软件学报 ISSN 1000-9825, CODEN RUXUEW
2026,37(5):1919−1935 [doi: 10.13328/j.cnki.jos.007545] [CSTR: 32375.14.jos.007545]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

交通场景多模态双阶反馈的三维目标检测方法*

唐文能[1], 李垚辰[1], 高笙景[1], 高 聪[1], 彭越涵[1], 刘跃虎[2]

[1](西安交通大学 软件学院, 陕西 西安 710049)

[2](人机混合增强智能全国重点实验室 (西安交通大学), 陕西 西安 710049)

通信作者: 李垚辰, E-mail: yaochenli@mail.xjtu.edu.cn

摘 要: 智能驾驶技术的最新进展主要体现在环境感知层面, 其中传感器数据融合对提升系统性能至关重要. 点云数据虽能提供精确三维空间描述, 但存在无序性和稀疏性; 图像数据则分布规则且稠密, 二者融合可弥补单模态检测的不足. 然而, 现有融合算法存在语义信息有限、模态交互不足等问题, 多模态三维目标检测在高精度检测方面仍有提升空间. 针对此问题, 提出一种多传感器融合方法: 利用 RGB 图像深度补全生成伪点云, 与真实点云结合以识别感兴趣区域. 关键改进包括: 采用可变形注意力的多层次特征提取, 自适应扩展感受野至目标区域; 利用二维稀疏卷积对伪点云进行高效特征提取, 发挥其图像域规则分布特性; 提出双阶反馈机制, 在特征级通过多模态交叉注意力解决数据对齐问题, 在决策级采用高效融合策略, 实现多阶段交互训练. 该方法有效解决了伪点云精度受限与计算量增大的矛盾, 显著提升了特征提取效率与检测精度. 在 KITTI 数据集的实验结果表明, 所提方法在三维交通要素检测任务中实现了更优的性能, 充分验证了算法的有效性, 为智能驾驶环境感知中的多模态融合提供了新思路.

关键词: 三维目标检测; 多模态融合; 注意力机制; 点云处理; 交通场景

中图法分类号: TP181

中文引用格式: 唐文能, 李垚辰, 高笙景, 高聪, 彭越涵, 刘跃虎. 交通场景多模态双阶反馈的三维目标检测方法. 软件学报, 2026, 37(5): 1919–1935. http://www.jos.org.cn/1000-9825/7545.htm

英文引用格式: Tang WN, Li YC, Gao SJ, Gao C, Peng YH, Liu YH. Multi-modal 3D Object Detection Method for Traffic Scenarios Based on Two-stage Feedback. Ruan Jian Xue Bao/Journal of Software, 2026, 37(5): 1919–1935 (in Chinese). http://www.jos.org.cn/1000-9825/7545.htm

Multi-modal 3D Object Detection Method for Traffic Scenarios Based on Two-stage Feedback

TANG Wen-Neng[1], LI Yao-Chen[1], GAO Sheng-Jing[1], GAO Cong[1], PENG Yue-Han[1], LIU Yue-Hu[2]

[1](School of Software Engineering, Xi'an Jiaotong University, Xi'an 710049, China)

[2](State Key Laboratory of Human-machine Hybrid Augmented Intelligence (Xi'an Jiaotong University), Xi'an 710049, China)

Abstract: The latest advancements in intelligent driving technology are primarily reflected in the environmental perception layer, where sensor data fusion is critical for enhancing system performance. Although point cloud data provides accurate 3D spatial descriptions, it suffers from unorderedness and sparsity. Image data, with its regular and dense distribution, can compensate for the limitations of single-modality detection when fused with point clouds. However, existing fusion algorithms face challenges such as limited semantic information and insufficient modal interaction, leaving room for improvement in high-precision multi-modal 3D object detection. To address this issue, this study proposes an innovative multi-sensor fusion method: generating pseudo-point clouds via depth completion from RGB images and combining them with real point clouds to identify regions of interest. It introduces three key improvements: (1) deformable attention-based

* 基金项目: 国家自然科学基金 (62473307)
本文由"多媒体智能理解与生成"专题特约编辑孙立峰教授、闵巍庆副研究员、马占宇教授、蒋树强研究员、彭宇新教授、田丰研究员、黄庆明教授推荐.
收稿时间: 2025-05-26; 修改时间: 2025-07-11; 采用时间: 2025-09-05; jos 在线出版时间: 2025-09-23
CNKI 网络首发时间: 2026-01-02

multi-layer feature extraction that adaptively expands the receptive field to target regions; (2) 2D sparse convolution for efficient pseudo-point cloud feature extraction leveraging their regular distribution in the image domain; and (3) a two-stage feedback mechanism employing multi-modal cross-attention at the feature level to solve data alignment issues and an efficient fusion strategy at the decision level for interactive training across different stages. These innovations effectively resolve the trade-off between pseudo-point cloud accuracy and computational load while significantly enhancing both feature extraction efficiency and detection accuracy. Experimental results on the KITTI dataset demonstrate the superior performance of the proposed method in 3D traffic object detection, validating its effectiveness and offering a new approach for multi-modal fusion in autonomous driving environmental perception.

Key words: 3D object detection; multi-modal fusion; attention mechanism; point cloud processing; traffic scenario

1 引 言

在智能交通系统[1]飞速发展的当下, 无人车作为其核心组成部分, 正逐渐从概念走向现实应用. 无人车要在复杂的交通场景中安全、高效地运行, 精准的环境感知是其关键前提[2]. 环境感知旨在实时、准确地获取车辆周边的各类信息, 包括目标物体的位置、形态、运动状态等, 从而为后续的路径规划、决策控制等模块提供可靠依据.

近年来, 随着智能驾驶技术的不断演进, 使用多传感器进行环境感知的场景日益增多[3]. 多传感器融合能够综合利用不同传感器的优势, 如激光雷达提供的高精度三维空间信息、摄像头捕获的丰富纹理色彩信息等, 从而为无人车构建更全面、准确的环境模型. 然而, 这也使得有效的传感器融合成为该领域的关键研究挑战. 若融合过程处理不当, 不同传感器数据之间会产生干扰, 进而对感知精度产生负面影响, 严重威胁无人车的行驶安全. 因此, 设计鲁棒的融合算法, 以实现最优检测性能, 对于推动智能交通和无人车技术的发展至关重要.

当前, 主流的点云三维目标检测方法在一些常用的基准数据集上已经能够实现较为理想的召回率, 表明这些方法已经具备了一定的粗略定位目标的能力. 但是, 由于点云数据自身稀疏性、不规则性等特性以及现有方法的局限性, 现有技术在实现更高精度的检测方面仍面临显著挑战. 为了突破这一局限, 研究者们开始探索引入其他传感器信息, 如摄像头所捕获的图像信息, 以此对点云数据进行补充, 从而有望获得更高精度的检测结果.

近期, 诸多研究者尝试利用图像深度补全[4]等任务对图像数据进行处理, 并将图像数据生成的伪点云与真实点云数据进行融合, 以此生成高质量的三维目标检测结果. 此类方法的优势在于可以沿用处理三维点云数据的方法来处理伪点云数据, 且与以往基于投影的方法不同, 其能够充分挖掘图像数据中稠密的上下文信息. 但不可忽视的是, 该方法也存在一些问题: 一方面, 相比于真实点云数据, 伪点云数据更加稠密, 数量更多, 这无疑大幅增加了处理所需的计算量, 对硬件设备的计算能力提出了更高要求; 另一方面, 由于深度补全过程中不可避免地存在误差, 导致伪点云难以精确地反映场景的三维真实分布, 从而影响检测的准确性.

基于上述观察与分析, 我们提出一种结合伪点云与体素特征的新型两阶段迭代融合方法. 该方法以点云与 RGB 图像为输入, 首先从真实点云生成粗检测结果, 随后对 RGB 图像进行深度补全以生成伪点云, 并利用感兴趣区域 (region of interest, RoI) 对其进行裁剪, 进而提取并融合不同传感器数据的特征. 值得一提的是, 除特征融合外, 我们设计的带多层次反馈的多模态融合算法可以进一步优化多模态结果, 有效提升检测的精度与稳定性.

本文主要贡献总结如下.

(1) 针对真实点云, 提出基于可变形注意力的多层次 RoI 特征提取方法, 将点特征扩展至 RoI 内部之外并自适应融入周围区域信息.

(2) 针对图像数据生成的伪点云特征提取, 提出基于二维稀疏卷积的高效 RoI 特征提取方法, 利用伪点云的二维分布规律性提升处理效率.

(3) 针对多模态融合, 提出两阶段迭代融合方法, 分别对特征和结果进行融合. 在特征融合阶段, 引入基于交叉注意力的方法解决数据错位导致的融合歧义问题; 同时提出一种简单高效的多模态检测结果融合方法以提升模型性能.

(4) 所提方法在车辆检测任务中展现出卓越性能, 在 KITTI[5]测试集上中等难度平均精度超越众多先进算法,

并在推理时间方面达到了最先进 (state-of-the-art, SOTA) 水平.

2 相关工作

2.1 基于点云的三维目标检测方法

基于原始点云的处理方法通常采用 PointNet 系列[6,7]进行逐点特征提取. 在此基础上, VoteNet 方法[8]揭示了表面点云分布导致的几何中心表征缺失问题, 提出基于偏移矢量场估计的特征重定位机制, 通过空间坐标变换将表面特征向潜在中心区域进行概率密度迁移. 两阶段方法如 PointRCNN[9]先进行前景分割, 再进行边界框回归. CT3D[10]则进一步通过通道注意力机制提升检测性能. 3DSSD 方法[11]发现特征传播解码层的计算冗余, 构建基于集合抽象层的轻量化预测头, 通过消除层级式特征融合机制, 在确保测量准确性的前提下提升计算效率. 针对深层网络采样性能退化问题, IASSD 方法[12]提出双重感知采样策略, 结合目标空间位置与类别先验信息改进采样鲁棒性. SASA 方法[13]则通过预测采样点置信度权重, 构建语义引导的加权采样算法, 增强关键区域特征保留能力.

在基于体素的方法中, 早期研究聚焦于空间离散化策略优化. 其中, PointPillars 方法[14]提出柱状单元空间离散化策略, 通过投影将三维点云压缩至伪二维表征空间. 并通过点云特征提取模块聚合柱状单元内的几何信息, 构建适用于二维卷积处理的特征矩阵. 针对特征矩阵中存在的稀疏分布问题, PillarNet[15]与 PillarNeXt[16]方法采用稀疏卷积替代常规卷积进行运算, 有效降低计算开销. VoxelNet 方法[17]采用均匀体素划分方案, 配合三维卷积神经网络实现结构化特征学习, 其分层特征聚合机制为后续研究奠定重要基础. 为提升计算效率, SECOND 方法[18]使用稀疏卷积, 仅对非空网格执行特征计算. 在检测范式革新方面, CenterNet[19]与 AFDet[20]方法提出去锚框预测, 消除传统方法中预定义锚框的几何约束. SASSD 方法[21]引入点云语义分割辅助任务, 通过多任务联合训练强化网络对目标表面几何结构的感知能力. CIASSD 方法[22]提出任务解耦补偿机制, 设计独立的交并比校准分支, 实现分类置信度与定位精度的动态对齐, 有效缓解因任务目标冲突导致的性能衰减问题. PVT-SSD 方法[23]则尝试融合原始点处理与体素化策略, 构建混合特征表征体系以提升检测精度. 在两阶段网格处理方法中, Pillar R-CNN[24] 和 Voxel R-CNN[25]通过添加感兴趣区域检测头和体素池化方法进行特征提取. Casa[26]和 TED[27]则引入级联预测和旋转不变特征提取以增强检测性能.

基于体素的方法通过离散化点云构建结构化特征表示, 其端到端架构在保留局部几何位移特征的同时, 有效平衡了信息完整性, 在运算效率与检测精度方面展现出较强的竞争力. 基于点云的方法虽可实现端到端训练并取得可观的检测精度, 但其核心依赖的最远点采样机制因本质为序列化采样流程, 存在并行计算受限问题.

受 Transformer 在视觉领域成功应用[28]的启发, 其泛化迁移能力引发三维感知领域的研究变革. DSVT[29]方法首次构建面向稀疏体素特征的自适应注意力学习框架, 其设计动态窗口自注意机制, 通过可变形窗口生成器动态划分非空体素区域, 减小计算复杂度. 针对基于 DETR[30]方法中三维检测器因对象查询监督模糊导致的局部假阳性问题, ConQueR[31]方法提出查询对比监督机制, 通过抑制非匹配查询置信度并强化最优匹配特征表达, 提升目标检测鲁棒性. OcTr[32]方法提出一种基于八叉树层级的注意力机制, 首层通过全局自注意力筛选关键特征标记构建动态八叉树结构, 采用递归式特征广播实现由粗到精的上下文建模, 在保证计算效率的同时增强全局表征能力.

2.2 基于多传感器融合的三维目标检测方法

基于多传感器融合的方法将来自不同传感器的数据进行融合, 以生成综合信息的表示. 当前主流方法多采用激光雷达与视觉传感器的跨模态融合策略. 其中, 激光雷达提供精确的三维几何表征, 构建具有绝对尺度的空间结构. 而视觉传感器则捕获高分辨率的纹理语义信息, 形成细粒度的场景上下文理解. 基于多模态融合的方法通过构建鲁棒的跨模态对齐机制, 可有效克服单模态存在的几何歧义与语义缺失的双重局限性, 进而构建鲁棒性更强的环境感知系统.

点级融合方法通过跨模态特征对齐实现细粒度特征增强, 但面临计算复杂度较高的问题. EPNet[33]方法设计无监督跨模态融合单元, 通过逐点级联图像语义特征优化点云表征, 无需依赖图像标注信息. EPNet++[34] 将点云数据投影到图像特征图上进行融合, 并利用融合后的点云特征对图像特征进行重新加权. PointPainting[35]方法通过跨模

态投影机制实现图像语义分割结果与点云几何特征的逐点关联. PointAugmenting[36]方法利用预训练的二维检测模型提取的点状特征来修饰点云, 之后进行三维物体检测. MVP[37]方法提出虚拟点云生成策略, 将检测结果转换为三维空间伪点云, 实现稀疏点云的密度补偿. 此类方法虽提升特征表达精度, 但由于逐点运算特性导致实时性受限.

特征级融合方法通过特征空间映射实现高效信息融合. MV3D[38]方法在俯视图特征空间进行跨模态特征拼接, 在计算效率与信息完整间取得平衡. AVOD[39]方法借鉴特征金字塔方法, 提取出的特征更加高效. 3D-CVF[40]使用自校准模块将图像特征转换为鸟瞰图视角, 再通过自适应门控融合模块进行融合. CLOCs[41]使用二维卷积处理稀疏矩阵以学习剔除错误预测. Fast-CLOCs[42]则引入 3D-Q-2D 模块, 使方法无需运行两个独立检测器即可工作. GOOD[43] 提出一种基于优化的晚期融合方法, 对不同维度的结果匹配进行分类和优化. F-PointNet[44]、PointFusion[45]和 F-ConvNet[46]使用二维检测器生成检测框, 将其投影到平截头体中进行点云裁剪. VPFNet[47]将感兴趣区域划分为网格并投影到立体图像特征上, 再将生成的图像特征与体素特征融合. SFD[48]方法提出双流编码架构, 通过区域建议网络生成高置信度三维候选框, 继而分别在点云体素空间与图像像素空间执行双流特征编码, 实现特征空间的几何一致性对齐. TransFusion[49]方法引入交叉注意力机制, 以点云候选框为查询实现图像特征动态聚合. DeepFusion[50]方法构建双向特征交互通道, 通过激光雷达与图像双向投影建立空间对应关系. 通过引入特征金字塔、自适应加权融合等技术, 持续优化跨模态特征对齐精度.

3 基于双阶反馈的多模态三维目标检测模型

如图 1 所示为我们提出方法的流程架构. 现有基于体素的三维目标检测方法首先被用于生成多传感器信息融合的感兴趣区域 (RoI). 这些感兴趣区域通过提取完整的点云特征并裁剪伪点云, 实现了处理负载与生成误差的双重降低. 此外, 本文引入了一种两阶段迭代融合方法, 以自适应地集成多模态特征与结果, 进而提升检测性能. 主干网络、特征提取以及多模态融合的具体细节将在后续章节中展开论述.

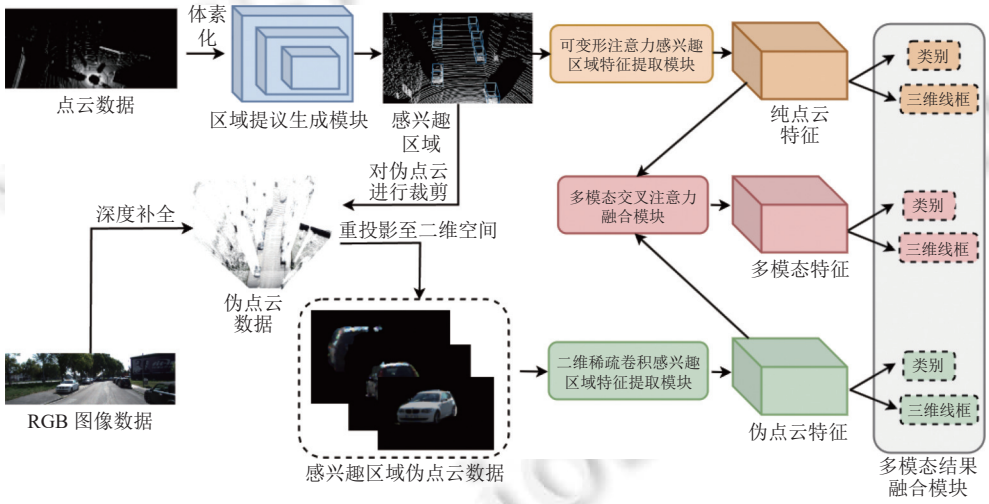

图 1 多模态双阶反馈的三维目标检测算法整体流程图

3.1 数据处理流程

本方法的输入数据包括点云、对应视角的 RGB 图像、传感器间的校准参数以及相机的内参和外参, 以实现高效的传感器融合.

在点云处理方面, 我们采用 SECOND 网络作为骨干网络, 该网络首先对点云进行体素化, 然后通过三维稀疏卷积和二维卷积生成区域建议 (region proposal). 对于生成的大量区域建议, 我们使用与 Voxel R-CNN 相同的参数设置, 通过非极大值抑制 (non-maximum suppression, NMS) 过滤冗余建议, 筛选后的建议称为感兴趣区域 (RoI), 作为第 2 阶段的输入.

对于图像数据, 我们采用 TWISE 方法[51]对 RGB 图像进行深度补全, 并将其投影到激光雷达坐标系. 该方法为图像数据提供了与点云可比的深度信息, 便于特征提取与融合. 此外, 通过将图像数据保持在三维空间中, 相比二维投影能够更好地利用密集的上下文信息.

在上述步骤的基础上, 我们利用点云生成的 RoI 对伪点云 (通过 RGB 图像深度补全技术生成的模拟点云数据) 进行裁剪, 确保后续阶段仅处理裁剪后的伪点云区域. 随后, 对各模态的 RoI 进行特征提取与融合, 最终利用多模态特征对 RoI 进行进一步优化.

3.2 点云特征提取

目前, 大多数现有方法在特征提取过程中往往仅聚焦于固定区域, 这一局限性使得区域提议网络 (region proposal network, RPN) 生成的粗略预测难以实现有效的检测调整. 具体而言, 由于模型未能充分考虑目标周围区域的上下文特征, 导致初始提议框的空间依赖性利用不足, 难以在后续处理中对目标的位置和类别进行精细化修正. 针对这一关键问题, 我们提出了一种基于可变形注意力的特征提取方法. 该方法突破了传统固定区域的约束, 赋予模型动态聚焦邻近物体特征的能力, 并通过自适应权重机制对不同位置的特征进行差异化聚合. 通过这种方式, 模型能够捕捉目标区域与周围环境之间多尺度的空间关联, 有效提取包含丰富细节和结构信息的精细化上下文特征. 这些经过增强的空间上下文特征为后续感兴趣区域 (RoI) 的调整提供了更具判别性的依据, 显著提升了边界框回归和分类的准确性. 图 2 展示了两种方法的对比原理.

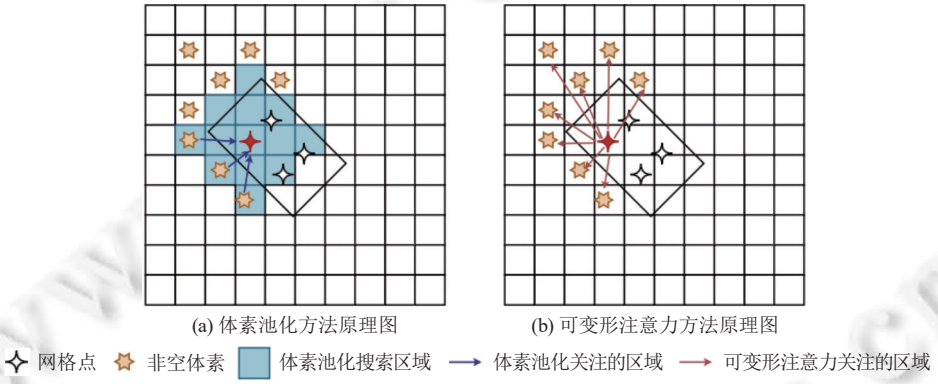

图 2 体素池化方法与可变形注意力方法原理对比

为了充分捕捉感兴趣区域 (RoI) 的环境上下文信息, 本方法将边界框按比例扩大, 以纳入更多周围环境信息. 随后, 系统在该扩展后的边界框内生成规则网格点, 与直接进行 RoI 特征聚合的方法相比, 此方式能够保留更多结构细节. 具体而言, 本方法将 RoI 在三维空间中均匀划分为 $G\times G\times G$ 大小的网格 (本实验中 G 设为 6), 并以网格质心 P_i 处的体素特征作为可变形注意力机制的查询向量. 通过可变形注意力机制, 模型将周围环境信息聚合为特征向量, 随后将这些特征向量组合, 生成适用于 RoI 的点云表示. 图 3 展示了可变形注意力机制的具体原理.

网格点的初始特征 f_i^l 被设置为其对应的体素特征. 每个网格点配备 K 个采样点, 采样位置初始时分布在网格点周围. 具体初始化方法如图 4 所示.

本方法利用位置嵌入和网格点特征来计算这些采样点的偏移量和注意力权重, 从而在特征聚合过程中实现对关键采样位置和特征的自适应聚焦. 采样点的偏移量通过两个多层感知机 (multi-layer perceptron, MLP) 层进行计算, 具体如下所示:

$$\Delta P_{i_k}^l = MLP\left(MLP\left(ReLU\left(MLP\left(f_i^l + PE\left(C_i^l\right)\right)\right)\right)\right) \tag{1}$$

其中, $\Delta P_{i_k}^l$ 表示采样点相对于其初始位置的三维坐标偏移量, $PE(C_i^l)$ 表示网格点坐标的位置嵌入. 此步骤的输出维度为 3, 而输入特征维度与特征图层级相对应. 基于采样点偏移量及其初始坐标, 可确定最终的采样点坐标, 进而检索对应的体素特征图特征.

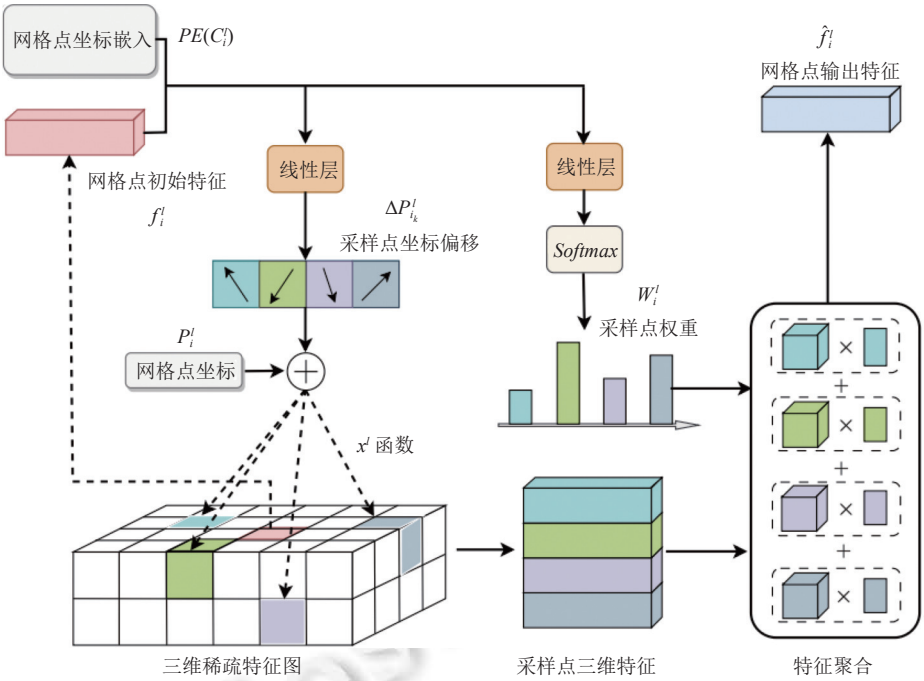

图 3 可变形注意力方法提取网格点特征具体原理

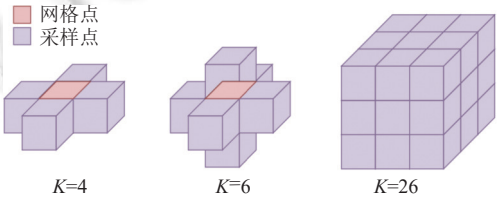

图 4 采样点初始化方法

对于采样点的注意力权重, 本方法利用一个多层感知机 (MLP) 层生成各采样点的权重, 并最终通过 *Softmax* 函数对生成的权重进行归一化处理. 具体方法定义如下:

$$W_i^l = Softmax\left(MLP\left(ReLU\left(MLP\left(f_i^l\right)\right)\right)\right) \tag{2}$$

其中, W_i^l 为特征图 F^l 上网格点 P_i 对应采样点的权重矩阵, 其维度大小为 K 且权重和为 1. 因此, 基于采样点的特征与注意力权重, 可通过计算采样点特征向量的加权和得到网格点的特征, 具体计算方法定义为:

$$\hat{f}_i^l = Concat_{l=1}^{L}\left(\sum_{k=1}^{K}\left(W_{i_k}^l \cdot x^l\left(P_i^l + \Delta P_{i_k}^l\right)\right)\right) \tag{3}$$

其中, $\hat{f}_i^l$ 为最终得到的网格点特征, $Concat_{l=1}^{L}$ 表示对各层级网格点获取的特征向量进行拼接操作, $W_{i_k}^l$ 为特征图 L 上感兴趣区域 (RoI) 内网格点 i 的采样点 k 计算得到的注意力权重, x^l 为利用采样点在特征图 F^l 上的坐标获取对应特征的方法.

可变形注意力方法通过捕捉 RoI 内各层级的网格点特征, 自适应调整采样位置与权重, 从而降低对初始预测的依赖. 然而, 由于点云数据的稀疏性, 这些特征可能缺乏足够细节以支持高质量预测. 针对这一问题, 本方法将 RGB 图像信息集成到 RoI 表示中, 以实现更全面的多模态特征提取.

3.3 图像特征提取

针对密集且精度有限的伪点云, 我们借助真实点云的粗定位能力进行处理. 首先基于真实点云数据生成 RoI,

并对伪点云进行裁剪, 仅保留目标周围区域. 伪点云裁剪的效果图如图 5(a) 所示. 该方法仅处理裁剪后的数据, 显著减少了点云规模, 同时最大程度降低了深度补全误差对检测结果的干扰. 随后, 研究重点转向对 RoI 内伪点云的特征提取.

(a) 伪点云裁剪效果图　　(b) 伪点云投影至二维平面效果图

图 5　伪点云裁剪效果图及投影至二维平面效果图

传统基于点的处理方法效率低下, 而基于体素的方法在处理复杂数据时存在困难, 均不适用于伪点云 RoI 的特征提取. 二维稀疏卷积通过将伪点云从三维空间重投影至二维图像空间, 利用卷积网络高效提取特征. 它避免了对空值位置的无效计算, 降低了计算复杂度, 加速了处理速度, 同时保留了三维特征信息和图像域的有序性. 为此, 本文提出将伪点云数据从三维空间重投影至二维图像空间, 并通过高效的二维稀疏卷积实现特征提取. 伪点云投影至二维平面的效果图如图 5(b) 所示. 在图 6 中, 展示了基于二维稀疏卷积进行特征提取的流程, 可以看出, 该方法在图像域中保留了三维特征信息, 同时充分利用了图像处理的高效性.

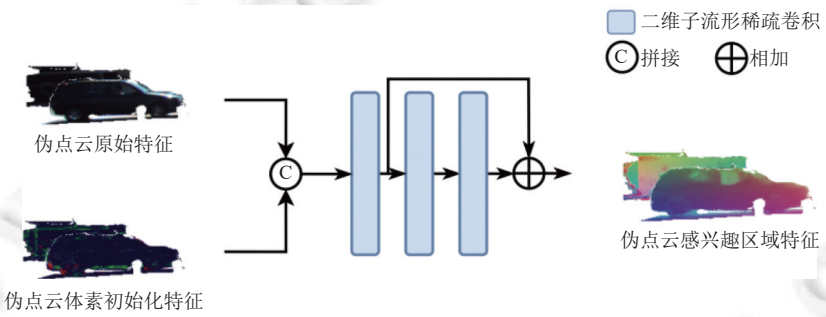

图 6　基于二维稀疏卷积的伪点云特征提取图

伪点云数据通过二维图像的深度补全生成, 其不仅包含三维点云的空间信息, 还包含伪点云在二维图像上的规则坐标以及颜色等信息. 因此, 伪点云的初始特征可描述为 (x,y,z,r,g,b,u,v), 其中 (x,y,z) 为伪点云在三维空间中的坐标, (r,g,b) 为伪点云的颜色信息, (u,v) 为伪点云对应像素在二维图像上的坐标.

为提升伪点云的表征能力并降低计算开销, 我们提出一种多模态伪点云特征初始化技术. 该方法利用插值体素特征对伪点云特征进行初始化, 具体通过计算伪点云的体素坐标, 快速识别其周围体素, 并基于邻域空实性和距离权重对特征进行聚合. 特征初始化过程在各层级体素特征上分别执行, 具体定义如下:

$$w_k^l = \frac{d_k^l * E_k^l}{\sum_{k=1}^{K}\left(d_k^l * E_k^l\right)} \tag{4}$$

$$f_i = Concat_{l=1}^{L}\left(\sum_{k=1}^{K}\left(w_k^l * f_k^l\right)\right) \tag{5}$$

其中, w_k^l 表示第 l 层中伪点云体素的第 k 个邻域体素权重, d_k^l 为该邻域体素到伪点云体素的距离, E_k^l 表示邻域体素是否为空 (空为 0, 否则为 1), *表示加权乘积操作, f_i 为初始化后的网格特征. 最终的伪点云特征表示为 (f_i,x,y,z,r,g,b,u,v).

利用初始化后的伪点云特征, 我们将数据投影至二维空间, 借助卷积网络实现高效特征提取, 充分利用二维空间的有序性和上下文信息. 通过应用二维稀疏卷积, 避免了对空值位置的无效计算, 从而降低了计算复杂度并加速处理速度. 本方法在推理时间方面达到了当前最优水平 (如图 7 所示). 我们采用子流形稀疏卷积来保留伪点云的形态特征, 并增强数据点之间的交互能力.

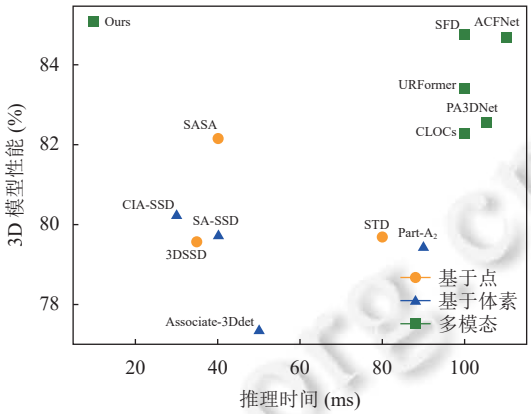

图 7 模型方法总体对比图

3.4 多模态双阶互馈融合

传统融合方法仅单一关注不同模态的特征或结果. 为此, 我们提出一种两阶段迭代融合方法, 该方法同时集成了多模态特征融合与多模态结果融合.

在特征融合方面, 当前不同传感器的特征以不同格式组织. 真实点云中的 RoI 特征表示为尺寸为 (B,N,G,G,G,C) 的张量, 其中 B 为批量大小, N 为 RoI 数量, G 为每个 RoI 的网格点数, C 为网格点特征长度. 对于伪点云数据, 其特征按点组织为 $(M,C+1)$ 的形式, 其中 M 为 RoI 内的伪点数量 (随批量变化), C 为特征维度, 额外的“1”表示场景编号以方便后期处理.

我们将不同模态的特征统一为网格点特征, 以促进多模态融合. 对于伪点云特征, 通过将 RoI 划分为均匀的 $(G\times G\times G)$ 网格, 并对每个网格内的伪点云特征应用最大池化操作, 将逐点数据转换为网格点特征.

为便于区分, 真实点云的网格点特征表示为 f_v, 伪点云的特征表示为 f_p. 为稳健融合不同传感器的特征, 我们采用跨模态融合方法. 在该方法中, 各模态的特征在注意力计算过程中作为另一模态的查询向量, 使特征能够自适应地聚合来自另一模态的相关信息. 随后将得到的特征进行拼接, 形成最终的融合特征 f_f, 该原理如图 8 所示. 与直接拼接相比, 我们的方法有效减少了由校准误差导致的融合模糊问题.

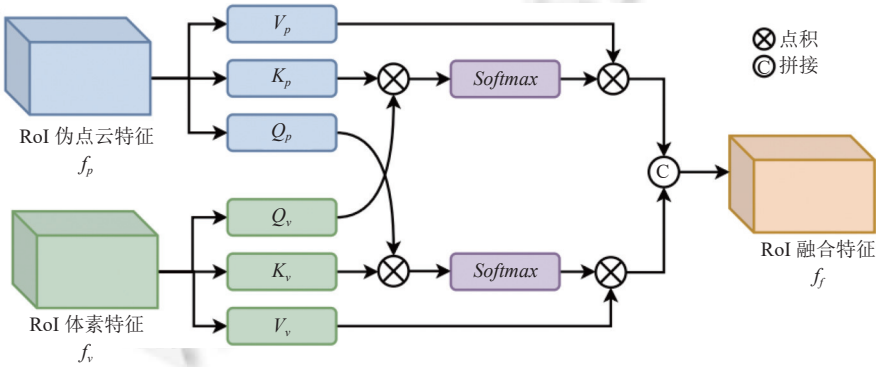

图 8 多模态交叉注意力融合模块原理图

除特征融合外, 受 CaSA 方法中的后融合策略与 Box Voting 算法启发, 我们提出一种高效轻量化的多模态后融合算法——多模态 Box Voting 算法. 该方法通过融合不同模态特征的预测结果生成最终检测结果, 其中各模态检测结果的融合采用简单的均值计算方法, 具体定义如下:

$$box^i = \frac{box_p^i + box_v^i + box_f^i}{3} \tag{6}$$

其中, box_p^i 表示第 1 个检测框的伪点云特征预测结果, 包含目标置信度及边界框参数等数据; box_v^i 表示体素特征的预测结果; box_f^i 表示融合特征的预测结果.

4 实验结果与分析

4.1 实验数据集

(1) KITTI 数据集

KITTI 数据集是智能驾驶领域最常用的数据集, 其提供了包括目标检测、目标跟踪、语义分割、深度补全、深度估计等方面的任务的标签以及评估方案, 可应用于各个与智能驾驶相关的视觉任务. KITTI 数据集中包含了多视角相机采集的彩色图像和灰度图像, 以及激光雷达采集的点云数据和各个传感器之间的标定参数, 图 9 展示了 KITTI 数据采集车辆的坐标系以及各个传感器间的位置关系. 数据集的采集场景主要包括城市道路、高速场景以及乡村道路. 数据以采集时间的形式进行组织, 共包含 7 481 帧训练数据和 7 581 帧测试数据. 训练数据包含标签, 测试数据不包含标签, 需将预测结果提交至官方服务器才可获得相关指标结果.

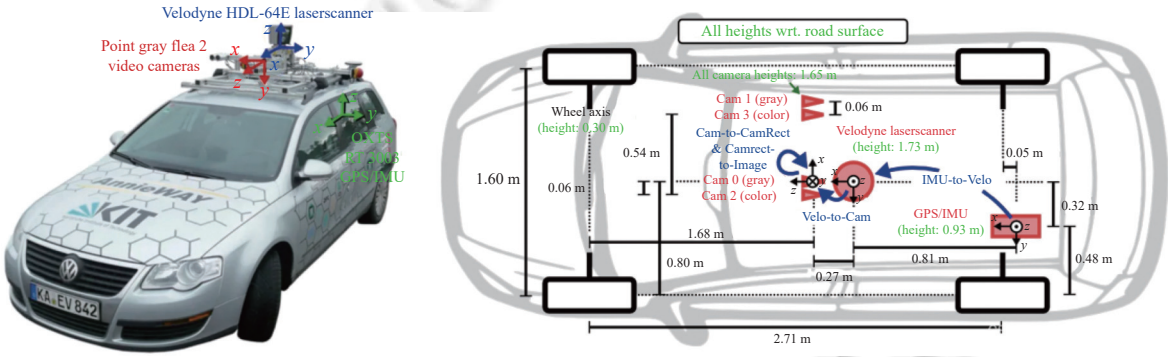

图 9 KITTI 数据集传感器坐标以及相对位置关系

在目标检测任务方面, KITTI 提供了二维图像数据的边界框标注和三维点云数据的边界框标注. 其中, 二维边界框包含目标在二维图像中的中心点位置以及尺寸等信息, 三维边界框除了包含目标在三维点云中的位置以及尺寸之外, 还包括目标的航向角. 此外, KITTI 数据集还对目标是否被截断以及是否被遮挡等进行标注. 在评估过程中, 根据目标的尺寸, 截断程度和遮挡程度, 将目标分为简单、中等、困难这 3 种难度, 并分别计算不同难度下的平均精度. 最终, 官方网站会以中等难度下的检测精度进行排序展示. 在三维检测任务中, 除了对三维边界框的检测精度进行评估之外, 还会对鸟瞰图视角下的检测框和航向角的回归精度进行评估.

(2) nuScenes 数据集

nuScenes 数据集是一个面向自动驾驶领域的大规模的公开数据集. 该数据集的场景来自新加坡和波士顿这两个车辆众多、交通环境复杂的城市, 共包含 1 000 个交通场景, 每个场景时长为 20 s. 这些场景经过精心挑选, 涵盖了多样化的驾驶行为、复杂的交通状况以及突发情况. nuScenes 数据集的丰富性和复杂性为研究者提供了开发新方法的机会, 以应对城市环境中每个场景可能包含数十个物体的挑战, 从而实现安全驾驶.

nuScenes 数据集总计包含约 15 h 的驾驶数据, 驾驶路线经过精心设计, 以捕捉具有挑战性的场景. 为了平衡

目标类别的分布, 数据集中特别增加了包含稀有目标 (如自行车) 的场景. 基于这些标准, 研究团队手动筛选了 1 000 个时长为 20 s 的场景, 并对其进行了精细的人工标注.

在传感器配置方面, nuScenes 数据集配备了 6 个摄像头 (camera), 分别位于车辆的前方 (front)、右前方 (front right)、左前方 (front left)、后方 (back)、右后方 (back right) 和左后方 (back left). 此外, 车顶 (top) 安装了一个激光雷达 (LiDAR) 以及 5 个毫米波雷达, 分别位于前方 (front)、右前方 (front right)、左前方 (front left)、右后方 (back right) 和左后方 (back left). 为了确保多传感器数据集的高质量, 研究团队还对每个传感器的外参 (extrinsics) 和内参 (intrinsics) 进行了精确校准. 这种严格的校准流程为融合多传感器的数据提供了可靠的基础.

4.2 实验评价指标

在三维目标检测任务中, KITTI 数据集对于三维边界框的评价指标为平均精度 (average precision, AP), 该指标是将模型在不同的召回率阈值下的平均精度作为最终的衡量指标. 此外, 在 KITTI 数据集中, 不同类别的目标 IoU 阈值设置也不同, 对于体积较大的目标, 例如汽车等, 其 IoU 阈值被设置为 0.7; 对于体积较小的目标, 例如行人等, 其 IoU 阈值被设置为 0.5. 在平均精度的计算方面, KITTI 选用 40 个不同的召回率阈值计算平均精度. 除了三维检测精度之外, KITTI 数据集还提供了将三维检测框投影到鸟瞰图视角下的检测精度计算以及航向角回归的精度等.

4.3 实现细节

在实验中, 我们将学习率设定为 0.001, 批量大小设定为 6, 且在单个 RTX 3090 GPU 上开展 40 个训练周期的训练工作, 最终选取召回率最高的检测结果用于测试集的提交. 非极大值抑制的置信度阈值 (该阈值对模型性能有着显著影响) 被设定为 0.5. 而其他所有参数均维持 OpenPCDet[52]中的默认设置不变.

4.4 对比实验结果

在本节中, 我们将所提方法与现有的先进多模态三维目标检测算法进行比较, 以验证其有效性. 我们在公开的 KITTI 数据集和 nuScenes 数据集上验证所提方法, 彰显出其有效性.

如表 1 所示, 所提方法在 KITTI 测试集上的性能优于所有现有方法. 其中, Graph-VoI 是 Graph R-CNN 通用框架与 SECOND (Voxel-based RPN) 及跨模态融合模块结合的多模态变体. 与基准方法 SFD 相比, 我们的方法显著提高了中等难度样本和高难度样本的检测精度. 然而, 对于简单样本, 检测精度略有下降. 这可能是因为简单样本的大量伪点云导致最大池化过程中丢失特征更多, 从而影响了检测性能.

表 1 本方法与其他代表性方法在 KITTI 测试集上的对比结果 (%)

方法	汽车 (IoU=0.7)			mAP
	简单	中等	困难	
ContFuse[53]	83.68	68.78	61.67	71.38
MV3D[38]	74.97	63.63	54.00	64.20
AVOD[39]	83.07	71.76	65.73	73.52
F-PointNet[44]	82.19	69.79	60.59	70.86
F-ConvNet[46]	85.88	76.51	68.08	76.82
MMF[54]	88.40	77.43	70.22	78.68
3D-CVF[40]	89.20	80.05	73.11	80.79
EPNet[33]	89.81	79.28	74.59	81.23
CLOCs[41]	88.94	80.67	77.15	82.25
MSF-MC[55]	89.63	80.06	75.83	81.84
Focals Conv[56]	90.55	82.28	77.59	83.47
VPFNet[47]	91.02	83.21	78.20	84.14
Graph-VoI[57]	91.89	83.27	77.78	84.31
EPFNet++[34]	91.37	81.96	76.71	83.35
MLF-DET[58]	91.18	82.89	77.89	83.99

表 1 本方法与其他代表性方法在 KITTI 测试集上的对比结果 (%)(续)

方法	汽车 (IoU=0.7)			mAP
	简单	中等	困难	
PA3DNet[59]	90.49	82.57	77.88	83.64
ACF-Net[60]	90.80	84.67	80.14	85.20
URFormer[61]	89.64	83.40	78.62	83.89
SFD (基准)[48]	**91.73**	84.76	77.92	84.80
本文方法	90.98	**85.07**	**80.16**	**85.40**
提升	−0.75	+0.31	+2.24	+0.6

此外, 本研究还进一步在其余交通要素检测方面进行了探究, 表 2 展示了本文方法与 SFD 方法的检测效果对比, 从表中可以看出, 本文方法的检测精度在各个类别的检测精度上均有提升, 特别是对于遮挡严重的困难种类的目标, 提升幅度更大, 这也一定程度上表明本研究所提方法的有效性.

表 2 本方法在 KITTI 数据集上骑行者和行人的检测效果 (%)

方法	骑行者			行人		
	简单	中等	困难	简单	中等	困难
SFD	89.56	72.83	63.29	72.41	65.39	58.39
本文方法	90.35	73.67	65.84	72.68	66.87	60.19
提升	+0.79	+0.84	+1.55	+0.27	+1.48	+1.80

为进一步验证本文方法的先进性和鲁棒性, 本研究基于 nuScenes 数据集进行了对比实验, 其结果如表 3 所示 (↑表示指标越高效果越好, ↓表示指标越低效果越好). 从表 3 可以看出, 本文方法在 nuScenes 数据集上表现十分出色. 尤其在平均精度 (mAP) 和 nuScenes 检测分数 (NDS) 这两项关键指标上, 本文方法的性能尤为突出. 这表明我们的模型在处理 nuScenes 数据集的复杂场景和多样目标时, 具备较高的精度和鲁棒性. 这些结果进一步证明, 本文方法不仅在 KITTI 数据集上表现优异, 在更具挑战性的 nuScenes 数据集上也能保持高水平性能, 充分展现了其良好的泛化能力和实际应用潜力.

表 3 本文方法与现有方法在 nuScenes 验证集上的结果对比

方法	mAP↑ (%)	NDS↑ (%)	mATE↓	mASE↓	mAOE↓	mAVE↓	mAAE↓
BEVDet	42.2	48.2	0.529	0.236	0.396	0.979	0.152
BEVFormer	44.5	53.5	0.582	0.256	0.375	0.378	0.129
Far3D	51.0	59.4	0.551	0.258	0.372	0.238	0.195
VoxelNeXt	60.5	66.7	0.301	0.252	0.405	0.216	0.185
CenterPoint	60.3	67.3	0.262	0.239	0.361	0.288	0.136
TransFusion	64.5	69.6	0.269	0.249	0.292	0.266	0.189
SparseFusion	70.4	72.8	0.273	0.258	0.342	0.277	0.138
CMT	70.3	72.9	0.294	0.260	0.323	0.271	0.131
FocalFormer3D	71.6	73.9	0.257	0.248	0.334	0.232	0.134
EVT	72.1	74.6	0.250	0.245	0.311	0.201	0.123
本文方法	**73.1**	**74.8**	**0.237**	**0.231**	**0.287**	0.251	0.128

4.5 消融实验结果

本节集中对各个提出模块的作用开展深入研究与分析. 所有实验结果均基于 KITTI 验证集获取, 沿用 KITTI 测试集的相同评估指标, 并依据 Recall 40 标准进行计算.

4.5.1 不同设计组件的使用效果

本节研究了每个所提模块对性能的影响. 表 4 中的实验结果表明, 可变形注意力模块提高了检测性能. 它使中

等难度样本的准确率提高了 1.75%, 困难样本的准确率提高了近 5%, 显著提升了在复杂情况下的检测效果. 此外, 引入伪点云数据使简单样本的准确率提高了 5% 以上, 并且在中等难度和困难样本上改善幅度更大, 因为其有助于弥补纯点云数据的稀疏性问题.

表 4 各模块对模型性能的影响 (%)

方法	简单	中等	困难
Voxel R-CNN[25]	89.41	84.52	78.93
+可变形注意力模块	89.55	86.27	83.83
+二维稀疏卷积模块	94.69	87.48	85.64
+双阶互馈模块	95.86	88.71	86.28

4.5.2 可变形注意力模块的效果

本节深入探讨了可变形注意力模块的作用, 实验结果如表 5 所示. 结果表明, 可变形注意力的初始化对模型性能有着显著影响. 在三维空间中随机初始化的采样点会导致性能较差, 因为网格点特征集中在网格位置附近. 因此, 在训练初期, 可变形注意力往往会捕捉到无关特征, 从而减缓收敛速度. 此外, 随着可变形注意力结构中采样点数量的增加, 性能也会得到提升.

表 5 可变形注意力模块对模型性能的影响 (%)

体素池化	可变形注意力			简单	中等	困难
	随机初始化	最近初始化	采样点数量			
√	—	—	—	89.41	84.52	78.93
—	√	—	6	87.36	83.65	76.29
—	—	√	6	89.37	85.31	80.74
—	—	√	26	**89.55**	**86.27**	**83.83**

4.5.3 伪点云特征提取的影响

本节主要研究伪点云特征提取过程中本研究所提方法对于模型性能的影响, 其中主要研究了伪点云的初始特征以及特征提取方法的影响. 本节的实验主要聚焦于伪点云的特征提取, 因此, 在真实点云和伪点云的特征融合方面, 统一采用本研究所提的基于多模态交叉注意力的方法进行融合. 对于伪点云的初始特征, 本研究主要包含两个部分, 分别是通过利用真实点云体素特征得到的多层次表征, 将这部分特征记为深层特征, 还包括伪点云已有的三维坐标、二维图像 RGB 和二维像素坐标特征, 将这部分特征记为浅层特征. 对于深层特征, 本节还研究了参与深层特征生成过程中参与插值的体素数量 K 对模型性能的影响. 本实验中使用的伪点云特征提取方法为二维稀疏卷积方案. 实验结果如表 6 所示.

表 6 伪点云不同的初始特征对模型性能的影响 (%)

浅层特征	深层特征			简单	中等	困难
	K=1	K=6	K=27			
—	—	—	—	89.55	86.27	83.83
√	—	—	—	93.28	86.83	84.37
—	—	√	—	94.92	87.39	85.39
√	√	—	—	95.34	87.69	85.75
√	—	√	—	**95.86**	**88.71**	**86.28**
√	—	—	√	95.55	88.24	85.86

表 6 中第 1 行的实验结果为未使用伪点云特征进行融合的结果. 从实验结果中可得, 对于所有的检测样本, 伪点云特征无论深层特征还是浅层特征均显著提升了模型的性能. 特别是对于简单样本, 伪点云特征提供了更加丰富的特征描述, 大幅提升了模型对于简单样本的检测能力. 从表 6 中还可以看出, 同时使用两类特征对于模型性能

的提升更加显著. 本节还深入研究了深层特征计算过程中邻居体素的数量 K 对于性能的影响. 其中, K=1 表示直接使用伪点云所属体素对应的真实点云的体素特征作为深层次特征, 不进行插值. K=6 表示使用距离伪点云体素最近的 6 个真实点云的体素特征进行插值, 依次类推. 实验结果表明, 本研究所使用的插值技术显著提升了模型的检测性能, 特别是当 K=6 时, 相比于不使用特征插值技术, 中等难度的检测样本提升近 1 个百分点, 其余类别的样本提升在 0.5 个百分点左右. 此外, 通过实验结果还观察到, 随着 K 值的增加, 模型检测性能稍有下降, 分析原因可能是因为参与计算的体素特征较多, 造成了一定的特征模糊.

此外, 我们还将提出的二维稀疏卷积方法 (SPConv2D) 与 SFD 中的 CPConv 方法以及直接使用 PointNet 的方法在检测精度、推理时间、计算复杂度和参数量方面进行了比较. 表 7 所示的实验结果表明, 与其他两种方法相比, PointNet 展现出较低的检测性能和推理效率. 这主要是由于伪点云数量众多, 以及 PointNet 的邻域搜索复杂度较高. 此外, 结果还表明, 我们提出的二维稀疏卷积方法在检测精度、效率、计算复杂度和参数量方面均优于 CPConv.

表 7 伪点云特征提取方法对模型性能的影响

方法	简单 (%)	中等 (%)	困难 (%)	每帧推理时间 (ms)	计算复杂度 (GFLOPs)	参数量 (M)
PointNet	95.06	87.91	85.01	504	42.6	8.8
CPConv	95.47	88.56	85.74	98	28.3	5.2
SPConv2D	95.86	88.71	86.28	53	19.7	4.9

4.5.4 两阶段迭代融合方法的效果

本节关注特征层面和结果层面的集成. 表 8 中的结果表明, 基于网格点的网格级融合因保留了更详细的感兴趣区域 (RoI) 特征, 其性能优于实例级融合. 在融合方法中, 基于拼接的融合优于基于相加的融合, 但这两者都因参数和深度补全误差而存在对齐问题. 所提出的两阶段迭代融合方法显著优于上述两种方法.

表 8 多模态特征融合对模型性能的影响 (%)

融合级别	拼接	相加	交叉注意力	简单	中等	困难
实例级别	√	—	—	94.74	87.72	85.49
	—	√	—	94.67	86.89	84.36
网格点级别	√	—	—	95.67	88.48	85.87
	—	√	—	95.49	88.19	85.71
	—	—	√	**95.86**	**88.71**	**86.28**

在结果融合领域, 本研究主要探究整合不同来源的检测结果对检测性能的影响. 实验结果如表 9 所示. 研究发现, 对于单个模态的检测结果, 伪点云数据由于比真实点云数据更密集, 因此在直接使用时可以产生具有竞争力的检测结果. 此外, 通过融合来自不同传感器的检测结果, 模型的检测性能得到进一步提升.

表 9 多模态结果融合对模型性能的影响 (%)

点云域结果	伪点云域结果	融合域结果	简单	中等	困难
√	—	—	89.55	86.27	83.83
—	√	—	91.33	87.02	85.31
—	—	√	95.27	88.12	85.37
√	—	√	95.39	88.48	85.46
—	√	√	95.42	88.53	85.46
√	√	√	**95.86**	**88.71**	**86.28**

4.6 结果可视化分析

为了更好地评估所提方法, 我们将其检测性能与 Voxel R-CNN 进行比较. 如图 10 所示, 其中红色边界框为基准框, 绿色边界框为本文方法检测框, 蓝色边界框为对比方法检测框, 红色虚线框为漏检的具体目标. 可以看出, 传

统基于点云的方法容易对灌木和建筑物等物体产生误检. 相比之下, 本文方法通过引入图像数据, 有效减少了此类误检. 此外, 对于远处或严重遮挡的目标, 本文方法能够准确检测这些物体, 而纯点云方法则难以应对. 这表明所提出的三维检测算法显著提高了检测性能, 增强了自动驾驶系统的环境感知能力.

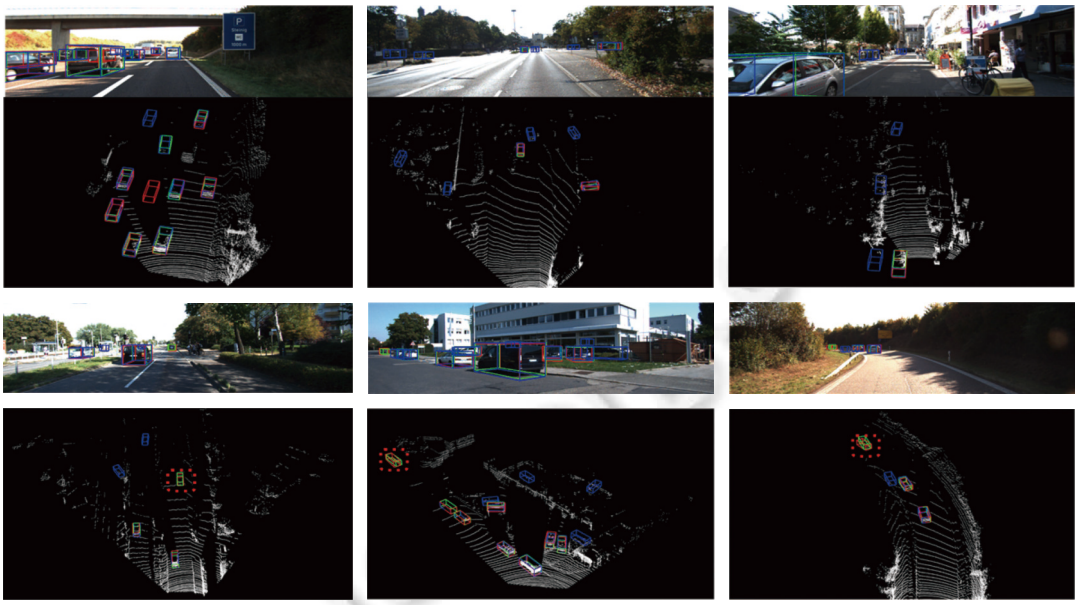

图 10　本文方法与对比方法可视化结果图

5 总　结

本文提出了一种高性能的多模态三维目标检测算法, 有效解决了道路交通场景中多模态融合面临的模态交互不足和特征不对齐等关键问题. 通过改进点云处理流程、优化图像特征提取机制以及设计双阶段互反馈的多模态融合策略, 我们构建了一个高效的三维目标检测框架, 为自动驾驶系统的环境感知提供了重要技术支持. 此外, 实验研究表明, 现有伪点云投影方法在二维平面处理过程中存在明显的局限性, 特别是在目标重叠和遮挡场景下的性能表现欠佳. 基于此发现, 我们提出未来研究将重点探索基于目标级裁剪的独立特征提取方法, 以期获得更精细的局部特征表征能力.

References

[1] Xiang YX, Li YK, Liu JQ, Wang XJ, Chen T, Tong ED, Niu WJ, Han Z. Quantified analysis of congestion situation in intelligent transportation towards frequency-reduced spoofing attack. Ruan Jian Xue Bao/Journal of Software, 2023, 34(2): 833–848 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/6416.htm [doi: 10.13328/j.cnki.jos.006416]

[2] Chen CT, Zhao JQ, Ye C, Deng R, Guan LT, Li DY. Inter-process asynchronous communication in autonomous vehicle based on shared memory. Ruan Jian Xue Bao/Journal of Software, 2017, 28(5): 1315–1325 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5144.htm [doi: 10.13328/j.cnki.jos.005144]

[3] Li GS, Liu Y, Zheng QB, Yang GL, Liu K, Wang Q, Diao XC. Review on multi-sensor data fusion research for unmanned aerial vehicles. Ruan Jian Xue Bao/Journal of Software, 2025, 36(4): 1881–1905 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/7273.htm [doi: 10.13328/j.cnki.jos.007273]

[4] Sun HF, Mu ZY, Qi Q, Wang JY, Liu C, Liao JX. Fast and accurate depth completion method based on dynamic gated fusion strategy. Ruan Jian Xue Bao/Journal of Software, 2023, 34(4): 1765–1778 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/6399.htm [doi: 10.13328/j.cnki.jos.006399]

[5] Geiger A, Lenz P, Urtasun R. Are we ready for autonomous driving? The KITTI vision benchmark suite. In: Proc. of the 2012 IEEE Conf. on Computer Vision and Pattern Recognition. Providence: IEEE, 2012. 3354–3361. [doi: 10.1109/CVPR.2012.6248074]

[6] Qi CR, Su H, Kaichun M, Guibas LJ. PointNet: Deep learning on point sets for 3D classification and segmentation. In: Proc. of the 2017 IEEE Conf. on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 77–85. [doi: 10.1109/CVPR.2017.16]

[7] Qi CR, Yi L, Su H, Guibas LJ. PointNet++: Deep hierarchical feature learning on point sets in a metric space. In: Proc. of the 31st Int'l Conf. on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 5105–5114.

[8] Qi CR, Litany O, He KM, Guibas LJ. Deep Hough voting for 3D object detection in point clouds. In: Proc. of the 2019 IEEE/CVF Int'l Conf. on Computer Vision. Seoul: IEEE, 2019. 9276–9285. [doi: 10.1109/ICCV.2019.00937]

[9] Shi SS, Wang XG, Li HS. PointRCNN: 3D object proposal generation and detection from point cloud. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 770–779. [doi: 10.1109/CVPR.2019.00086]

[10] Sheng HL, Cai SJ, Liu Y, Deng B, Huang JQ, Hua XS, Zhao MJ. Improving 3D object detection with channel-wise Transformer. In: Proc. of the 2021 IEEE/CVF Int'l Conf. on Computer Vision. Montreal: IEEE, 2021. 2723–2732. [doi: 10.1109/ICCV48922.2021.00274]

[11] Yang ZT, Sun YN, Liu S, Jia JY. 3DSSD: Point-based 3D single stage object detector. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 11037–11045. [doi: 10.1109/CVPR42600.2020.01105]

[12] Zhang YF, Hu QY, Xu GQ, Ma YX, Wan JW, Guo YL. Not all points are equal: Learning highly efficient point-based detectors for 3D LiDAR point clouds. In: Proc. of the 2022 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 18931–18940. [doi: 10.1109/CVPR52688.2022.01838]

[13] Chen C, Chen Z, Zhang J, Tao DC. SASA: Semantics-augmented set abstraction for point-based 3D object detection. In: Proc. of the 36th AAAI Conf. on Artificial Intelligence. AAAI Press, 2022. 221–229. [doi: 10.1609/aaai.v36i1.19897]

[14] Lang AH, Vora S, Caesar H, Zhou LB, Yang J, Beijbom O. PointPillars: Fast encoders for object detection from point clouds. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 12689–12697. [doi: 10.1109/CVPR.2019.01298]

[15] Shi GS, Li RF, Ma C. PillarNet: Real-time and high-performance pillar-based 3D object detection. In: Proc. of the 17th European Conf. on Computer Vision. Tel Aviv: Springer, 2022. 35–52. [doi: 10.1007/978-3-031-20080-9_3]

[16] Li JY, Luo CX, Yang XD. PillarNeXt: Rethinking network designs for 3D object detection in LiDAR point clouds. In: Proc. of the 2023 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023. 17567–17576. [doi: 10.1109/CVPR52729.2023.01685]

[17] Zhou Y, Tuzel O. VoxelNet: End-to-end learning for point cloud based 3D object detection. In: Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 4490–4499. [doi: 10.1109/CVPR.2018.00472]

[18] Yan Y, Mao YX, Li B. SECOND: Sparsely embedded convolutional detection. Sensors, 2018, 18(10): 3337. [doi: 10.3390/s18103337]

[19] Duan KW, Bai S, Xie LX, Qi HG, Huang QM, Tian Q. CenterNet: Keypoint triplets for object detection. In: Proc. of the 2019 IEEE/CVF Int'l Conf. on Computer Vision. Seoul: IEEE, 2019. 6568–6577. [doi: 10.1109/ICCV.2019.00667]

[20] Ge RZ, Ding ZZ, Hu YH, Wang Y, Chen SJ, Huang L, Li Y. AFDet: Anchor free one stage 3D object detection. arXiv:2006.12671, 2020.

[21] He CH, Zeng H, Huang JQ, Hua XS, Zhang L. Structure aware single-stage 3D object detection from point cloud. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 11870–11879. [doi: 10.1109/CVPR42600.2020.01189]

[22] Zheng W, Tang WL, Chen SJ, Jiang L, Fu CW. CIA-SSD: Confident IoU-aware single-stage object detector from point cloud. In: Proc. of the 35th AAAI Conf. on Artificial Intelligence. AAAI Press, 2021. 3555–3562. [doi: 10.1609/aaai.v35i4.16470]

[23] Yang HH, Wang WX, Chen MH, Lin BB, He T, Chen H, He XF, Ouyang WL. PVT-SSD: Single-stage 3D object detector with point-voxel Transformer. In: Proc. of the 2023 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023. 13476–13487. [doi: 10.1109/CVPR52729.2023.01295]

[24] Shi GS, Li RF, Ma C. Pillar R-CNN for point cloud 3D object detection. arXiv:2302.13301, 2023.

[25] Deng JJ, Shi SS, Li PW, Zhou WG, Zhang YY, Li HQ. Voxel R-CNN: Towards high performance voxel-based 3D object detection. In: Proc. of the 35th AAAI Conf. on Artificial Intelligence. AAAI Press, 2021. 1201–1209. [doi: 10.1609/aaai.v35i2.16207]

[26] Wu H, Deng JH, Wen CL, Li X, Wang C, Li J. CasA: A cascade attention network for 3-D object detection from LiDAR point clouds. IEEE Trans. on Geoscience and Remote Sensing, 2022, 60: 5704511. [doi: 10.1109/TGRS.2022.3203163]

[27] Wu H, Wen CL, Li W, Li X, Yang RG, Wang C. Transformation-equivariant 3D object detection for autonomous driving. In: Proc. of the 37th AAAI Conf. on Artificial Intelligence. Washington: AAAI Press, 2023. 2795–2802. [doi: 10.1609/aaai.v37i3.25380]

[28] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. Attention is all you need. In: Proc. of the 31st Int'l Conf. on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 6000–6010.

[29] Wang HY, Shi C, Shi SS, Lei M, Wang S, He D, Schiele B, Wang LW. DSVT: Dynamic sparse voxel Transformer with rotated sets. In: Proc. of the 2023 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023. 13520–13529. [doi: 10.1109/CVPR52729.2023.01299]

[30] Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A, Zagoruyko S. End-to-end object detection with Transformers. In: Proc. of the 16th European Conf. on Computer Vision. Glasgow: Springer, 2020. 213–229. [doi: 10.1007/978-3-030-58452-8_13]

[31] Zhu BJ, Wang Z, Shi SS, Xu H, Hong LQ, Li HS. ConQueR: Query contrast voxel-DETR for 3D object detection. In: Proc. of the 2023 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023. 9296–9305. [doi: 10.1109/CVPR52729.2023.00897]

[32] Zhou C, Zhang YN, Chen JX, Huang D. OcTr: Octree-based Transformer for 3D object detection. In: Proc. of the 2023 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023. 5166–5175. [doi: 10.1109/CVPR52729.2023.00500]

[33] Huang TT, Liu Z, Chen XW, Bai X. EPNet: Enhancing point features with image semantics for 3D object detection. In: Proc. of the 16th European Conf. on Computer Vision. Glasgow: Springer, 2020. 35–52. [doi: 10.1007/978-3-030-58555-6_3]

[34] Liu Z, Huang TT, Li BL, Chen XW, Wang X, Bai X. EPNet++: Cascade bi-directional fusion for multi-modal 3D object detection. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2023, 45(7): 8324–8341. [doi: 10.1109/TPAMI.2022.3228806]

[35] Vora S, Lang AH, Helou B, Beijbom O. PointPainting: Sequential fusion for 3D object detection. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 4603–4611. [doi: 10.1109/CVPR42600.2020.00466]

[36] Wang CW, Ma C, Zhu M, Yang XK. PointAugmenting: Cross-modal augmentation for 3D object detection. In: Proc. of the 2021 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 11789–11798. [doi: 10.1109/CVPR46437.2021.01162]

[37] Yin TW, Zhou XY, Krähenbühl P. Multimodal virtual point 3D detection. In: Proc. of the 35th Int'l Conf. on Neural Information Processing Systems. Curran Associates Inc., 2021. 1261.

[38] Chen XZ, Ma HM, Wan J, Li B, Xia T. Multi-view 3D object detection network for autonomous driving. In: Proc. of the 2017 IEEE Conf. on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 6526–6534. [doi: 10.1109/CVPR.2017.691]

[39] Ku J, Mozifian M, Lee J, Harakeh A, Waslander SL. Joint 3D proposal generation and object detection from view aggregation. In: Proc. of the 2018 IEEE/RSJ Int'l Conf. on Intelligent Robots and Systems (IROS). Madrid: IEEE, 2018. 1–8. [doi: 10.1109/IROS.2018.8594049]

[40] Yoo JH, Kim Y, Kim J, Choi JW. 3D-CVF: Generating joint camera and LiDAR features using cross-view spatial feature fusion for 3D object detection. In: Proc. of the 16th European Conf. on Computer Vision. Glasgow: Springer, 2020. 720–736. [doi: 10.1007/978-3-030-58583-9_43]

[41] Pang S, Morris D, Radha H. CLOCs: Camera-LiDAR object candidates fusion for 3D object detection. In: Proc. of the 2020 IEEE/RSJ Int'l Conf. on Intelligent Robots and Systems (IROS). Las Vegas: IEEE, 2020. 10386–10393. [doi: 10.1109/IROS45743.2020.9341791]

[42] Pang S, Morris D, Radha H. Fast-CLOCs: Fast camera-LiDAR object candidates fusion for 3D object detection. In: Proc. of the 2022 IEEE/CVF Winter Conf. on Applications of Computer Vision. Waikoloa: IEEE, 2022. 3747–3756. [doi: 10.1109/WACV51458.2022.00380]

[43] Shen BQ, Dai SW, Chen YY, Xiong R, Wang Y, Jiao YM. GOOD: General optimization-based fusion for 3D object detection via LiDAR-camera object candidates. arXiv:2303.09800, 2023.

[44] Qi CR, Liu W, Wu CX, Su H, Guibas LJ. Frustum PointNets for 3D object detection from RGB-D data. In: Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 918–927. [doi: 10.1109/CVPR.2018.00102]

[45] Xu DF, Anguelov D, Jain A. PointFusion: Deep sensor fusion for 3D bounding box estimation. In: Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 244–253. [doi: 10.1109/CVPR.2018.00033]

[46] Wang ZX, Jia K. Frustum ConvNet: Sliding frustums to aggregate local point-wise features for amodal 3D object detection. In: Proc. of the 2019 IEEE/RSJ Int'l Conf. on Intelligent Robots and Systems (IROS). Macao: IEEE, 2019. 1742–1749. [doi: 10.1109/IROS40897.2019.8968513]

[47] Zhu HQ, Deng JJ, Zhang Y, Ji JM, Mao QY, Li HQ, Zhang YY. VPFNet: Improving 3D object detection with virtual point based LiDAR and stereo data fusion. IEEE Trans. on Multimedia, 2023, 25: 5291–5304. [doi: 10.1109/TMM.2022.3189778]

[48] Wu XP, Peng L, Yang HH, Xie L, Huang CX, Deng CQ, Liu HF, Cai D. Sparse fuse dense: Towards high quality 3D detection with depth completion. In: Proc. of the 2022 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 5418–5427. [doi: 10.1109/CVPR52688.2022.00534]

[49] Bai XY, Hu ZY, Zhu XG, Huang QQ, Chen YL, Fu HB, Tai CL. TransFusion: Robust LiDAR-camera fusion for 3D object detection with Transformers. In: Proc. of the 2022 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 1080–1089. [doi: 10.1109/CVPR52688.2022.00116]

[50] Li YW, Yu AW, Meng TJ, Caine B, Ngiam J, Peng DY, Shen JY, Lu YF, Zhou D, Le QV, Yuille A, Tan MX. DeepFusion: LiDAR-camera deep fusion for multi-modal 3D object detection. In: Proc. of the 2022 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 17161–17170. [doi: 10.1109/CVPR52688.2022.01667]

[51] Imran S, Liu XM, Morris D. Depth completion with twin surface extrapolation at occlusion boundaries. In: Proc. of the 2021 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 2583–2592. [doi: 10.1109/CVPR46437.2021.00261]

[52] OpenPCDet Development Team. OpenPCDet: An open-source toolbox for 3D object detection from point clouds. 2020. https://github.com/open-mmlab/OpenPCDet

[53] Liang M, Yang B, Wang SL, Urtasun R. Deep continuous fusion for multi-sensor 3D object detection. In: Proc. of the 15th European Conf. on Computer Vision (ECCV). Munich: Springer, 2018. 663–678. [doi: 10.1007/978-3-030-01270-0_39]

[54] Liang M, Yang B, Chen Y, Hu R, Urtasun R. Multi-task multi-sensor fusion for 3D object detection. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 7337–7345. [doi: 10.1109/CVPR.2019.00752]

[55] Wang ZJ, Zhao Z, Jin Z, Che ZP, Tang J, Shen CM, Peng YX. Multi-stage fusion for multi-class 3D lidar detection. In: Proc. of the 2021 IEEE/CVF Int'l Conf. on Computer Vision Workshops. Montreal: IEEE, 2021. 3113–3121. [doi: 10.1109/ICCVW54120.2021.00347]

[56] Chen YK, Li YW, Zhang XY, Sun J, Jia JY. Focal sparse convolutional networks for 3D object detection. In: Proc. of the 2022 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 5418–5427. [doi: 10.1109/CVPR52688.2022.00535]

[57] Yang HH, Liu ZL, Wu XP, Wang WX, Qian W, He XF, Cai D. Graph R-CNN: Towards accurate 3D object detection with semantic-decorated local graph. In: Proc. of the 17th European Conf. on Computer Vision. Tel Aviv: Springer, 2022. 662–679. [doi: 10.1007/978-3-031-20074-8_38]

[58] Lin ZW, Shen YQ, Zhou SP, Chen ST, Zheng NN. MLF-DET: Multi-level fusion for cross-modal 3D object detection. In: Proc. of the 32nd Int'l Conf. on Artificial Neural Networks. Heraklion: Springer, 2023. 136–149. [doi: 10.1007/978-3-031-44195-0_12]

[59] Wang ML, Zhao L, Yue YF. PA3DNet: 3-D vehicle detection with pseudo shape segmentation and adaptive camera-LiDAR fusion. IEEE Trans. on Industrial Informatics, 2023, 19(11): 10693–10703. [doi: 10.1109/TII.2023.3241585]

[60] Tian YL, Zhang XJ, Wang X, Xu JT, Wang JG, Ai R, Gu WH, Ding WP. ACF-Net: Asymmetric cascade fusion for 3D detection with LiDAR point clouds and images. IEEE Trans. on Intelligent Vehicles, 2024, 9(2): 3360–3371. [doi: 10.1109/TIV.2023.3341223]

[61] Zhang GX, Xie J, Liu L, Wang ZP, Yang KH, Song ZY. URFormer: Unified representation LiDAR-camera 3D object detection with Transformer. In: Proc. of the 6th Chinese Conf. on Pattern Recognition and Computer Vision (PRCV). Xiamen: Springer, 2024. 401–413. [doi: 10.1007/978-981-99-8435-0_32]

附中文参考文献

[1] 相迎宵, 李轶珂, 刘吉强, 王潇瑾, 陈彤, 童恩栋, 牛温佳, 韩臻. 面向降频污染攻击的智能交通拥堵态势量化分析. 软件学报, 2023, 34(2): 833–848. http://www.jos.org.cn/1000-9825/6416.htm [doi: 10.13328/j.cnki.jos.006416]

[2] 陈存铜, 赵君峤, 叶晨, 邓蓉, 管林挺, 李德毅. 基于共享内存的智能无人车进程间消息异步传输机制. 软件学报, 2017, 28(5): 1315–1325. http://www.jos.org.cn/1000-9825/5144.htm [doi: 10.13328/j.cnki.jos.005144]

[3] 李庚松, 刘艺, 郑奇斌, 杨国利, 刘坤, 王强, 刁兴春. 无人机多传感器数据融合研究综述. 软件学报, 2025, 36(4): 1881–1905. http://www.jos.org.cn/1000-9825/7273.htm [doi: 10.13328/j.cnki.jos.007273]

[4] 孙海峰, 穆正阳, 戚琦, 王敬宇, 刘聪, 廖建新. 基于动态门控特征融合的轻量深度补全算法. 软件学报, 2023, 34(4): 1765–1778. http://www.jos.org.cn/1000-9825/6399.htm [doi: 10.13328/j.cnki.jos.006399]

作者简介

唐文能, 硕士生, 主要研究领域为计算机视觉, 三维目标检测.

李垚辰, 博士, 副教授, 博士生导师, CCF 专业会员, 主要研究领域为计算机视觉, 智能交通, 具身智能.

高笙景, 硕士生, 主要研究领域为计算机视觉, 三维目标检测.

高聪, 硕士生, 主要研究领域为计算机视觉, 三维目标检测.

彭越涵, 硕士生, 主要研究领域为计算机视觉, 三维目标检测.

刘跃虎, 博士, 教授, 博士生导师, 主要研究领域为计算机视觉, 具身智能.