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Madrid, Spain: IEEE, 2018. 1–8
 - 202 Cai Q, Pan Y W, Yao T, Ngo C W, Mei T. ObjectFusion: Multi-modal 3D object detection with object-centric fusion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France: IEEE, 2023. 18067–18076
 - 203 Liang M, Yang B, Chen Y, Hu R, Urtasun R. Multi-task multi-sensor fusion for 3D object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA: IEEE, 2019. 7345–7353
 - 204 Wu X P, Peng L, Yang H H, Xie L, Huang C X, Deng C Q, et al. Sparse fuse dense: Towards high quality 3D detection with depth completion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE, 2022. 5418–5427
 - 205 Pang S, Morris D, Radha H. CLOCs: Camera-LiDAR object candidates fusion for 3D object detection. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Las Vegas, USA: IEEE, 2020. 10386–10393
 - 206 Pang S, Morris D, Radha H. 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 (WACV). Waikoloa, USA: IEEE, 2022. 187–196
 - 207 Bai X Y, Hu Z Y, Zhu X G, Huang Q Q, Chen Y L, Fu H B, et al. Transfusion: Robust LiDAR-camera fusion for 3D object detection with Transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE, 2022. 1090–1099
 - 208 Yan J J, Liu Y F, Sun J J, Jia F, Li S L, Wang T C, et al. Cross modal Transformer: Towards fast and robust 3D object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France: IEEE, 2023. 18268–18278
 - 209 Yang Z Y, Chen J Q, Miao Z W, Li W, Zhu X T, Zhang L. DeepInteraction: 3D object detection via modality interaction. In: Proceedings of the 36th International Conference on Neural

- Information Processing Systems. New Orleans, USA: Curran Associates, 2022. Article No. 145
- 210 Yang Z Y, Song N, Li W, Zhu X T, Zhang L, Torr P H S. DeepInteraction++: Multi-modality interaction for autonomous driving. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025, 47(8): 6749–6763
- 211 Chen X Y, Zhang T Y, Wang Y, Wang Y L, Zhao H. FUTR3D: A unified sensor fusion framework for 3D detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Vancouver, Canada: IEEE, 2023. 172–181
- 212 Zhang H C, Liang L, Zeng P X, Song X, Wang Z. SparseLIF: High-performance sparse LiDAR-camera fusion for 3D object detection. In: Proceedings of the 18th European Conference on Computer Vision. Milan, Italy: Springer, 2025. 109–128
- 213 Xie Y C, Xu C F, Rakotosaona M J, Rim P, Tombari F, Keutzer K, et al. SparseFusion: Fusing multi-modal sparse representations for multi-sensor 3D object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France: IEEE, 2023. 17591–17602



汪凌峰 北京化工大学信息科学与技术学院教授。2013 年获得中国科学院自动化研究所博士学位。主要研究方向为机器学习理论, 计算机视觉。

E-mail: lfwang@mail.buct.edu.cn

(**WANG Ling-Feng** Professor at the College of Information Science

and Technology, Beijing University of Chemical Technology. He received his Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences in 2013. His research interests include machine learning theory and computer vision.)



但啟涛 北京化工大学信息科学与技术学院硕士研究生。2023 年获得北京化工大学学士学位。主要研究方向为自动驾驶 3D 感知。

E-mail: 2023200802@buct.edu.cn

(**DAN Qi-Tao** Master student at the College of Information Science

and Technology, Beijing University of Chemical Technology. He received his bachelor degree from Beijing University of Chemical Technology in 2023. His main research interest is 3D perception for autonomous driving.)



朱耀辉 北京化工大学信息科学与技术学院副教授。2021 年获得中国科学院计算技术研究所博士学位。主要研究方向为多模态感知, 感知规划一体化自动驾驶。

E-mail: yaohui.zhu@buct.edu.cn

(**ZHU Yao-Hui** Associate profess-

or at the College of Information Science and Technology, Beijing University of Chemical Technology. He received his Ph.D. degree from the Institute of Computing Technology, Chinese Academy of Sciences in 2021. His research interests include multi-modal perception and integrated perception and planning for autonomous driving.)



陈思涵 北京化工大学信息科学与技术学院硕士研究生。2022 年获得北京化工大学学士学位。主要研究方向为自动驾驶。

E-mail: 2022200815@buct.edu.cn

(**CHEN Si-Han** Master student at the College of Information Science

and Technology, Beijing University of Chemical Technology. She received her bachelor degree from Beijing University of Chemical Technology in 2022. Her main research interest is autonomous driving.)



张瀚清 北京化工大学信息科学与技术学院硕士研究生。2023 年获得北京化工大学学士学位。主要研究方向为图像生成, 图像质量评估。

E-mail: 2023200794@buct.edu.cn

(**ZHANG Han-Qing** Master student at the College of Information

Science and Technology, Beijing University of Chemical Technology. He received his bachelor degree from Beijing University of Chemical Technology in 2023. His research interests include image generation and image quality assessment.)



王坤峰 北京化工大学信息科学与技术学院教授。2008 年获得中国科学院自动化研究所博士学位。主要研究方向为计算机视觉, 多模态感知和智能无人系统。本文通信作者。

E-mail: wangkf@buct.edu.cn

(**WANG Kun-Feng** Professor at

the College of Information Science and Technology, Beijing University of Chemical Technology. He received his Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences in 2008. His research interests include computer vision, multi-modal perception, and intelligent unmanned systems. Corresponding author of this paper.)