智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

自动驾驶中深度学习的三维目标检测方法综述

吴一全, 蔡佳琦

引用本文:
吴一全, 蔡佳琦. 自动驾驶中深度学习的三维目标检测方法综述[J]. 智能系统学报, 2026, 21(2): 297-320.
WU Yiquan, CAI Jiaqi. Deep learning-based 3D object detection for autonomous driving:a comprehensive review[J]. *CAAI Transactions on Intelligent Systems*, 2026, 21(2): 297-320.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202504021

DOI: 10.11992/tis.202504021
网络出版地址: https://link.cnki.net/urlid/23.1538.tp.20250818.1757.002

自动驾驶中深度学习的三维目标检测方法综述

吴一全, 蔡佳琦
(南京航空航天大学 电子信息工程学院, 江苏 南京 211106)

摘　要: 自动驾驶技术的快速发展对车辆感知系统准确性和实时性的要求日益提升。三维目标检测作为车辆感知系统的核心组成部分, 对于确保行车安全和提升驾驶体验至关重要。首先将三维目标检测算法按传感器所获取的数据类型分为 3 类: 视觉算法 (包括基于二维特征和三维特征的子类)、激光点云算法 (涵盖网格化点云、原始点云和混合点云)、基于多传感器的算法 (按照网络串行融合和并行融合的方式进行分类)。据此总结了具体算法的特点、贡献及局限性。随后, 介绍了典型三维目标检测数据集及其评价指标, 并比较了代表性算法在不同数据集上的性能。最后, 分析了当前技术面临的挑战, 并对未来发展方向进行了展望。

关键词: 自动驾驶; 三维目标检测; 深度学习; 点云; 多传感器融合; 卷积神经网络; 数据集; 性能评价指标
中图分类号: TP391.41　　**文献标志码**: A　　**文章编号**: 1673−4785(2026)02−0297−24

中文引用格式: 吴一全, 蔡佳琦. 自动驾驶中深度学习的三维目标检测方法综述 [J]. 智能系统学报, 2026, 21(2): 297–320.
英文引用格式: WU Yiquan, CAI Jiaqi. Deep learning-based 3D object detection for autonomous driving: a comprehensive review[J]. CAAI transactions on intelligent systems, 2026, 21(2): 297–320.

Deep learning-based 3D object detection for autonomous driving: a comprehensive review

WU Yiquan, CAI Jiaqi
(School of Electronic Information Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China)

Abstract: The rapid advancement of autonomous driving technology has increasingly heightened the demands for the accuracy and real-time performance of vehicle perception systems. 3D Object Detection, as a core component of vehicle perception systems, is of vital importance for ensuring driving safety and enhancing the driving experience. Firstly, 3D object detection algorithms are categorized into three types based on the data types acquired by sensors: Visual algorithms encompass subcategories based on 2D and 3D features; LiDAR point cloud algorithms cover grid-based point clouds, raw point clouds, and hybrid point cloud approaches; multi-sensor-based algorithms are classified based on the modes of serial and parallel fusion of the network. Accordingly, the features, contributions, and limitations of specific algorithms are summarized. Subsequently, typical 3D object detection datasets and their evaluation metrics are reviewed, and the performance of representative algorithms on different datasets is compared. Finally, the current technical challenges are analyzed, and the future development directions are prospected.

Keywords: autonomous driving; 3D object detection; deep learning; point cloud; multi-sensor fusion; convolutional neural network; dataset; performance evaluation metrics

作为幅员辽阔但人口分布呈现显著集聚特征的国家, 我国 2024 年城市通勤监测数据显示[1], TOP10 城市中大部分的通勤高峰的拥堵指数较上年有所提升。自动驾驶汽车依靠人工智能、视觉计算、雷达、监控装置和全球定位系统协同合作[2], 能够提升驾驶安全性, 优化交通系统效率, 并为用户节省时间。自动驾驶车辆自主行驶时, 需要对周围三维 (3D) 场景进行感知, 而 3D 目标检测用于获取目标在 3D 空间中的位置和类别信息, 是自动驾驶感知系统的基础, 对后续的路径规划、运动预测、碰撞避免具有重要指导作用。

在自动驾驶场景中, 相机和激光雷达通常分别用于获取 RGB 图像和 3D 点云信息。传统的

收稿日期: 2025−04−24.　**网络出版日期**: 2025−08−19.
基金项目: 国家自然科学基金项目 (61573183).
通信作者: 吴一全. E-mail: nuaaimage@163.com.

3D 目标检测方法通常以图像作为输入，借助 2D 目标检测模型提取 2D 边界框，再通过深度估计将其映射到 3D 空间。2D 目标检测算法中的特征提取器和分类器是手动设计的，图像特征通过人工选取，这些特征在面对多变的环境时缺乏足够的适应性，因此在复杂的场景中表现不佳。随着深度学习的发展，研究者凭借其擅长处理大规模图像数据的优点逐步用卷积神经网络替代手工特征提取器，目标检测算法向基于深度学习的方向发展，检测精度、速度不断提升。相比于传统手工设计特征的方法，基于深度学习的方法避免了繁琐的手工设计过程，能够自动学习更具有区分力的深度特征，并将特征提取和分类器学习统一在一个框架中，实现端到端的学习。

研究人员进行了广泛而深入的研究，致力于探索 3D 目标检测领域的各个维度。目前，相关综述文献情况如下：文献 [3-8] 分别介绍了 3D 目标检测中某一类 (视觉、点云、多模态) 的深度学习检测方法，探讨了这些方法的原理、优势及局限性。然而，单一类别的综述方式在展现全面性和跨领域视角时稍显局限。文献 [9-12] 在自动驾驶领域针对基于深度学习的 3D 目标检测研究进行了综述，具有通用性，但涉及的方法不够全面深入。其中文献 [9-11] 缺少对各类算法的归纳总结，未能全面梳理各类算法的核心思想、优缺点及相互之间的关系。文献 [12] 对自动驾驶中的 3D 目标检测技术进行了广泛讨论，但并未对该领域数据集进行总结。而本文在前人研究的基础上，系统梳理了深度学习在 3D 目标检测中的应用，围绕基于视觉、激光点云与多传感器融合三大方法，构建了统一的分类框架。同时，针对不同检测方法的性能特性，本文进一步归纳了各类算法之间的性能差异，分析其核心设计理念与技术路线。此外，本文还对现有主流数据集进行了系统总结，分析了数据集特性对算法评估与应用场景适应性的影响，力求为相关领域研究者提供更加系统全面的参考。

本文梳理和分析了大量基于深度学习的 3D 目标检测方法，旨在帮助研究人员可以快速、系统地了解相关技术。考虑至此，文中做了以下安排：第 1 部分概述基于视觉的 3D 目标检测技术，按照检测时使用的特征分为 2D 特征和 3D 特征两个方面进行概括总结；第 2 部分叙述基于激光点云的 3D 目标检测技术，根据对点云数据的处理形式不同将其分为 3 种类型，基于网格化点云的方法、基于原始点云的方法、基于混合点云的方法；第 3 部分阐述了基于多传感器融合的 3D 目标检测方法，介绍激光点云与图像融合及毫米波点云与图像融合两种不同的组合方式，依照串行融合和并行融合框架对其分类；第 4 部分梳理了 3D 目标检测常用数据集以及性能评估指标，并对比了典型算法的性能。最后，针对现有的 3D 目标检测方法的不足做了展望。

1　基于视觉的 3D 目标检测方法

早期的 3D 目标检测任务本质上是 2D 自动驾驶图像信息与现实道路场景几何和深度信息的一种结合。因此在研究初期，学者将 3D 任务解耦为 2D 图像检测和其他参数的预测。根据使用的特征类型，基于视觉的检测算法可分为基于 2D 特征和基于 3D 特征两类。

1.1　基于 2D 特征的 3D 目标检测方法

此类方法的处理流程通常包括 3 个步骤：首先，从获取的车辆图像中提取 2D 特征；然后，利用这些特征得到车辆在图像平面中的类别和 2D 位置信息；最后，通过几何推算或深度学习方法将 2D 位置信息映射到 3D 空间，最终获得检测目标的 3D 位置和姿态信息。可进一步细分为 3 类技术路径：空间几何约束、图像模板识别和其他方法。基于 2D 特征的 3D 目标检测流程如图 1 所示。

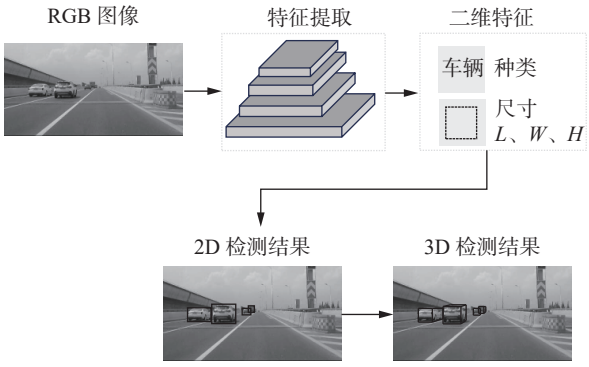


图 1　基于 2D 特征的 3D 目标检测流程

Fig. 1　3D object detection process based on 2D features

1.1.1　空间几何约束

车辆和行人在交通道路上的位置关系，如遮挡、截断和尺寸变化，对几何约束至关重要。通过分析这些关系可推断车辆和行人的 3D 位置以及形状。

Mousavian 等[13] 在原始 2D 检测框架上提出了 Deep3DBox，添加分支提取局部坐标系中的车辆旋转角度，并减小了子类别平均尺寸的偏差。在获取车辆的 3D 结构信息时，仅使用 2D 边界框会带来表征模糊问题，Li 等[14] 将车辆可见表面的

视觉特征作为网络输入, 有效缓解了这一问题。为了提高 2D-3D 特征的匹配度, M3DSSD(monocular 3D single stage object detector)[15] 将形状和中心作为特征对齐的主要参数, 减少了预测结果与特征图之间的差距。

自动驾驶车辆配备的双目摄像机通过组合不同焦距和基线的立体模块, 提供更广泛的感知能力。Qin 等[16] 提出的 TLNet(triangulation learning network) 建立了车辆图像与感兴趣区域 (region of interest, ROI) 之间的对应关系, 从而实现 3D 空间中的三角测量和目标检测。LIGA (learning LiDAR geometry aware)-Stereo[17] 引导双目 3D 检测模型学习更多车辆的几何信息。针对自动驾驶产业的实时性需求, 迟旭然等[18] 通过双层特征融合网络缩短特征之间的传递路径, 从而缩短检测时间。

1.1.2　图像模板识别

现有机动车大多由已知的物体部件组成, 因此研究人员利用尺寸信息估计车辆距离。3DOP (3D object proposals)[19] 以立体车辆图像对作为输入, 采用类似 Fast R-CNN (regions with convolutional neural network)[20] 的两步检测过程, 将 2D 图像的像素级坐标投影至 3D 空间计算点云, 实现目标位置回归。Chabot 等[21] 利用细粒度 3D 模型区分汽车类型, 使用相似模板的顶点信息完成匹配。3D-R-CNN[22] 从 CAD 模型集合中学习低维形状空间, 结合车辆的形状先验, 将图像区域映射为物体的 3D 形状和位姿。Kreiss 等[23] 通过点检测模型[24] 定位街道上的行人, 同时预测随机不确定性, 提升无人驾驶中的行人安全性。Li 等[25] 和 Cai 等[26] 均将单目 3D 检测转化为稀疏关键点检测问题。由于此类方法忽略了车辆的整体形状, AutoShape[27]基于 n 个 2D 关键点, 建模车辆的形状特征, 在 KITTI 测试集上平均精度 (average precision, AP) 较基准提升 2.72%。在此基础上, 近期研究逐步从固定模板方法转向可学习的几何先验建模。Shuai 等[28] 提出了一种关键点监督的形状流形, 使得模型能够灵活适应跨类别的几何形态; Duan 等[29] 则构建了基于立体关键点对齐优化的可变形模板, 有效提升了在遮挡与视角变化场景下的姿态估计精度。

1.1.3　其他方法

除了上述方法, 研究人员利用不确定性理论估计单一像素点, 以预测车辆中心点位置。Chen 等[30] 提出的 MonoPair(monocular 3D object detection using pairwise spatial relationships) 采用全局优化的思想, 将深度估计中的不确定性纳入预测过程。Ma 等[31] 利用异方差偶然不确定性估计中心位置。Zhang 等[32] 提出的 MonoFlex 结合利用高度比预测物体深度的思想与不确定性理论, 通过集成学习的方式估计车辆的中心位置。汪萌等[33] 引入深度和航向角的不确定性, 提高了远距离车辆和斜向目标的检测精度。

Huang 等[34] 和 Li 等[35] 全局深度信息集成到 Transformer 中, 利用注意力机制融合多点输入的时空信息, 以提高视觉感知模型的性能。Wang 等[36] 基于知识蒸馏, 提出多相机鸟瞰图 (bird's-eye-view, BEV) 检测框架, 通过引导学生模型模仿预训练激光雷达检测器所提取的特征, 增强学生模型的表达能力。

1.2　基于 3D 特征的 3D 目标检测方法

此类方法的处理流程通常包括 3 个步骤: 首先, 从获取的车辆图像中提取 2D 特征或将图像转换为 3D 伪点云数据; 然后, 实现 2D-3D 特征升维或利用 3D 数据提取目标的 3D 特征; 最后, 在 3D 空间中实现汽车和行人的定位、识别和姿态估计, 最终获得目标的 3D 信息。可进一步细分为两类技术路径: 3D 深度估计和伪激光点云。基于 3D 特征的 3D 目标检测流程如图 2 所示。

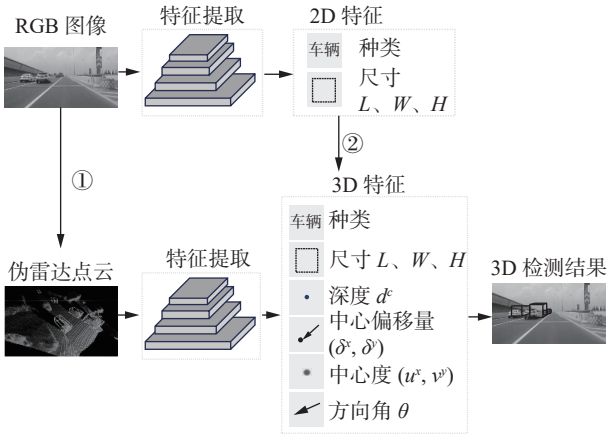


图 2　基于 3D 特征的 3D 目标检测流程

Fig. 2　3D object detection process based on 3D features

1.2.1　3D 深度估计

目标车辆在交通道路中与传感器之间的距离信息是实现精确目标检测的一个关键因素。MF3D (multi-level fusion based 3D object detection)[37] 通过集成 MonoDepth[38] 实现深度推断。D4LCN(depth-guided dynamic-depthwise-dilated local convolutional network)[39] 构建了一个由深度图引导的特征提取网络, 从 RGB 图像中提取更丰富的深度信息。Peng 等[40] 提出的 DID-M3D(decoupling instance depth for monocular 3D object detection) 将深度信息

解耦为车辆视觉深度和车辆属性深度的组合，实现了对车辆深度信息的精确估计。此外，OFT(orthographic feature transform)[41] 的思想是将前视图转换为 BEV，在 BEV 空间中进行目标检测。PETR (position embedding transformation)[42] 将 3D 坐标信息编码到图像特征中，生成具有 3D 位姿感知的特征，进行端到端的目标检测。

双目相机的匹配误差对双目深度估计至关重要。Žbontar 等[43] 提出将匹配过程转化为计算两个车辆图像的相似度，并利用监督信号指导神经网络提取适用于匹配任务的图像特征，从而提升匹配精度。为进一步优化匹配误差计算过程，GC-Net(geometry and context network)[44]引入几何和上下文引导信息，提高了深度估计的准确性和鲁棒性。

1.2.2　伪激光点云

伪激光点云是一种通过从图像中预测深度信息并生成与激光雷达点云相似的 3D 点云数据的技术。Wang 等[45] 提出的 Pseudo-LiDAR 利用 DRON（deep reinforcement objective-oriented network）[46] 或 PSMNET（pyramid stereo matching network）[47] 进行深度估计，然后结合原始图像和深度信息生成伪激光点云，实现了车辆 3D 检测框的回归。ForeSeE(foreground-background separated monocular depth estimation)[48] 将交通环境中的前景与背景分离，使用独立的优化目标和解码器分别估计前景车辆与背景环境的深度信息，解决了伪激光雷达定位不准确的问题。Li 等[49] 提出的 CG (confidence guided)-Stereo 将深度估计网络的置信度作为 3D 车辆检测器中的软注意力机制，引导检测网络关注高质量深度信息的点云区域。Hossain 等[50]基于多输入残差网络编码器，提出了一种自监督立体深度估计方法。Oh 等[51] 通过立体图像生成 3D 边界框和伪 LiDAR 点云，显著减少了 LiDAR 数据标注的时间与成本。

1.3　小结

表 1 对比了几类基于视觉的 3D 目标检测算法的贡献与局限性。当前，自动驾驶领域仍面临小样本学习、大规模计算需求和复杂场景应用等挑战，解决这些问题已成为学术界和工业界共同关注的焦点。传统视觉算法主要依赖低级特征，难以获取高级语义特征及复杂场景信息。随着深度学习技术的不断发展，基于视觉的方法在场景理解和目标识别方面取得了显著进展，进一步提升了基于视觉的方法在车辆检测中的重要性。尽管基于视觉和基于激光点云的方法在性能上存在差异，但由于相机相较于激光雷达成本更低，因此在某些场景下相机仍然是激光雷达的一种替代选择。

表 1　基于视觉的 3D 目标检测算法对比

Table 1　Comparison of vision-based 3D object detection algorithms

类别		文献	具体方法	贡献	局限性
单目	空间几何约束	[13]	分别预测车辆尺寸，转角与得分	解决过度约束的优化问题	依赖于2D边界框的准确检测
		[14]	利用2D边框估计3D车辆中心	消除了表征模糊	检测精度一般
		[15]	两步特征对齐方法	解决特征不匹配问题	算法复杂度高
	图像模板匹配	[21]	将车辆部件与CAD模型匹配	提高了检测性能	深度估计方面仍然面临挑战
		[22]	提出一种快速逆图形网络	将CAD模型体素化	用于梯度近似的有限差分会在训练期间引起歧义
		[26]	建模车辆的形状信息	可精确估计目标车辆的位置	复杂样本的改进效果有限
	其他方法	[30]	采用全局优化的思想	提高遮挡车辆检测的准确性	计算复杂度较高
		[33]	引入深度和航向角的不确定性	提升斜向目标检测精度	需进一步轻量化改进
		[34]	引入Transformer	提高了检测精度	需要额外LiDAR数据辅助训练
		[36]	引入教师–学生模型	改善了跨模态知识转移	模态和模型结构差异限制知识提炼
	3D深度估计	[39]	通过深度图引导特征提取网络	克服了传统2D卷积无法捕获物体多尺度信息的问题	引入无法消除的深度估计误差
		[41]	将RGB图像投影为BEV	缓解前视图带来的遮挡问题	反向映射结果不准确
	伪激光点云	[45]	通过单目/双目相机生成伪点云	首个引入伪点云的方法	精度受限，难以处理反射和遮挡
		[48]	将前景与背景分离	提升了前景目标的深度估计性能	距离较远时，目标检测效果一般

续表 1

类别		文献	具体方法	贡献	局限性
双目	空间几何约束	[16]	建立3D锚框和感兴趣区域之间的对应关系	对目标车辆进行三角测量	耗费大量计算资源
		[17]	使用基于LiDAR的探测器引导基于立体的探测器	增强整体的几何和语义表示	—
	图像模板匹配	[19]	采用两步检测过程	以立体图像对作为输入	检测精度一般
	3D深度估计	[43]	将匹配过程转化为计算两个图像块的相似度	首个用于立体匹配的深度学习方法	需要大量的训练数据和计算资源
		[44]	使用3D CNN聚合4D成本体	提高深度估计的准确性和鲁棒性	在复杂场景表现不佳
	伪激光点云	[51]	通过立体图像生成3D边界框和伪LiDAR点云	支持自动生成伪LiDAR注释数据	类间表现不均衡

2　基于激光点云的 3D 目标检测方法

点云可以通过大量的数据点表达车辆表面的特征信息以及空间分布信息。相对于 2D 图像, 点云能更有效地还原真实的交通道路信息, 修饰车辆的 3D 特征, 而且受到光照、姿态变化的影响更小。因此, 如何高效利用点云数据是近年研究人员关注的重点。

2.1　基于网格化点云的 3D 目标检测方法

2.1.1　基于 2D 视图的 3D 目标检测方法

不同视图能够表现点云的不同特点, 将 3D 点云数据从不同“视角”投影至 2D 平面, 可以获得具有互补信息的特征表示, 从而充分挖掘点云的空间结构信息。常用的投影方式主要包括前视图、距离视图及 BEV。通过融合来自不同视角的特征, 可以提升点云处理任务的准确性和鲁棒性。

Li 等[52] 将点云转换到与车辆图片相似的前视图, 利用“点云伪图像”进行目标检测。考虑到前视图中多个点云可能映射到同一图像坐标, PIXOR(pixel-wise oriented object detector)[53] 将点云转换为紧凑的车辆 BEV 表示, 避免了前视图的遮挡问题。然而 BEV 会丢失大量纵轴信息, 因此 BirdNet[54] 将点云投影转化为包含高度、强度和密度的三通道 BEV 特征图, 利用 2D 检测器结合高度信息生成 3D 检测框。Meyer 等[55] 提出的 LaserNet 设计了基于距离视图的检测框架, 通过融合多模态分布实现目标检测。然而, 由于尺度变化和车辆遮挡问题, 距离视图在 3D 物体检测中的表现不如预期。依赖单一视图进行环境感知极易造成关键信息的丢失, 仅依靠前视摄像头的单一视角, 可能无法及时察觉侧后方突然出现的车辆或行人。为此, 研究人员提出结合多个视图进行联合检测的方案。Deng 等[56] 提出了 H3D (hallucinated hollow-3D) R-CNN, 将前视图的语义特征和原始点云特征传递至鸟瞰图进一步提取特征, 形成双视图的 3D 表示。Sun 等[57] 提出了 RSN (range sparse net), 结合距离视图和鸟瞰图, 采用双阶段检测, 提高远距离车辆的检测精度。

将点云数据转化为多个 2D 视图, 自动驾驶系统能获取更全面准确的路况信息, 极大提升决策的准确性与行车安全性, 从而构建出完整、精准的环境模型, 辅助自动驾驶车辆做出更可靠的决策, 有效降低事故风险。

2.1.2　基于体素的 3D 目标检测方法

在自动驾驶中, 点云数据因场景动态性和传感器扫描机制的不连续性而呈现不规则分布, 例如车辆、行人和障碍物的位置多变导致点云密度不均匀。将点云数据转换为紧凑形状的体素是处理不规则点云数据常用的方法, 其数据表现形式通常为体素 (Voxel) 或柱体 (Pillar), 如图 3 所示。

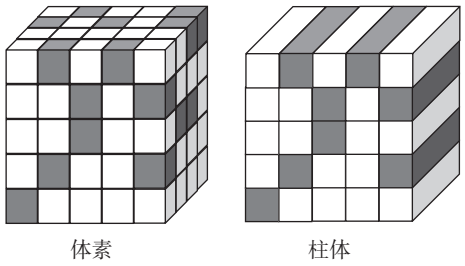


图 3　数据表现形式

Fig. 3　Manifestation form of data

Zhou 等[58] 提出了端到端网络 VoxelNet, 将车辆点云划分为体素, 通过扩展感受野聚合体素特征, 提升上下文形状描述。但算法计算复杂度较高, 实时性较差。因此, Yan 等[59] 提出了 SECOND (sparsely embedded convolutional detection), 引入仅对非空体素进行操作的稀疏卷积, 从而提高检测速度。PointPillars[60] 将点云数据划分为不同的柱体, 学习点云数据的特征。Li 等[61] 设计通过膨胀卷积处理不同尺寸的车辆, 增强特征提取的深度

与精度，改善了 PointPillars[60] 对小目标的检测精度较低的问题。然而，柱体编码方法会丢失部分细粒度信息，并且点云特征提取能力有限，因此，后续研究倾向于将点云编码为体素，这是基于效率和精度之间的权衡。为了能充分利用体素信息，Kuang 等[62] 提出了 Voxel-FPN（voxel-based feature pyramid network），其框架中的编码器自底向上提取多尺度体素特征，解码器自顶向下融合这些特征，并关联至最终自动驾驶系统检测输出。TANet(triple attention network)[63] 通过融合通道、点和体素注意力，实现了对点云特征的多层次建模。Chen 等[64] 提出的 Voxelnext 基于稀疏体素特征进行 3D 目标检测，避免了传统方法中依赖手工设计以及稀疏到密集转换的额外计算成本。

Zheng 等[65] 提出的 CIA-SSD(confident iou-aware single-stage object detector) 引入置信度校正模块对齐定位精度和分类置信度，提升检测精度。SST(single stride)[66] 借助 Transformer 模型解决了单步架构中感受野不足的问题，但占用内存较大。为此，He 等[67] 提出了基于体素的集合注意力模块，能够建模任意大小标记集群的长期依赖关系。SA-SSD(structure aware single-stage 3D object detection)[68] 将每个点视为非零点并量化为张量索引，在不增加额外开销的情况下具有更好的定位性能。考虑到自动驾驶道路点云中的背景冗余，Ada3D[69] 选择性过滤冗余 3D 体素和 2D BEV 特征，并通过稀疏性保留批归一化减少背景干扰。

将点云降维为规则的体素和柱体是自动驾驶中常用的高效点云处理方式，能简化数据结构并提升检测性能。然而，降采样操作虽然减少了数据量，但基于体素的方法主要关注车辆的空间位置等几何信息，可能丢失重要的车辆纹理特征；同时，为保持体素结构完整性而进行的无效填充操作，增加了额外的计算负担。

2.2　基于原始点云的 3D 目标检测方法

在自动驾驶场景中，激光雷达点云因车辆动态性和交通环境干扰呈现不规则分布，基于原始点云的方法通过直接在 3D 点云数据上进行特征提取，可以更精准地保留车辆细节，提升检测精度。3 种特征提取网络结构如图 4 所示。

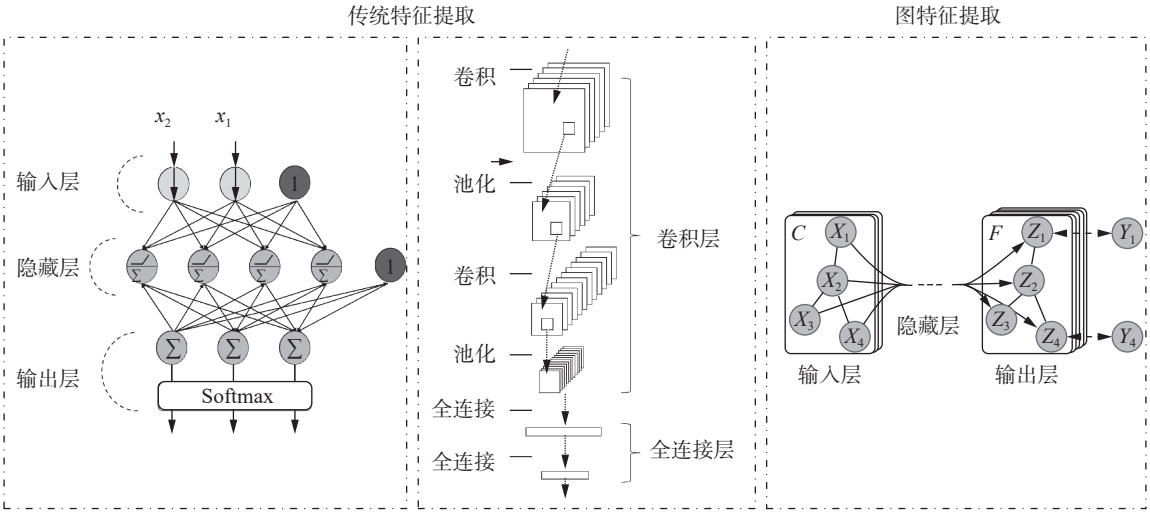


图 4　3 种特征提取网络结构

Fig. 4　Three feature extraction network architectures

2.2.1　基于多层感知机的 3D 目标检测方法

基于多层感知机的 3D 目标检测方法通过处理环境点云，识别出车辆、行人、障碍物等目标信息，为自动驾驶提供基础环境信息。

Qi 等[70] 提出了直接以点云作为输入的 PointNet，采用最大池化策略解决点云的无序性问题，并通过学习转移矩阵实现点云规范化。然而，PointNet 在捕捉车辆局部结构特征方面存在不足，限制了其在复杂场景中的泛化能力。因此后续研究中的 PointNet++[71] 利用分层提取特征的思想，增强了网络的鲁棒性。STD(sparse-to-dense)[72] 利用能涵盖任意角度的球形锚框生成更精确的候选框，从而减少锚框数量。针对恶劣天气条件下雨雪颗粒对激光雷达点云的干扰问题，陈熙源等[73] 提出了一种基于 KD 树（K-dimensional tree）的空间滤波方法。该方法通过计算不同欧氏距离下离群点的马氏距离，有效识别并去除点云中的雨雪噪声，显著提升了自动驾驶系统在复杂气象条件下的感知可靠性和安全性。Objformer[74] 通过实例特征编码器提取车辆的几何结构特征和语义信

息, 并通过实例交互模块实现标签引导的特征融合, 从而增强了对车辆表面属性的鲁棒性表征能力。

在自动驾驶感知任务中, 由于车辆前景点在整个场景中所占比例较小, 采用随机下采样策略可能进一步减少前景点的数量, 从而加剧点云数据的稀疏性问题, 影响目标检测的准确性。因此 SASA(semantics augmented set abstraction)[75] 为距离指标附加语义权重, 使采样点在保留更多车辆的同时覆盖更广范围。IA-SSD[76] 利用 2 种实例感知下采样策略, 分层选取前景车辆。王理嘉等[77] 通过多帧点云配准实现稀疏点云稠密化, 使用融合帧点云提升低线束激光雷达的检测精度。HEDNet (hierarchical encoder-decoder network)[78] 通过构建编码器−解码器结构, 有效捕捉点云特征之间的长程依赖关系, 增强了特征表达的全局一致性。

2.2.2　基于卷积神经网络的 3D 目标检测方法

卷积神经网络能够有效地提取点云中的空间信息, 并通过深度学习模型学习复杂的特征表示。PointCNN[79] 通过学习输入车辆点的 X 变换, 实现将无序点云数据有序化。CenterPoint[80] 将 3D 目标检测转化为关键点估计问题, 直接预测车辆中心点及其属性。针对点云的稀疏性和物体遮挡导致的车辆结构信息不完整问题, 涂新奎等[81] 提出的 SSG(symmetric shape generation)-R-CNN, 为每个前景车辆预测镜像对称点, 从而恢复车辆的完整结构。陶乐等[82] 基于 CenterPoint[80], 引入编码解码稀疏模块、自校正卷积和大核注意力模块, 增强了 3D 特征提取和点云特征聚合, 在 nuScenes 数据集上算法的平均精度均值和 NDS(nuScenes detection score) 相较于基准模型分别提升了 5.97% 和 3.62%。周昊等[83] 通过球面投影将点云转换为深度图像, 结合 YOLO (you only look once) v7 有效补全体素化过程中的信息损失。

2.2.3　基于图神经网络的 3D 目标检测方法

由于点云特有的几何特性, 研究者们开始尝试应用图神经网络 (graph neural network, GNN) 探索非结构化点云数据的几何关系。Point GNN [84] 将点云中的每个点表示为图的节点, 点之间通过边连接表示。通过多层图神经网络的堆叠, 从而逐步提取和聚合点云的几何信息。Najibi 等[85] 提出的 DOPS(detect objects and predict shapes) 利用图卷积聚合每个点的目标框估算, 并使用特定分支预测车辆的 3D 形状编码表示。PC (point cloud)-RGNN[86] 设计的局部−全局注意力机制和多尺度图上下文聚合增强了点之间的关系编码, 解决了遮挡区域检测效果较差的问题。

在自动驾驶感知系统中, 直接处理原始点云数据, 摒弃体素化或降采样等操作, 有助于精准捕捉车辆的细节与形态特征, 避免关键信息的丢失。然而, 多层感知机在提取车辆局部细节特征方面仍存在局限, 基于 CNN 的方法在处理不规则点云时效率较低, 而 GNN 虽能通过点间关联增强场景理解, 但计算开销较高。此外, 自动驾驶场景下的点云数据通常具有高度不规则性和稀疏性, 且易受动态环境中的噪声干扰, 这些挑战仍需进一步优化解决。

2.3　基于混合点云的 3D 目标检测方法

基于体素的方法具有较高的计算效率, 基于点云的方法计算成本较高。混合点云方法结合两者优势, 在提高检测精度的同时优化计算效率, 在 3D 目标检测任务中展现出广阔的应用前景。

Liu 等[87] 提出的 PV(point-voxel)Conv 以点的形式存储车辆数据, 并通过体素化卷积提高局部感知能力, 解决了点云的不规则和稀疏性带来的数据访问耗时问题。HVPR(hsaybrid voxel-point representation)[88] 通过记忆模块强化点特征, 使其在语义上类似体素特征, 形成伪图像形式的混合 3D 表示。PV-RCNN[89] 将多尺度交通场景体素特征编码为关键点。由于前景车辆和交通背景之间存在严重不平衡, Wu 等[90]提出了 PV-RCNN++, 通过语义点−体素特征交互提升检测性能。MPV (multi point-voxel)Conv[91] 采用 3D 卷积神经网络, 增强了点特征和体素特征之间的信息关联和非线性。鉴于点云和体素的融合方式过于简单会造成特征语义混淆, 李虎辰等[92] 设计了点分支和体素分支分别提取车辆粗粒度和细粒度特征, 从而增强车辆点云的全局特征表达能力。PVC(point-voxel dual-channel fusion with cascade point estimation)-SSD[93] 用无锚点方法对点和体素进行双通道融合编码, 实现了局部和全局特征的建模。

混合点云方法在复杂的交通环境中展现出强大的检测能力, 实现了计算效率与识别精度的平衡。通过融合车辆点云的细粒度特征与体素的高效计算结构, 该方法能够精准捕捉目标形状与边缘信息, 并快速完成交通目标定位。然而, 体素参数的优化及点云与体素之间的有效信息传递限制了其在动态交通场景中的进一步应用。

2.4　小结

通过上述文献分析可见, 近年来学者们主要针对以下问题展开研究: 点云的稀疏性导致自动驾驶目标检测精度不足; 点云在不同数据形式转

换过程中存在信息丢失，未能充分利用空间信息；模型参数量过大，难以满足实际工业需求等。此外，研究的重点逐渐倾向在满足产业化检测精度的同时，提高检测的速率，从而实现自动驾驶中的安全、高效的目标检测。实际应用中，需要根据具体需求和资源限制选择合适的算法，并进一步优化网络以提高性能和效果。回顾近年来基于激光点云的算法研究进展，图 5 按时间顺序，梳理了近几年标志性的激光点云检测方法。表 2 给出了各类算法的特点。

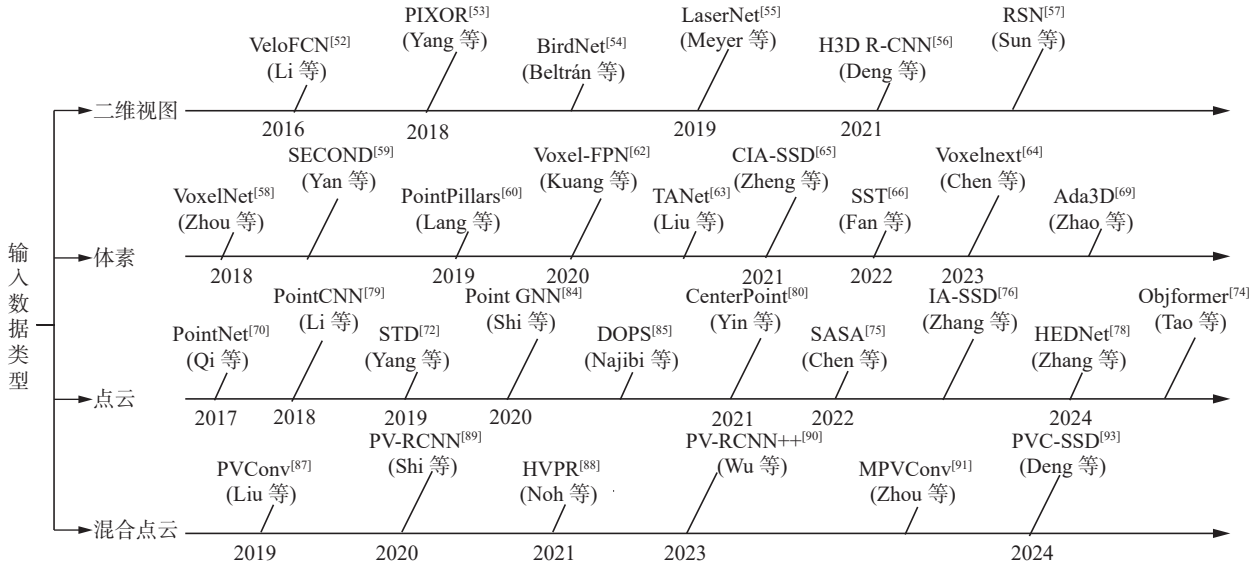


图 5　激光点云目标检测方法

Fig. 5　LiDAR point cloud object detection method

表 2　基于激光点云算法对比

Table 2　Comparison of algorithms based on LiDAR point clouds

类别	文献	具体方法	贡献	局限性
基于2D视图	[52]	将点云转换到前视图	降低了计算成本	前视图会丢失部分信息
	[53]	将点云转换到BEV	避免了前视图的遮挡问题	丢失大量纵轴信息
	[54]	点云投影转化三通道BEV特征图	提供高精度和远距离物体检测	检测效果一般
	[55]	将点云转换到距离视图	用距离视图完成3D目标检测任务	尺度变化和遮挡问题导致检测效果差
	[56]	双视图表示点云	实现与基于点的方法相当高的精度	计算量大
	[57]	在距离视图上进行前景分割	计算量小，检测效率高	只关注单帧点云的处理
基于体素	[58]	端到端的网络框架	提升上下文形状描述	运算量较大，且速度不佳
	[59]	引入稀疏3D卷积	计算量和内存占用减少	行人和骑车者检测性能较差
	[60]	将点云划分为柱体	使用2D卷积实现端到端学习	对小物体的检测性能较差
	[61]	三分支扩展卷积网络	使用膨胀卷积处理不同尺寸物体	丢失细粒度信息
	[63]	引入三元注意力模块	增强目标的关键信息	受3D卷积局部感受野限制
	[66]	基于Transformer架构	解决单步架构中感受野不足	网络模型占用内存大
基于点云	[70]	直接对点云进行特征提取操作	提取到更多的细节特征	消耗大量计算成本
	[72]	用基于点的球形锚框生成候选框	降低计算量	不进行上采样，损失性能
	[75]	为距离指标附加语义权重	保留更多的前景点	采样易受噪声点云的影响
	[76]	实例感知下采样策略	分层选取感兴趣对象的前景点	类别分布不均衡时，精度低
	[79]	学习输入点的X变换	解决了点云的不规则性问题	—
	[81]	基于对称形状生成	恢复目标的完整结构	—
	[84]	基于图GNN	引入图表示方法	对点云稀疏区域和点云遮挡区域的识别较差
基于混合点云	[86]	结合点云补全模块	提高了远距离稀疏点云的检测精度	模型需要进一步轻量化
	[91]	多点体素卷积	增强了点和体素之间的信息关联	计算复杂度高
	[93]	无锚点方法	有效融合了目标的局部和全局特征	—

3　基于多传感器融合的 3D 目标检测方法

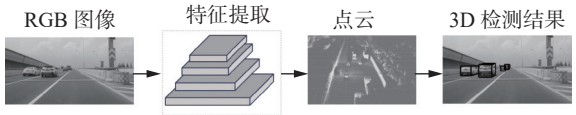
自动驾驶车辆依赖于多种传感器的协同工作, 以实现精准的环境感知。视觉传感器提供丰富的纹理信息, 激光雷达通过点云数据提供精确的深度感知, 而毫米波雷达则能在恶劣天气条件下确保稳定的目标检测。通过融合不同传感器的数据, 自动驾驶系统能够显著提高目标检测的精度和鲁棒性, 从而增强车辆在复杂驾驶环境中的适应能力与安全性。

3.1　多模态数据融合框架

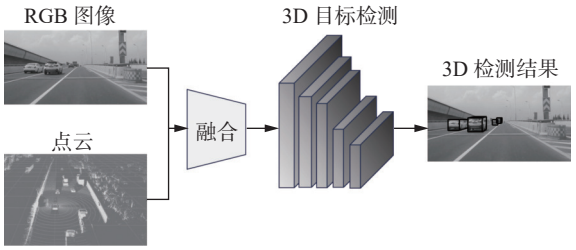
针对多传感器融合的 3D 目标检测, 可以将模型分为串行融合和并行融合两种方式。

3.1.1　串行融合

串行融合使用 2D 检测模型的结果缩小 3D 目标检测的搜索空间, 从而减少点云中的背景点, 如图 6(a) 所示。每个传感器的特征提取和融合按照特定的顺序进行, 后续阶段的特征提取严重依赖于前一阶段的输出。这种逐步整合传感器信息的方式, 能提高目标检测的效率和精度, 尤其是在复杂的驾驶场景中, 帮助系统更准确地识别和定位目标。



(a) 串行融合



(b) 并行融合

图 6　多模态融合方法

Fig. 6　Multimodal fusion method

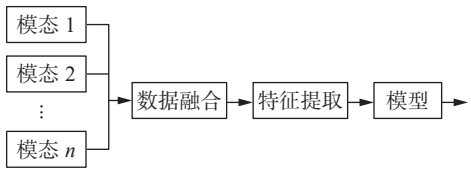
3.1.2　并行融合

在并行融合中, 各个传感器的特征提取和融合同时进行, 如图 6(b) 所示, 每个传感器的数据经过独立的特征提取模块。根据融合的时机, 可以将并行融合方法分为数据融合、特征融合和决策融合 3 种方式, 如图 7 所示。

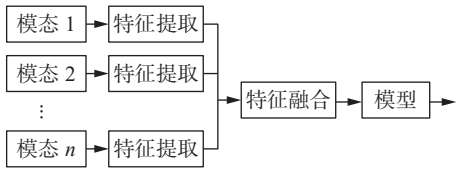
1) 数据融合: 也称为早融合, 是指在输入阶段将来自不同模态的数据进行融合。数据融合能够充分利用不同模态数据的丰富信息, 有助于提高自动驾驶系统的性能和鲁棒性, 尤其在复杂环境下的目标检测与障碍物识别中, 但过于依赖标定矩阵和对齐效果。

2) 特征融合: 也称为深度融合。在特征提取阶段, 将各传感器提取的特征通过拼接、注意力机制或自适应融合等连接成一个高维特征向量。这种方法在自动驾驶中的多模态感知任务中应用广泛, 通过深度融合不同传感器的特征, 能够提升目标检测、语义分割等任务的精度和鲁棒性。特征融合可以灵活地处理不同模态之间的关系, 但需要额外的网络设计和训练。

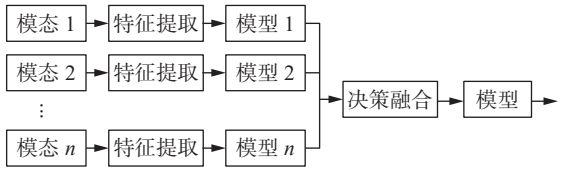
3) 决策融合: 也称为后期融合, 不同模态的数据分别经过特征提取和决策处理后, 通过多个预测类别标签的代数组合规则对模型输出结果进行融合。决策融合的目标是通过整合独立决策的优势, 提升自动驾驶系统的整体性能与鲁棒性。然而, 决策融合无法有效捕捉不同传感器数据间的内在关联, 可能在某些复杂场景中表现不如特征融合。



(a) 数据融合



(b) 特征融合



(c) 决策融合

图 7　并行融合

Fig. 7　Parallel fusion

3.2　主流的融合算法

自动驾驶汽车通常配备多个传感器, 在目标检测任务中, 融合传感器可以利用不同传感器的优势, 获得更准确的目标检测结果, 提高目标检测的准确性和鲁棒性。以下将从不同的传感器组合角度概述不同传感器融合之间的特点。

3.2.1　基于激光点云与图像融合的方法

激光雷达能够提供精确的距离和几何信息, 对于地面检测和障碍物定位非常有效, 但缺乏色彩信息, 在分类方面可能存在局限。结合激光雷

达与相机，可以利用相机的高分辨率图像信息进行物体识别和场景理解，进而提高自动驾驶感知的全面性和准确性。目前，研究者们主要聚焦于点云稀疏性、特征对齐及多模态融合策略等方面的改进。

Chen 等[94] 提出了 MV3D(multi-view 3D object detection)，将激光雷达点云和图像数据深度融合，映射到 BEV、前视图和 RGB 图像 3 个维度，采用多视图对点云进行编码。AVOD(aggregate view object detection)[95] 网络则结合 BEV 与 RGB 图像，利用特征金字塔网络和几何约束提升整体检测性能。为更精准地定位 3D 车辆，F-PointNets[96] 将 2D 检测框投影到点云中，形成视锥体后进行距离语义分割和边界框预测。该方法对前景点分割效果依赖较大，且层次结构复杂。因此，Wang 等[97] 提出 F(frustum)-ConvNet，通过滑动视锥体机制聚合局部点向量特征，增强局部点云的特征表达。而 PointPainting[98] 利用来自图像的语义分割信息辅助点云信息，提高了检测效果。

1) 点云稀疏性

针对激光雷达采集的交通场景点云数据稀疏性导致的 3D 目标检测困难问题，Liang 等[99] 提出的 MMF(multi-task multi-sensor fusion) 通过补全伪点云的方式增强点云稠密性，并结合路面估计和图像特征实现特征融合。但这种融合策略比较简单，容易忽略图像和点云中的对象部分对应关系。基于此，Wu 等[100] 提出了 SFD(sparse fuse dense)，采用细粒度网格化策略，融合原始点云与伪点云的感兴趣区域特征，提升稀疏区域目标识别精度。黄漫等[101] 则结合 IP-Basic 算法，在消除噪声的同时将稀疏雷达点云补全成稠密深度图。Xie 等[102] 提出了使用稀疏候选对象产生稀疏表达的 SparseFusion，使用轻量级注意力模块融合各模态特征，同时使用跨模态信息转移方法，避免各单一模态的缺陷对融合阶段造成负向影响。

2) 特征对齐

在自动驾驶中，对齐不同模态的特征一直是提升多模态感知精度的关键。Zhang 等[103] 提出了 CAT-Det(contrastively augmented Transformer for detection) 算法，引入 Transformer 架构实现局部与全局特征的融合。Gunn 等[104] 使用交叉注意力机制将相机与激光雷达特征结合，进一步提升特征关联性。Liu 等[105] 提出 BEVFusion，通过将激光雷达和摄像头数据统一映射到共享的鸟瞰视图空间，同时保留了几何与语义信息，提升了自动驾驶系统对复杂场景的理解能力。为了充分利用车辆图像中的语义信息，PVF (perspective voxel fusion)-DectNet[106] 将每个点划分为单独的透视体素，投影到图像特征图上，实现了体素与图像特征的融合。DeepFusion[107]利用交叉注意机制动态捕获点云和图像之间的相关性，增强了多模态信息的互补性。周治国等[108]使用多层多模态融合，解决了视角错位与特征匹配问题。

3) 融合方法

Liu 等[109] 提出的 PVConvNet 框架对点云和图像进行体素特征级融合。金宇锋等[110] 设计双域融合特征区域建议模块融合激光点云和图像特征。PCDR-DFF(point cloud diversity representation and dual feature fusion)[111] 结合多点融合模型与嵌入稀疏 3D-U Net 的多特征融合机制，提高了自动驾驶系统目标检测的准确性。考虑到在特征提取过程中某一模态失效影响整体系统模型，王五岳等[112] 将激光雷达和红外相机作为独立分支，通过门控注意力机制融合两种模态特征，从而实现高效的 3D 目标检测。董钰婷等[113] 提出采用自适应加权融合激光雷达和相机，消除不同传感器感知能力之间的差异，从而实现了激光雷达与相机的高效融合。Mono-BEVFusion(monocular-bird's eye view fusion)[114] 将点云数据和 RGB 数据构建为 BEV 特征，进一步增强自动驾驶系统目标检测的稳定性和可靠性。

改进数据融合算法，提高传感器配准和特征对齐的准确性，同时研究更有效的跨模态特征融合方法，是未来提升目标检测性能的发展方向。

3.2.2　基于毫米波点云与图像融合的方法

激光雷达使用激光束进行环境感知，而毫米波雷达则使用无线电波进行探测。相比于激光雷达，低成本的毫米波雷达具有生产成本更低、不受恶劣天气条件的限制、具有较强的远距离检测能力，能提供车辆速度信息等优势。因此，研究人员一直在寻求将图像和毫米波点云融合的新方法，提升自动驾驶系统的感知能力。

CenterFusion[115] 提出了一种将雷达探测点与目标中心点关联的方法，通过生成联合特征图来实现目标属性的回归。Bansal 等[116] 独立提取图像中的纹理和语义信息，同时融合雷达数据，避免了某一模态失效对自动驾驶系统性能的影响，实现了可靠的特征提取。

由于雷达精度低且存在测量歧义，Kim 等[117] 提出了 CRAFT(camera-radar 3D object detection with spatio-contextual fusion Transformer) 框架。该框架采用软极坐标关联方法，实现了图像与雷达

点的有效匹配, 解决了 3D 坐标系与车辆空间属性之间的差异问题。CRN(camera radar net)[118] 通过雷达辅助的视图变换将图像特征转换到 BEV 空间, 并利用多模态特征聚合层在 BEV 空间中对图像和雷达特征进行深度融合, 有效解决了输入模态之间的空间不对齐问题。车俐等[119] 通过空间对齐和扩展毫米波雷达与相机信息, 避免了自动驾驶中天气和光照变化对目标检测的影响。Liu 等[120] 提出了一种稀疏 3D 物体查询的雷达与图像特征融合方法, 该方法在特征采样与聚合时保留了原始雷达特征的完整性, 为自动驾驶决策提供了全面且精确的 3D 目标检测支持。

综上所述, 现有方法主要采用特征级融合, 通过对雷达点云与图像特征的对齐实现多源信息的有效融合。部分研究进一步设计了专用网络模块, 以优化融合过程并提升信息传递的完整性。这些方法显著提升了自动驾驶系统在复杂环境中的感知能力, 尤其在毫米波雷达与图像融合方面取得了重要进展。

3.3　小结

表 3 对比了几类基于多传感器融合算法的特点。研究表明, 通过融合多模态数据, 能够有效应对自动驾驶场景中的尺度与视角变化问题, 提供更全面、鲁棒的目标检测结果。相较于单一模态的检测方法, 基于多传感器融合的检测方法在检测精度、推理速度及实时性能等方面均展现出显著优势。未来研究应进一步聚焦于融合算法的高性能优化、融合策略的精度提升, 以及在动态复杂环境下的鲁棒性增强, 从而推动自动驾驶系统向实用化与智能化方向持续发展。

表 3　基于多传感器融合的算法对比

Table 3　Comparison of algorithms based on multi-sensor fusion

融合类型	融合结构	文献	具体方法	贡献	局限性
激光点云与图像	数据融合	[95]	结合鸟瞰图与RGB图像	计算成本不高	融合数据方式较简单
		[105]	将多模态数据映射至BEV空间	保留几何与语义信息	遮挡导致密集特征不准确
	特征融合	[94]	多视角投影融合	多模态融合	小尺寸目标检测效果一般
		[99]	补全伪点云	利用深度补全辅助目标检测	检测精度依赖于深度补全
		[103]	引入Transformer	减小了模态之间的差异	网络模型内存占用大
		[100]	细粒度网格化融合	提升了稀疏区域的检测精度	网络复杂计算量大
		[107]	交叉注意机制	实现点云与像素精确配准	计算复杂度高
		[112]	门控注意力机制	模型有效、易于扩展	—
	决策融合	[102]	借稀疏候选对象产生稀疏表达	跨模态对齐对象中心特征	严重依赖单个分支特征质量
	串行融合	[96]	将2D检测框投影为视椎体	更精准地定位3D目标	对前景点分割效果敏感
		[97]	采用滑动视锥体机制	增强了点云的特征表达	强依赖于2D网络的检测输出
毫米波点云和视图	特征融合	[115]	通过视椎法关联检测点	解决毫米波雷达分辨率低的问题	相较激光雷达方法性能不足
		[116]	独立提取图像中的纹理、语义信息	避免某一模态失效影响整体模型	独立提取特征操作耗时
		[117]	软极坐标关联方法	实现了稳健而专注的融合	网络模型效率较低
	串行 融合	[118]	利用点云信息增强摄像头BEV特征	有效整合了点云的位置信息	网络性能受点云变化影响

4　数据集及性能指标评价

3D 目标检测离不开大规模的数据集, 为了模拟和训练自动驾驶系统, 用于无人驾驶 3D 目标检测的数据集层出不穷。本节介绍了常用的自动驾驶 3D 目标检测数据集, 列举了典型数据集的评估指标, 并对经典算法的性能指标进行了比较。

4.1　常用数据集

KITTI 数据集[121] 是自动驾驶和计算机视觉领域常用的公开数据集, 包括广角相机和激光雷达传感器, 提供大规模城市道路场景的多模态数据, 涵盖立体图像、光流、视觉测距、3D 物体检测和 3D 跟踪等评测任务。该数据集包含 7 481 个训练样本和 7 518 个测试样本, 涵盖包括汽车、行人和骑行者在内的 8 个目标类别。由于测试数据集标签不开放, 研究人员通常将训练集划分为训练和验证子集。

nuScenes 数据集[122] 由 nuTonomy 公司发布, 提供多传感器的全景 360°采集数据。该数据集包含超过 1 000 个场景, 每个场景持续约 20 s, 总数

据量超过 100 h。每个样本包括车辆、行人等目标的标注信息以及路面、车道线和交通标志等环境标注。除了雷达和摄像头数据外，nuScenes 还包括毫米波雷达数据，并覆盖夜间及不同天气条件下的场景，数据多样且丰富。

Waymo Open 数据集[123]由配备 5 个激光雷达和 5 个高分辨率针孔相机的采集车辆收集，覆盖城市街道、高速公路和乡村道路等环境。与 KITTI 数据集相比，Waymo Open 数据集提供更全面的 360°目标标注，并根据激光雷达点云数量将目标分为 LEVEL_1 和 LEVEL_2 两个级别。

ApolloScape 数据集[124]是目前行业内环境最复杂、标注最精准的 3D 自动驾驶公开数据集，由百度 Apollo 自动驾驶平台发布。该数据集涵盖多个城市在不同照明条件下的驾驶场景，提供稠密且精确的点云数据。

KAIST 数据集[125]主要用于行人检测与跟踪。其数据采集设备包括 RGB 摄像头、热成像相机、立体视觉、激光雷达和 GPS 惯导传感器，涵盖校园、街道和乡村等交通场景，包含日出、早晨、下午、日落、夜晚和黎明不同时间段数据。然而，该数据集的数据来源局限于单一城市和季节，且使用的相机存在一定的图像噪音。

DAIR-V2X 数据集[126]是全球首个大规模多模态车路协同自动驾驶数据集，涵盖车端和路端相机、LiDAR 等传感器数据。数据采集于北京市高级别自动驾驶示范区，覆盖 10 km 城市道路、10 km 高速公路和 28 个路口，并考虑了不同天气与时间段。数据集包含 71 253 帧原始图像和 71 254 帧点云数据，附带标注、时间戳和标定文件等辅助信息。

除上述介绍的常用数据集外，还有如 Cityscapes 3D [127]数据集、Argoverse2 [128]数据集以及基于仿真平台构建的虚拟数据集等，专门用于自动驾驶的 3D 目标检测任务，进一步丰富了该领域在复杂城市环境中的研究场景与测试条件。

针对 3D 目标检测任务，现有主流数据集在传感器类型、数据规模、标注精度及场景复杂度等方面存在显著差异。以 KITTI 为代表的早期数据集主要采用单一模态传感器，数据量有限，适用于验证轻量级模型在简单道路环境下的检测性能。而 nuScenes、Waymo Open 等多模态数据集，融合了激光雷达与多视角相机信息，并涵盖更为复杂的城市交通场景，使得算法能够在更真实的应用场景下进行性能评估。此外，数据集在标注粒度上的差异，如 BEV 标注、3D 立体框标注等，也影响了检测算法对空间几何特性的建模能力。因此，合理选择与目标应用场景特性相匹配的数据集，是实现算法评估公平性与应用适应性的重要前提。为便于读者查阅与使用，表 4汇总了常用数据集的数据信息及下载地址。

表 4　公共数据集列表

Table 4　List of public data sets

数据集	图像	点云	3D框	类别	拍摄环境	下载地址
KITTI[121]	200 KB	1.2 MB	80 KB	8	白天	http://www.cvlibs.net/datasets/kitti/raw_data.php
nuScenes[122]	1.4 MB	390 KB	1.4 MB	23	白天+夜晚+雨天	https://www.nuscenes.org/nuscenes#download
Waymo Open[123]	12 MB	230 KB	12 MB	4	白天+夜晚+雨天	https://waymo.com/open/download/
ApolloScape[124]	144 KB	29 KB	70 KB	35	白天+夜晚	https://apolloscape.auto/scene.html#to_down_href
KAIST[125]	95 KB	—	—	3	白天+夜晚	https://sites.google.com/view/complex-urban-dataset/download-lidar#h.sa42osfdnwst
Cityscapes 3D[127]	5 KB	0	—	8	白天	https://www.cityscapes-dataset.com/downloads/
Argoverse2[128]	2.7 MB	150 KB	—	26	白天+夜晚+雨天	https://www.argoverse.org/av2.html#lidar-link
Lyft L5[129]	240 KB	46 KB	1.3 MB	9	白天+夜晚+雨天	https://level5.lyft.com/dataset/?source=post_page
DAIR-V2X[126]	73 KB	73 KB	1.4 MB	15	白天+夜晚+雨天	https://thudair.baai.ac.cn/index
PandaSet[130]	41 KB	16 KB	—	28	白天+夜晚+雨天	https://scale.com/resources/download/pandaset
The H3D Dataset[131]	83 KB	27 KB	1.1 MB	8	—	https://usa.honda-ri.com/H3D
STCrowd[132]	158 KB	219 KB	219 KB	1	白天+夜晚+雨天	https://github.com/4DVLab/STCrowd.git
SynLiDAR[133]	—	19 482 MB	—	32	—	https://github.com/xiaoaoran/SynLiDAR

4.2　典型数据集性能指标评价

4.2.1　KITTI 数据集性能指标评价

KITTI 使用平均精度 (average precision, AP) 衡量边界框匹配性能，平均方向相似性 (average orientation similarity, AOS) 用于衡量预测的 3D 物体边界框与真实标注框之间的方向差异。将预测

的 3D 边界框与真实标注框的交并比 (intersection over inion, IoU) 和 3D 阈值进行比较, 若 IoU 大于阈值, 则为真阳性 (true positive, TP), 否则为假阳性 (false positive, FP)。据此计算精确率 P 和召回率 R, 具体计算公式为

$$P=\frac{N_{\mathrm{TP}}}{N_{\mathrm{TP}}+N_{\mathrm{FP}}}$$

$$R=\frac{N_{\mathrm{TP}}}{N_{\mathrm{TP}}+N_{\mathrm{FN}}}$$

式中: N_{TP}为真阳性样本数, N_{FP}为假阳性样本数, N_{FN}为假阴性样本数。精确率反映了真实正样本在分类器确定的正样本中所占的比例, 召回率表示被正确判定为正例的样本占所有实际正例的样本的比例。在计算平均精度时, 通常会使用不同的精确率阈值, 如骑行者和步行者类别要求正例的 IoU 超过 50%、车类别要求正例的 IoU 超过 70% 等, 据此来计算相应的精确率−召回率曲线, 然后通过对这些曲线进行插值, 计算平均精度。平均方向相似性 (AOS) 计算公式为

$$S_{\mathrm{AO}}=\frac{1}{11}\sum_{r\in\{0,0.1,\cdots,1\}}\max_{\tilde{R}:\tilde{R}\geq R}s(\tilde{R})$$

式中 R 表示召回率。在因变量 R 下, 方向相似性 $s\in[0,1]$被认定是所有预测样本与 ground truth 余弦距离的归一化, 其计算公式为

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

其中: $D(R)$表示在召回率 R 下所有预测为正样本的集合, $\Delta_{\theta}^{(i)}$表示目标 i 的预测角度与真值之间的差。为了惩罚多个检出匹配到同一个真值, 如果检出目标 i 已经匹配到真值, 则设定$\delta_i=1$, 否则 $\delta_i=0$。表 5 给出了不同算法在 KITTI 验证集上的检测性能。

表 5　不同算法在 KITTI 验证集上的检测性能

Table 5　Detection performance of different algorithms on KITTI validation set　　%

类型	方法	AP_{BEV}(IoU=0.5)			AP_{BEV}(IoU=0.7)		
		简单	中等	困难	简单	中等	困难
Mono	MF3D[37]	55.02	36.73	31.27	22.03	13.63	11.60
	M3DSSD[15]	—	—	—	34.51	26.20	23.40
	MonoPair[30]	61.06	47.63	41.92	24.12	18.17	15.76
	MonoDTR[34]	69.04	52.47	45.90	33.33	25.35	21.68
Stereo	3DOP[19]	55.04	41.25	34.55	12.63	9.49	7.59
	TLNet[16]	62.46	45.99	41.92	29.22	21.88	18.83
	Pseudo-LiDAR[45]	89.80	77.60	68.20	72.80	51.80	44.00
	CG-Stereo[49]	97.04	88.58	80.34	87.31	68.69	65.80
View-based	VeloFCN[52]	79.68	63.82	62.80	40.14	32.08	30.47
Voxel-based	VoxelNet[52]	—	—	—	89.60	84.81	78.57
	SECOND[59]	—	—	—	89.96	87.07	79.66
	Voxel-FPN[62]	—	—	—	90.20	87.92	86.27
	CIA-SSD[65]	—	—	—	—	—	—
Point-based	STD[72]	—	—	—	90.50	88.50	88.10
	PC-RGNN[86]	—	—	—	—	—	—
多传感器	MV3D[94]	—	—	—	86.55	78.10	76.67
	AVOD[95]	88.53	83.79	77.90	—	—	—
	F-PointNets[96]	—	—	—	88.16	84.02	76.44
	F-ConvNet[97]	—	—	—	90.23	88.79	86.84
	SFD[100]	—	—	—	—	—	—

续表 5

类型	方法	AP_{3D}(IoU=0.5)/%			AP_{3D}(IoU=0.7)/%		
		简单	中等	困难	简单	中等	困难
Mono	MF3D[37]	47.88	29.48	26.44	10.53	5.96	5.39
	M3DSSD[15]	—	—	—	27.77	21.67	18.28
	MonoPair[30]	55.38	42.39	37.99	16.28	12.30	10.42
	MonoDTR[34]	64.03	47.32	42.20	24.52	18.57	15.51
Stereo	3DOP[19]	46.04	34.63	30.09	6.55	5.07	4.10
	TLNet[16]	59.51	43.71	37.99	18.15	14.26	13.72
	Pseudo-LiDAR[45]	89.50	75.50	66.30	59.40	39.80	33.50
	CG-Stereo[49]	90.58	87.01	79.76	76.17	57.82	54.63
View-based	VeloFCN[52]	67.92	57.57	52.56	15.20	13.66	15.98
Voxel-based	VoxelNet[52]	—	—	—	81.97	65.46	62.85
	SECOND[59]	—	—	—	87.43	76.48	69.10
	Voxel-FPN[62]	—	—	—	88.27	77.86	75.84
	CIA-SSD[65]	—	—	—	90.04	79.81	78.80
Point-based	STD[72]	—	—	—	89.70	79.80	79.30
	PC-RGNN[86]	—	—	—	90.94	81.43	80.45
多传感器	MV3D[94]	—	—	—	71.29	62.68	56.56
	AVOD[95]	81.94	71.88	66.38	—	—	—
	F-PointNets[96]	—	—	—	83.76	70.92	63.65
	F-ConvNet[97]	—	—	—	89.02	78.80	77.09
	SFD[100]	—	—	—	95.47	88.56	85.74

4.2.2　nuScenes 数据集性能指标评价

在使用 nuScenes 数据集进行 3D 目标检测时，使用 AP 作为评估指标，但与 KITTI 不同的是，AP 的阈值匹配不使用 IoU 来计算，而是用在鸟瞰图投影的中心距离 d 来计算。这样做分离了物体的尺寸和方向对 AP 计算的影响。在计算 AP 时，将低于 0.1 的召回率和精确率区间用 0 代替，可以降低这些部分噪声的影响。d 的取值范围为{0.5, 1, 2, 4}m，则可得到平均精度均值（mean average precision, mAP），计算公式为

$$P_{\mathrm{mA}}=\frac{1}{|C||D|}\sum_{c\in C}\sum_{d\in D}P_{c,d}$$

式中：C 表示数据集中所有的目标类别，D 表示用于匹配 正确检测到的目标中心距离阈值集合。

由于 mAP 仅考虑了包围框的位置信息，不包括尺寸和方向，因此 nuScenes 还设计了一系列的误差度量，包括：1) 平均平移误差 (average translation error, ATE)，是 2D 欧几里德中心距离，单位为 m；2) 平均尺度误差 (average scale error, ASE)，是 $1-I_{\mathrm{oU}}$，其中 I_{oU} 是角度对齐后的 3D 交并比；3) 平均角度误差 (average orientation error, AOE)，是预测值和真实值之间最小的偏航角差，所有的类别角度偏差都在 360°内，除了障碍物这个类别的角度偏差在 180°内；4) 平均速度误差 (average velocity error, AVE)，是 2D 速度差的 L_2 范数；5) 平均属性误差 (average attribute error, AAE)，定义为 $1-A_{\mathrm{cc}}$，其中 A_{cc} 为分类别分类精度。

对于每个正确检测度量，计算所有类的平均正确检测到目标的度量值 P_{mT}。计算公式为

$$P_{\mathrm{mT}}=\frac{1}{|C|}\sum_{c\in C}P_c$$

除了 mAP 外，nuScenes 还提出了一个指标 NDS，该指标使用 P_{mT} 计算，计算公式为

$$S_{\mathrm{ND}}=\frac{1}{10}\left[5P_{\mathrm{mA}}+\sum_{P_{\mathrm{mT}}\in\mathrm{TP}}(1-\min(1,P_{\mathrm{mT}}))\right]$$

NDS 一半基于检测性能 (mAP)，而另一半基于检测性能根据位置、大小、方向、属性和速度度量的检测质量 (ATE、ASE、AOE、AVE、AAE)。表 6 给出了不同算法在 nuScenes 测试集上的检测性能。

表 6　不同算法在 nuScenes 测试集上的检测性能

Table 6　Detection performance of different algorithms on the nuScenes test set　%

类型	方法	mAP	NDS
Mono	DistillBEV[36]	52.5	61.2
Point-based	SASA[75]	45.0	61.0
Voxel-based	Voxelnext[64]	56.5	64.5
多传感器	PointPainting[97]	46.4	58.1
	SparseFusion[102]	73.8	72.0
	BEVFusion[105]	69.2	71.8
	CRN[118]	57.5	62.4

4.2.3　Waymo Open 数据集评价标准

在 3D 目标检测任务中, Waymo Open 数据集的检测基准依然使用传统的 AP, 不过增加了一个包含方向角信息的 APH, 多考察了角度检测效果, 计算公式为

$$P_{\mathrm{AH}}=100\int_0^1 \max\{h(R')|R'\geqslant R\}\mathrm{d}R$$

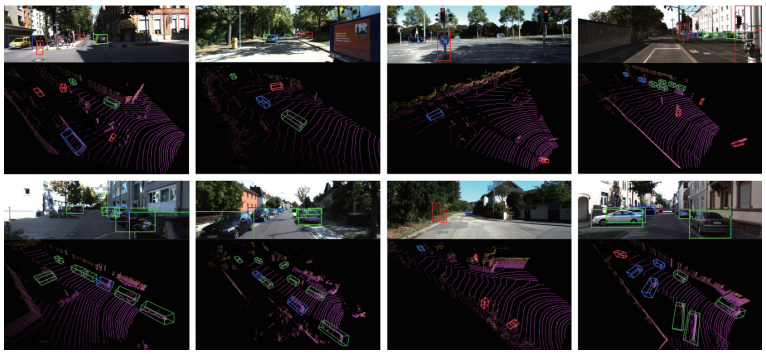
式中: R 是召回率; $h(R')$类似 P-R 曲线的 $p(R)$, 但在每个点都增加了一个权重, 即为 $\dfrac{\min(|\hat{\theta}-\theta|,2\pi-|\hat{\theta}-\theta|)}{\pi}$; $\hat{\theta}$和θ分别表示预测的方向角和真值方向角。表 7 给出一些代表性算法在 Waymo Open 数据集上的 3D 目标检测性能。

为更直观地展示代表性方法的性能差异, 图 8 给出了几种方法在 KITTI 与 nuScenes 数据集上的定性对比结果, 以补充表格中的定量指标。

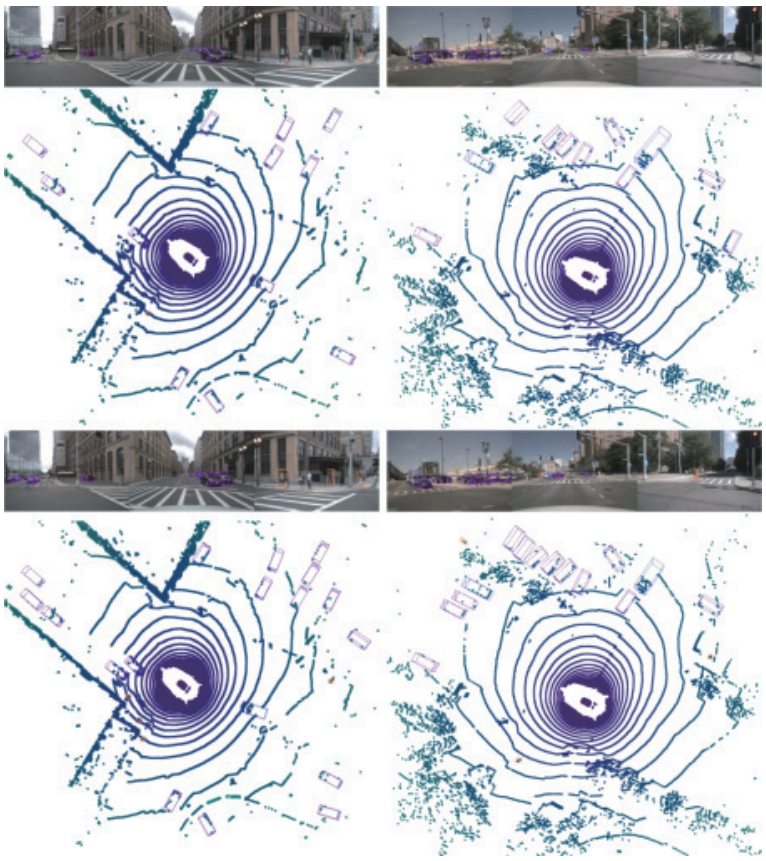
表 7　不同算法在 Waymo Open 数据集上的检测性能

Table 7　Detection performance of different algorithms on Waymo Open Datast validation set　%

类型	方法	LEVEL_1 mAP/mAPH	LEVEL_2 mAP/mAPH
View- based	H3D R-CNN[56]	75.15/—	66.14/—
Point- based	IA-SSD[76]	70.53/69.67	61.55/60.80
Voxel-based	SECOND[59]	72.27/71.69	63.85/63.33
	Ada3D[69]	78.66/78.21	69.98/69.57
Point-Voxel-based	PV-RCNN[89]	77.51/76.89	68.98/68.41
多传感器	DeepFusion[109]	81.89/80.48	76.91/75.54



(a) clos[134] 在 KITTI 数据集上与 SECOND[59] 比较



(b) VirConv-T[135] 在 nuScenes 数据集中与 PVConvNet[109] 比较

图 8　典型算法在 KITTI 和 nuScenes 数据集上的定性对比

Fig. 8　Qualitative comparison of representative algorithms on KITTI and nuScenes datasets

如图 8(a) 所示, clos[134] 与 SECOND[59] 在 KITTI 数据集上的对比表明, 前者在小目标和远距离目标的检测上表现出更明显的优势, 而 SECOND 在中距离和大目标的检测上更为稳定, 二者各有优势; 如图 8(b) 所示, VirConv-T[135] 与 PVConvNet[109] 在 nuScenes 数据集上的对比显示, 前者在复杂场景中对目标结构的保持能力更强, 而 PVConvNet[109] 在较开阔区域的目标定位精度上表现更好, 体现了不同方法在不同场景下的适用性。

5　展望

随着无人驾驶技术逐步迈向实际应用, 高精度、高效率的目标检测已成为自动驾驶感知系统的核心能力之一。然而, 自动驾驶场景下的 3D 目标检测任务面临诸多挑战: 首先, 传感器数据易受外部环境 (如雨雪天气、光照变化等) 干扰; 其次, 场景视野的动态变化以及目标尺寸过小、目标间相互遮挡等问题进一步增加了检测难度。这些挑战使得提升复杂环境下的目标检测精度、增强算法的鲁棒性, 以及降低外部环境对检测性能的影响, 成为当前领域亟待解决的关键技术难题。基于对现有算法研究现状的分析, 未来研究可能聚焦于以下方向:

1) 基于图像的深度估计技术

鉴于相机相较于激光雷达的成本优势, 基于图像的 3D 目标检测方法仍然有广阔的发展前景。立体感知与深度估计技术作为该领域的核心技术, 能够更精确地重建场景中物体的 3D 结构, 为自动驾驶系统提供更丰富的环境信息。结合深度学习与立体视觉, 通过深度神经网络学习立体匹配的特征表示并优化算法, 已成为提升深度估计精度与鲁棒性的有效途径。这种方法不仅能够提取图像中的关键特征, 还能自适应地优化匹配策略, 从而适应复杂多变的驾驶环境。

2) 多模态数据融合

不同传感器的数据在格式、分辨率、坐标系等方面存在异构性, 有效的多传感器融合需要解决异构数据的校准、配准与融合问题, 以确保数据的一致性与准确性。在融合过程中, 传感器数据的权重分配、可信度评估以及融合策略的选择对系统性能具有重要影响。因此, 开发高效的多模态融合算法并解决计算复杂性问题, 是未来研究的重要方向之一。

3) 模型压缩与优化

自动驾驶技术的快速发展对计算资源与存储空间的需求日益增加。为实现自动驾驶系统在更多车型上的部署, 模型压缩与优化技术成为关键。深度学习提供了多种模型压缩与优化方法, 如剪枝、量化、知识蒸馏等。这些技术能够在保持模型性能的同时, 显著降低对硬件资源的需求, 从而降低成本并提升其普及率。

4) 弱监督学习与无监督学习

标记大规模点云数据是一项昂贵且耗时的任务, 当前大多数 3D 目标检测方法依赖于大量标注数据, 限制了其在不同环境中的适应性。弱监督学习方法如弱标签、部分标签或辅助任务) 可以在有限标注数据下辅助点云特征学习。无监督学习方法则能够从未标注的点云数据中自动学习特征表示与数据分布。结合强化学习与无监督学习, 智能体可以通过与环境的交互学习目标检测策略, 逐步优化并适应复杂驾驶场景中的检测需求。

5) 数据集的构建

数据集的质量与多样性对 3D 目标检测算法的训练与评估具有关键影响。高质量数据集应提供准确的标注信息, 涵盖车辆、行人、自行车等多类目标, 并包括多样化的场景、天气和交通状况, 以增强模型的泛化能力与鲁棒性。同时, 边缘场景与极端样本的引入有助于测试系统在复杂环境下的稳定性。

值得关注的是, 近年来部分研究开始探索构建虚拟数据集, 以降低数据采集和标注成本。此类数据集借助 CARLA、Unreal Engine 等仿真平台生成, 具备高可控性和可扩展性, 能够覆盖多种复杂或危险场景, 常用于预训练或数据增强。然而, 由于合成数据与真实场景在传感器特性、纹理细节、噪声分布等方面存在差异, 模型在虚拟数据上训练后直接迁移至真实环境时往往会出现性能下降, 即所谓“域间差距”问题。缩小合成与真实数据之间的域差距, 同时将虚拟数据更有效地引入训练和评估流程, 是未来研究的重要方向。

6) 端到端多任务联合学习

未来的 3D 目标检测研究将更加重视多任务协同学习。通过端到端的联合框架, 同时整合目标检测、目标跟踪、语义分割以及行为预测等任务, 不仅可以共享特征表示, 减少冗余计算, 还能增强不同任务间的信息交互与融合, 提升整体感知的准确性和鲁棒性。这种多任务学习方式有助于构建更全面的环境理解, 为自动驾驶系统提供更丰富的语义信息和动态决策支持。此外, 多任务联合学习还能促进模型泛化能力的提升, 适应更加复杂和多变的实际应用场景, 推动智能驾驶

技术向更高层次发展。

7) 实时自适应系统与边缘智能

随着自动驾驶对实时性和计算资源效率的严格要求, 设计具备在线自适应能力的轻量级模型成为未来研究重点。此类模型能够根据环境变化和任务需求, 动态调整参数和计算流程, 实现高效且稳定的性能表现。同时, 结合边缘计算与云端智能的协同架构, 可以有效分配计算任务, 降低终端设备的负载, 减少延迟, 提升系统的实时响应能力。边缘智能不仅能够保障数据的隐私与安全, 还能在网络条件不佳时保持感知系统的稳定运行。未来, 实时自适应与边缘智能的深度融合将成为自动驾驶感知系统的重要发展方向, 为实现大规模商业化应用奠定坚实基础。

通过在自动驾驶领域的持续研究与技术突破, 未来的自动驾驶系统将能够更好地适应复杂多变的道路环境, 显著提升目标检测的精度、速度与鲁棒性, 从而推动无人驾驶技术的广泛应用与普及。

6　结束语

自动驾驶技术的迅猛发展对 3D 目标检测算法的准确性和实时性提出了更高的要求。本文深入探讨了基于深度学习的 3D 目标检测算法的最新进展, 并根据所使用的传感器类型将其划分为三大类: 基于视觉、基于激光点云和基于多传感器融合。基于视觉的方法通过融合深度估计、几何特征提取等手段, 间接获取 3D 信息, 虽然计算高效, 但精度存在一定的限制。激光点云方法则直接处理点云数据, 能够提供高精度的 3D 空间信息, 但其处理复杂度较高。多传感器融合策略则通过整合不同传感器的优势, 旨在精度与效率之间找到最佳平衡点, 展现了巨大的潜力。本文还介绍了常用的自动驾驶数据集以及算法在这些数据集上的评价指标, 并对一些代表性算法在不同数据集上的性能进行了比较。通过分析当前 3D 目标检测技术面临的问题和挑战, 对未来进一步的研究工作进行了展望。通过本文的总结和分析, 期望能为 3D 目标检测的研究提供一定的参考。

参考文献:

[1] 百度地图, 北京交通发展研究院, 清华大学数据科学研究院交通大数据研究中心, 等. 2024 年度中国城市交通报告[R]. 北京: 百度地图, 2024: 7−8.
BAIDU Maps, Beijing Traffic Development Research Institute, Tsinghua University Data Science Research Institute Traffic Big Data Center, et al. 2024 Annual China Urban Traffic Report[R]. Beijing: Baidu Maps, 2024: 7−8.

[2] 段伟. 汽车自动驾驶技术简述[J]. 中国自动识别技术, 2024(2): 66−68.
DUAN Wei. A brief introduction to automobile autonomous driving technology[J]. China automatic identification technology, 2024(2): 66−68.

[3] 郭毅锋, 吴帝浩, 魏青民. 基于深度学习的点云三维目标检测方法综述[J]. 计算机应用研究, 2023, 40(1): 20−27.
GUO Yifeng, WU Dihao, WEI Qingmin. Overview of single-sensor and multi-sensor point cloud 3D target detection methods[J]. Application research of computers, 2023, 40(1): 20−27.

[4] 李佳男, 王泽, 许廷发. 基于点云数据的三维目标检测技术研究进展[J]. 光学学报, 2023, 43(15): 296−312.
LI Jianan, WANG Ze, XU Tingfa. Three-Dimensional Object Detection Technology Based on Point Cloud Data[J]. Acta optica sinica, 2023, 43(15): 296−312.

[5] 曹家乐, 李亚利, 孙汉卿, 等. 基于深度学习的视觉目标检测技术综述[J]. 中国图象图形学报, 2022, 27(6): 1697−1722.
CAO Jiale, LI Yali, SUN Hanqin, et al. A survey on deep learning based visual object detection[J]. Journal of image and graphics, 2022, 27(6): 1697−1722.

[6] 贾明达, 杨金明, 孟维亮, 等. 融合点云与图像的环境目标检测研究进展[J]. 中国图象图形学报, 2024, 29(6): 1765−1784.
JIA Minda, YANG Jinming, MENG Weiliang, et al. 2024. Survey on the fusion of point clouds and images for environmental object detection[J]. Journal of image and graphics, 2024, 29(6): 1765−1784.

[7] CUI Yaodong, CHEN Ren, CHU Wenbo, et al. Deep learning for image and point cloud fusion in autonomous driving: a review[J]. IEEE Transactions on Intelligent Transportation Systems, 2021, 23(2): 722−739.

[8] 陈慧娴, 吴一全, 张耀. 基于深度学习的三维点云分析方法研究进展[J]. 仪器仪表学报, 2023, 44(11): 130−158.
CHEN Huixian, WU Yiquan, ZHANG Yao. Research progress on 3D point cloud analysis methods based on deep learning[J]. Chinese journal of scientific instrument, 2023, 44(11): 130−158.

[9] 周燕, 许业文, 蒲磊, 等. 自动驾驶场景下的图像三维目标检测研究进展[J]. 计算机科学, 2024, 1−18.
ZHOU Yan, XU Yewen, PU Lei, et al. Research progress on image 3D target detection in autonomous driv-

ing scenarios[J]. Computer science, 2024, 1−18.

[10] DROBNITZKY M, FRIEDERICH J, EGGER B, et al. Survey and systematization of 3D object detection models and methods[J]. The visual computer, 2024, 40(3): 1867−1913.

[11] 任柯燕, 谷美颖, 袁正谦, 等. 自动驾驶 3D 目标检测研究综述[J]. 控制与决策, 2023, 38(4): 865−889.
REN Keyan, GU Meiyin, YUAN Zhengqian, et al. Review of research on 3D target detection in autonomous driving[J]. Control and decision, 2023, 38(4): 865−889.

[12] 张新宇, 徐子贤, 闫冬梅, 等. 基于深度学习的 3D 目标检测算法综述[J]. 控制工程, 2024, 31(3): 526−534.
ZHANG Xinyu, XU Zixian, YAN Dongmei, et al. Review of 3D object detection algorithms based on deep learning[J]. Control engineering, 2024, 31(3): 526−534.

[13] MOUSAVIAN A, ANGUELOV D, FLYNN J, et al. 3D bounding box estimation using deep learning and geometry[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017: 7074−7082.

[14] LI Buyu, OUYANG Wanli, Lu Sheng, et al. GS3D: an efficient 3D object detection framework for autonomous driving[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 1019−1028.

[15] LUO Shujie, DAI Hang, SHAO Ling, et al. M3DSSD: monocular 3D single stage object detector[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021: 6145−6154.

[16] QIN Zengyi, WANG Jinglu, LU Yan. Triangulation learning network: from monocular to stereo 3D object detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 7615−7623.

[17] GUO Xiaoyang, SHI Shaoshuai, WANG Xiaogang, et al. LIGA-Stereo: learning LiDAR geometry aware representations for stereo-based 3D detector[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021: 3153−3163.

[18] 迟旭然, 裴伟, 朱永英, 等. Fast Stereo-RCNN 三维目标检测算法[J]. 小型微型计算机系统, 2022, 43(10): 2157−2161.
CHI Xuran, PEI Wei, ZHU Yongying, et al. Fast Stereo-RCNN 3D target detection algorithm[J]. Mini-micro computer systems, 2022, 43(10): 2157−2161.

[19] HEN Xiaozhi, KUNDU K, ZHU Yukun, et al. 3D object proposals using stereo imagery for accurate object class detection[J]. IEEE transactions on pattern analysis and machine intelligence, 2017, 40(5): 1259−1272.

[20] GIRSHICK R. Fast R-CNN[C]//Proceedings of the IEEE International Conference on Computer Vision. Santiago: IEEE, 2015: 1440−1448.

[21] CHABOT F, CHAOUCH M, RABARISOA J, et al. Deep MANTA: a coarse-to-fine many-task network for joint 2D and 3D vehicle analysis from monocular image [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017: 2040−2049.

[22] KUNDU A, LI Yin, REHG J M. 3D-RCNN: instance-level 3D object reconstruction via render-and-compare[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 3559−3568.

[23] KREISS S, BERTONI L, ALAHI A. PifPaf: composite fields for human pose estimation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 11977−11986.

[24] BERTONI L, KREISS S, ALAHI A. MonoLoco: monocular 3D pedestrian localization and uncertainty estimation[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Long Beach: IEEE, 2019: 6861−6871.

[25] LI Peixuan, ZHAO Huaici, LIU Pengfei, et al. RTM3D: real-time monocular 3D detection from object keypoints for autonomous driving[C]//European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2020: 644−660.

[26] CAI Yingjie, LI Buyu, JIAO Zeyu, et al. Monocular 3D object detection with decoupled structured polygon estimation and height-guided depth estimation[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2020: 10478−10485.

[27] LIU Zongdai, ZHOU Dingfu, LU Feixiang, et al. AutoShape: real-time shape-aware monocular 3D object detection[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021: 15641−15650.

[28] SHUAI Qingyao, ZHANG Chi, YANG Kaizhi, et al. DPF-Net: combining explicit shape priors in deformable primitive field for unsupervised structural reconstruction of 3D objects[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2023: 14321−14329.

[29] DUAN Fan, YU Jiahao, CHEN Li. T-CorresNet: template guided 3D point cloud completion with correspondence pooling query generation strategy[C]//European Conference on Computer Vision. Milan: Springer Nature Switzerland, 2024: 90−106.

[30] CHEN Yongjian, TAI Lei, SUN Kai, et al. MonoPair: monocular 3D object detection using pairwise spatial relationships[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Virtual Conference: IEEE, 2020: 12093−12102.

[31] MA Xinzhu, ZHANG Yinmin, XU Dan, et al. Delving into localization errors for monocular 3D object detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021: 4721−4730.

[32] ZHANG Yunpeng, LU Jiwen, ZHOU Jie, et al. Objects are different: flexible monocular 3D object detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021: 3289−3298.

[33] 汪萌, 诸兵. 不确定性建模在 2D 和 3D 目标检测中的应用[J]. 系统工程与电子技术, 2023, 45(8): 2370−2376.
WANG Meng, ZHU Bing. Application of uncertainty modeling in 2D and 3D target detection[J]. Systems engineering and electronics, 2023, 45(8): 2370−2376.

[34] HUANG K C, WU T H, SU H T, et al. MonoDTR: monocular 3D object detection with depth-aware Transformer[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022: 4012−4021.

[35] LI Zhiqi, WANG Wenhai, LI Hongyang, et al. BEVFormer: learning bird's-eye-view representation from multi-camera images via spatiotemporal Transformers [EB/OL]. (2022−03−31)[2025−04−24]. https://arxiv.org/abs/2203.17270.

[36] WANG Zeyu, LI Dingwen, LUO Chenxu, et al. DistillBEV: boosting multi-camera 3D object detection with cross-modal knowledge distillation[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2023: 8637−8646.

[37] XU Bin, CHEN Zhenzhong. Multi-level fusion based 3D object detection from monocular images[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 2345−2353.

[38] GODARD C, MAC AODHA O, BROSTOW G J. Unsupervised monocular depth estimation with left-right consistency[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017: 270−279.

[39] DING Mingyu, HUO Yuqi, YI Hongwei, et al. Learning depth-guided convolutions for monocular 3D object detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Virtual Conference: IEEE, 2020: 1000−1001.

[40] PENG Liang, WU Xiaopei, YANG Zheng, et al. DID-M3D: decoupling instance depth for monocular 3D object detection[C]//European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022: 71−88.

[41] RODDICK T, KENDALL A, CIPOLLA R. Orthographic feature transform for monocular 3D object detection [EB/OL]. (2018−11−20)[2025−04−24]. https://arxiv.org/abs/1811.08188.

[42] LIU Yingfei, WANG Tiancai, ZHANG Xiangyu, et al. PETR: position embedding transformation for multi-view 3D object detection[C]//European conference on computer vision. Cham: Springer Nature Switzerland, 2022: 531−548.

[43] ŽBONTAR J, LECUN Y. Stereo matching by training a convolutional neural network to compare image patches[J]. Journal of machine learning research, 2016, 17(65): 1−32.

[44] KENDALL A, MARTIROSYAN H, DASGUPTA S, et al. End-to-end learning of geometry and context for deep stereo regression[C]//Proceedings of the IEEE International Conference on Computer Vision. Venice: IEEE, 2017: 66−75.

[45] WANG Yan, CHAO Weilun, GARG D, et al. Pseudo-LiDAR from visual depth estimation: bridging the gap in 3D object detection for autonomous driving[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 8445−8453.

[46] FU Huan, GONG Mingming, WANG Chaohui, et al. Deep ordinal regression network for monocular depth estimation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 2002−2011.

[47] CHANG Jiaren, CHEN Yongshen. Pyramid stereo matching network[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 5410−5418.

[48] WANG Xinlong, YIN Wei, KONG Tao, et al. Task-aware monocular depth estimation for 3D object detec-

tion[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2020: 12257−12264.

[49] LI Chengyao, KU J, WASLANDER S L. Confidence guided stereo 3D object detection with split depth estimation[C]//2020 IEEE/RSJ International Conference on Intelligent Robots and Systems(IROS). Las Vegas: IEEE, 2020: 5776−5783.

[50] HOSSAIN S, LIN Xianke. Efficient stereo depth estimation for pseudo-LiDAR: a self-supervised approach based on multi-input ResNet encoder[J]. Sensors, 2023, 23(3): 1650.

[51] OH C, JANG Y, SHIM D, et al. Automatic pseudo-LiDAR annotation: generation of training data for 3D object detection networks[J]. IEEE access, 2024.

[52] LI Bo, ZHANG Tianlei, XIA Tian. Vehicle detection from 3D LiDAR using fully convolutional network [EB/OL]. (2016−08−29)[2025−04−24]. https://arxiv.org/abs/1608.07916.

[53] YANG Bin, LUO Wenjie, URTASUN R. PIXOR: real-time 3D object detection from point clouds[C]//Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 7652−7660.

[54] BELTRÁN J, GUINDEL C, MORENO F M, et al. BirdNet: a 3D object detection framework from LiDAR information[C]//2018 21st International Conference on Intelligent Transportation Systems(ITSC). Miami: IEEE, 2018: 3517−3523.

[55] MEYER G P, LADDHA A, KEE E, et al. LaserNet: an efficient probabilistic 3D object detector for autonomous driving[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 12677−12686.

[56] DENG Jiajun, ZHOU Wengang, ZHANG Yanyong, et al. From multi-view to hollow-3D: hallucinated hollow-3D R-CNN for 3D object detection[J]. IEEE transactions on circuits and systems for video technology, 2021, 31(12): 4722−4734.

[57] SUN Pei, WANG Weiyue, CHAI Yuning, et al. RSN: range sparse net for efficient, accurate LiDAR 3D object detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021: 5725−5734.

[58] ZHOU Yin, TUZEL O. VoxelNet: end-to-end learning for point cloud based 3D object detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 4490−4499.

[59] YAN Yan, MAO Yuxing, LI Bo. Second: sparsely embedded convolutional detection[J]. Sensors, 2018, 18(10): 3337.

[60] LANG A H, VORA S, CAESAR H, et al. Pointpillars: fast encoders for object detection from point clouds[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 12697−12705.

[61] LI Haisheng, LU Yanling. 3D object detection based on point cloud in automatic driving scene[J]. Multimedia tools and applications, 2024, 83(5): 13029−13044.

[62] Wang Bei, An Jianping, Cao Jiayan, et al. Voxel-FPN: multi-scale voxel feature aggregation for 3D object detection from LIDAR point clouds[J]. Sensors, 2020, 20(3): 704.

[63] LIU Zhe, ZHAO Xin, HUANG Tengteng, et al. TANet: robust 3D object detection from point clouds with triple attention[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2020: 11677−11684.

[64] CHEN Yukang, LIU Jianhui, ZHANG Xiangyu, et al. VoxelNeXt: fully sparse voxelnet for 3D object detection and tracking[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023: 21674−21683.

[65] ZHENG Wu, TANG Weiliang, CHEN Sijin, et al. CIA-SSD: confident IOU-aware single-stage object detector from point cloud[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2021: 3555−3562.

[66] FAN Lue, PANG Ziqi, ZHANG Tianyuan, et al. Embracing single stride 3D object detector with sparse Transformer[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022: 8458−8468.

[67] HE Chenhang, LI Ruihuang, LI Shuai, et al. Voxel set Transformer: a set-to-set approach to 3D object detection from point clouds[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022: 8417−8427.

[68] HE Chenhang, ZENG Hui, HUANG Jianqiang, et al. Structure aware single-stage 3D object detection from point cloud[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Virtual Conference: IEEE, 2020: 11873−11882.

[69] ZHAO Tianchen, NING Xuefei, HONG Ke, et al. Ada3D: exploiting the spatial redundancy with adaptive

inference for efficient 3D object detection[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2023: 17728−17738.

[70] QI C R, SU Hao, MO Kaichun, et al. PointNet: deep learning on point sets for 3D classification and segmentation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Montreal: IEEE, 2017: 652−660.

[71] QI C R, YI Li, SU Hao, et al. PointNet++: deep hierarchical feature learning on point sets in a metric space [J]. Advances in neural information processing systems, 2017, 30.

[72] YANG Zetong, SUN Yanan, LIU Shu, et al. STD: sparse-to-dense 3D object detector for point cloud[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019: 1951−1960.

[73] 陈熙源, 戈明明, 姚志婷, 等. 雨雪天气下的激光雷达滤波算法研究[J]. 仪器仪表学报, 2023, 44(7): 172−181.
CHEN Xiyuan, GE Mingming, YAO Zhiting, et al. Filtering algorithm of LiDAR in rainy and snowy weather [J]. Chinese journal of scientific instrument, 2023, 44(7): 172−181.

[74] TAO Manli, ZHAO Chaoyang, TANG Ming, et al. Objformer: boosting 3D object detection via instance-wise interaction[J]. Pattern Recognition, 2024, 146: 110061.

[75] CHEN Chen, CHEN Zhe, ZHANG Jing, et al. SASA: semantics-augmented set abstraction for point-based 3D object detection[J]. Proceedings of the AAAI conference on artificial intelligence, 2022, 36(1): 221−229.

[76] ZHANG Yifan, HU Qingyong, XU Guoquan, et al. Not all points are equal: learning highly efficient point-based detectors for 3D LiDAR point clouds[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022: 18953−18962.

[77] 王理嘉, 于欢, 刘守印. 动态环境中多帧点云融合算法及三维目标检测算法研究[J]. 计算机应用研究, 2023, 40(3): 909−913.
WANG Lijia, YU Huan, LIU Shouyin. Research on multi-frame point cloud fusion and 3D target detection algorithms in dynamic environments[J]. Application research of computers, 2023, 40(3): 909−913.

[78] ZHANG Gang, CHEN Junnan, GAO Guohuan, et al. HEDNet: a hierarchical encoder-decoder network for 3D object detection in point clouds[J]. Advances in neural information processing systems, 2024, 36.

[79] LI Yangyan, BU Rui, SUN Mingchao, et al. PointCNN: convolution on X-transformed points[J]. Advances in neural information processing systems, 2018, 31.

[80] YIN Tianwei, ZHOU Xingyi, KRAHENBUHL P. Center-based 3D object detection and tracking[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021: 11784−11793.

[81] 涂新奎, 郑少武, 于善虎, 等. 基于对称形状生成的三维目标检测网络[J]. 仪器仪表学报, 2023, 44(6): 252−263.
TU Xinkui, ZHENG Shaowu, YU Shanhu, et al. 3D target detection network based on symmetric shape generation[J]. Chinese journal of scientific instrument, 2023, 44(6): 252−263.

[82] 陶乐, 王海, 蔡英凤, 等. 面向自动驾驶场景的多目标点云检测算法[J]. 汽车工程, 2024, 46(7): 1208−1218, 1238.
TAO Le, WANG Hai, CAI Yingfeng, et al. Multi-object point cloud detection algorithm for autonomous driving scenarios[J]. Automotive engineering, 2024, 46(7): 1208−1218, 1238.

[83] 周昊, 齐洪钢, 邓永强, 等. 融合点云深度信息的 3D 目标检测与分类[J]. 中国图象图形学报, 2024, 29(8): 2399−2412.
ZHOU Hao, QI Honggang, DENG Yongqiang, et al. 3D target detection and classification using fused point cloud depth information[J]. Journal of image and graphics, 2024, 29(8): 2399−2412.

[84] SHI Weijing, RAJKUMAR R. Point-GNN: graph neural network for 3D object detection in a point cloud[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Virtual Conference: IEEE, 2020: 1711−1719.

[85] NAJIBI M, LAI Guangda, KUNDU A, et al. DOPS: learning to detect 3D objects and predict their 3D shapes [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Virtual Conference: IEEE, 2020: 11913−11922.

[86] ZHANG Yanan, HUANG Di, WANG Yunhong. PC-RGNN: point cloud completion and graph neural network for 3D object detection[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2021: 3430−3437.

[87] LIU Zhijian, TANG Haotian, LIN Yujun, et al. Point-voxel CNN for efficient 3D deep learning[J]. Advances in neural information processing systems, 2019, 32.

[88] NOH J, LEE S, HAM B. HVPR: hybrid voxel-point representation for single-stage 3D object detection[C]//Pro-

ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021: 14605−14614.

[89] SHI Shaoshuai, GUO Chaoxu, JIANG Li, et al. PV-RCNN: point-voxel feature set abstraction for 3D object detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Virtual Conference: IEEE, 2020: 10529−10538.

[90] WU Peng, GU Lipeng, YAN Xuefeng, et al. PV-RCNN++: semantical point-voxel feature interaction for 3D object detection[J]. The visual computer, 2023, 39(6): 2425−2440.

[91] ZHOU Wei, ZHANG Xiaodan, HAO Xin, et al. Multi point-voxel convolution(MPVConv) for deep learning on point clouds[J]. Computers & graphics, 2023, 112: 72−80.

[92] 李虎辰, 管海燕, 雷相达, 等. 基于点−体素一致性约束的城市激光雷达点云分类[J]. 中国激光, 2024, 51(13): 251−264.
LI Huchen, GUAN Haiyan, LEI Xiangda, et al. Urban LiDAR point cloud classification based on point-voxel consistency constraints[J]. Chinese journal of lasers, 2024, 51(13): 251−264.

[93] DENG Pengzhen, ZHOU Li, CHEN Jie. PVC-SSD: point-voxel dual-channel fusion with cascade point estimation for anchor-free single-stage 3D object detection[J]. IEEE sensors journal, 2024.

[94] CHEN Xiaozhi, MA Huimin, WAN Ji, et al. Multi-view 3D object detection network for autonomous driving[C]//Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017: 1907−1915.

[95] KU Jason, MOZIFIAN M, LEE J, et al. Joint 3D proposal generation and object detection from view aggregation[C]//2018 IEEE/RSJ International Conference on Intelligent Robots and Systems(IROS). Madrid: IEEE, 2018: 1−8.

[96] QI C R, LIU Wei, WU Chenxia, et al. Frustum pointnets for 3D object detection from RGB-D data[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 918−927.

[97] WANG Zhixin, JIA Kui. Frustum ConvNet: sliding frustums to aggregate local point-wise features for amodal 3D object detection[C]//2019 IEEE/RSJ International Conference on Intelligent Robots and Systems. Osaka: IEEE, 2019: 1742−1749.

[98] VORA S, LANG A H, HELOU B, et al. Pointpainting: sequential fusion for 3D object detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Virtual Conference: IEEE, 2020: 4604−4612.

[99] LIANG Ming, YANG Bin, CHEN Yun, et al. Multi-task multi-sensor fusion for 3D object detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 7345−7353.

[100] WU Xiaopei, PENG Liang, YANG Honghui, et al. Sparse fuse dense: towards high quality 3D detection with depth completion[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022: 5418−5427.

[101] 黄漫, 黄勃, 高永彬. 引入深度补全与实例分割的三维目标检测[J]. 传感器与微系统, 2021, 40(1): 129−132.
HUANG Man, HUANG Bo, GAO Yongbin. 3D target detection with depth completion and instance segmentation[J]. Sensors and microsystems, 2021, 40(1): 129−132.

[102] XIE Yichen, XU Chenfeng, RAKOTOSAONA M J, et al. Sparsefusion: Fusing multi-modal sparse representations for multi-sensor 3D object detection[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2023: 17591−17602.

[103] ZHANG Yanan, CHEN Jiaxin, HUANG Di. CAT-Det: contrastively augmented Transformer for multi-modal 3D object detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022: 908−917.

[104] GUNN J, LENYK Z, SHARMA A, et al. Lift-Attend-Splat: bird's-eye-view camera-LiDAR fusion using Transformers[EB/OL]. (2023−12−22)[2025−04−24]. https://arxiv.org/abs/2312.14919.

[105] LIU Zhijian, TANG Haotian, AMINI A, et al. BEVFusion: multi-task multi-sensor fusion with unified bird's-eye view representation[C]//2023 IEEE International Conference on Robotics and Automation(ICRA). London: IEEE, 2023: 2774−2781.

[106] WANG Ke, ZHOU Tianqiang, ZHANG Zhichuang, et al. PVF-DectNet: multi-modal 3D detection network based on perspective-voxel fusion[J]. Engineering applications of artificial intelligence, 2023, 120: 105951.

[107] LI Yingwei, YU A W, MENG Tianjian, et al. DeepFusion: Lidar-camera deep fusion for multi-modal 3D object detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022: 17182−17191.

[108] 周治国, 马文浩. 一种多层多模态融合 3D 目标检测方法[J]. 电子学报, 2024, 52(3): 696−708.
ZHOU Zhigui, MA Wenhao. A multi-layer multi-modal fusion 3D target detection method[J]. Acta electronica sinica, 2024, 52(3): 696−708.

[109] LIU Huaijin, DU Jixiang, ZHANG Yong, et al. PVConvNet: pixel-voxel sparse convolution for multimodal 3D object detection[J]. Pattern recognition, 2024, 149: 110284.

[110] 金宇锋, 陶重犇. 基于 Transformer 的融合信息增强 3D 目标检测算法[J]. 仪器仪表学报, 2023, 44(12): 297−306.
JIN Yufeng, TAO Zhongben. Fusion-enhanced 3D target detection algorithm based on Transformer[J]. Chinese journal of scientific instrument, 2023, 44(12): 297−306.

[111] XIA Chenxing, LI Xubing, GAO Xiuju, et al. PCDR-DFF: multi-modal 3D object detection based on point cloud diversity representation and dual feature fusion[J]. Neural computing and applications, 2024: 1−18.

[112] 王五岳, 徐召飞, 曲春燕, 等. 基于红外与激光雷达融合的鸟瞰图空间三维目标检测算法[J]. 光子学报, 2024, 53(1): 73−84.
WANG Wuyue, XU Zhaofei, QU Chunyan, et al. 3D target detection algorithm in bird's-eye view space based on infrared and LiDAR fusion[J]. Acta photonica sinica, 2024, 53(1): 73−84.

[113] 董钰婷, 官磊. 基于自适应加权融合激光雷达和相机的三维目标检测方法[J]. 计算机应用, 2024, 44(S1): 250−255.
DONG Yyuting, GUAN Lei. 3D target detection method based on adaptive weighted fusion of LiDAR and camera[J]. Computer applications, 2024, 44(S1): 250−255.

[114] 李文礼, 喻飞, 石晓辉, 等. BEV 特征下激光雷达和单目相机融合的目标检测算法研究[J]. 计算机工程与应用, 2024, 60(11): 182−193.
LI Wenli, YU Fei, SHI Xiaohui, et al. Target detection algorithm based on BEV features for LiDAR and monocular camera fusion[J]. Computer engineering and applications, 2024, 60(11): 182−193.

[115] NABATI R, QI Hairong. CenterFusion: center-based radar and camera fusion for 3D object detection[C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. Piscataway: IEEE, 2021: 1527−1536.

[116] BANSAL K, RUNGTA K, BHARADIA D. RadSegNet: a reliable approach to radar camera fusion[EB/OL]. (2022−08−08)[2025−04−24]. https://arxiv.org/abs/2208.03849.

[117] KIM Y, KIM S, CHOI J W, et al. CRAFT: camera-radar 3D object detection with spatio-contextual fusion Transformer[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2023: 1160−1168.

[118] KIM Y, SHIN J, KIM S, et al. CRN: camera radar net for accurate, robust, efficient 3D perception[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2023: 17615−17626.

[119] 车俐, 吕连辉, 蒋留兵. AF-CenterNet: 基于交叉注意力机制的毫米波雷达和相机融合的目标检测[J]. 计算机应用研究, 2024, 41(4): 1258−1263.
CHE Li, LYU Lianhui, JIANG Liubing. AF-CenterNet: Cross-attention mechanism-based millimeter-wave radar and camera fusion for target detection[J]. Application research of computers, 2024, 41(4): 1258−1263.

[120] LIU Xiang, LI Zhenglin, ZHOU Yang, et al. Camera–radar fusion with modality interaction and radar Gaussian expansion for 3D object detection[J]. Cyborg and bionic systems, 2024, 5: 0079.

[121] GEIGER A, LENZ P, URTASUN R. Are we ready for autonomous driving? the kitti vision benchmark suite[C]//2012 IEEE Conference on Computer Vision and Pattern Recognition. Providence: IEEE, 2012: 3354−3361.

[122] CAESAR H, BANKITI V, LANG A H, et al. nuScenes: a multimodal dataset for autonomous driving[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020: 11621−11631.

[123] SUN Pei, KRETZSCHMAR H, DOTIWALLA X, et al. Scalability in perception for autonomous driving: Waymo open dataset[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Virtual Conference: IEEE, 2020: 2446−2454.

[124] HUANG Xinyu, WANG Peng, CHENG Xinjing, et al. The apolloscape open dataset for autonomous driving and its application[J]. IEEE transactions on pattern analysis and machine intelligence, 2019, 42(10): 2702−2719.

[125] CHOI Y, KIM N, HWANG S, et al. KAIST multi-spectral day/night data set for autonomous and assisted driving[J]. IEEE Transactions on Intelligent Transportation Systems, 2018, 19(03): 934−948.

[126] HOUSTON J, ZUIDHOF G, BERGAMINI L, et al. One thousand and one hours: Self-driving motion prediction dataset[C]//Conference on Robot Learning. Palo Alto: PMLR, 2021: 409−418.

[127] YU Haibao, LUO Yizhen, SHU Mao, et al. DAIR-V2X: a large-scale dataset for vehicle-infrastructure cooperative 3D object detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022: 21361−21370.

[128] GÄHLERT N, JOURDAN N, CORDTS M, et al. Cityscapes 3D: Dataset and benchmark for 9 dof vehicle detection[EB/OL]. (2020−06−14)[2025−04−24]. https://arxiv.org/abs/2006.07864.

[129] WILSON B, QI W, AGARWAL T, et al. Argoverse 2: Next generation datasets for self-driving perception and forecasting[EB/OL]. (2023−01−02)[2025−04−24]. https://arxiv.org/abs/2301.00493.

[130] XIAO Pengchuan, SHAO Zhenlei, HAO S, et al. PandaSet: advanced sensor suite dataset for autonomous driving[C]//2021 IEEE International Intelligent Transportation Systems Conference(ITSC). Indianapolis: IEEE, 2021: 3095−3101.

[131] PATIL A, MALLA S, GANG H, et al. The H3D dataset for full-surround 3D multi-object detection and tracking in crowded urban scenes[C]//2019 International Conference on Robotics and Automation(ICRA). Montreal: IEEE, 2019: 9552−9557.

[132] CONG Peishan, ZHU Xinge, QIAO Feng, et al. STCrowd: a multimodal dataset for pedestrian perception in crowded scenes[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022: 19608−19617.

[133] XIAO Aoran, HUANG Jiaxing, GUAN Dayan, et al. Transfer learning from synthetic to real LiDAR point cloud for semantic segmentation[EB/OL]. (2021−07−12)[2025−04−24]. https://arxiv.org/abs/2107.05399.

[134] PANG Su, MORRIS D, RADHA H. CLOCs: Camera-LiDAR object candidates fusion for 3D object detection[C]//2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Las Vegas: IEEE, 2020: 10386−10393.

[135] WU Hai, WEN Chenglu, SHI Shaoshuai, et al. Virtual sparse convolution for multimodal 3D object detection[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. Vancouver: IEEE, 2023: 21653−21662.

作者简介：

吴一全，教授，主要研究方向为视觉检测与图像测量、视频处理与智能分析。主持国家自然科学基金等项目 48 项。发表学术论文 350 余篇。E-mail：nuaaimage@163.com。

蔡佳琦，硕士研究生，主要研究方向为计算机视觉、图像处理。E-mail：Caij-q@nuaa.edu.cn。

[责任编辑：刘冰洁]