Journal of Image and Signal Processing 图像与信号处理, 2025, 14(2), 173-184
Published Online April 2025 in Hans. https://www.hanspub.org/journal/jisp
https://doi.org/10.12677/jisp.2025.142017

Hans汉斯

基于深度学习的3D目标检测技术综述

邵　博[1]，徐仕琦[1]，张子琦[1]，杨琳倩[1]，高炜晴[2]，何嘉懿[1]，秦傲雪[1]，吴茜茵[1]

[1]贵州大学大数据与信息工程学院，贵州　贵阳
[2]贵州大学数学与统计学院，贵州　贵阳

收稿日期：2025年2月27日；录用日期：2025年3月18日；发布日期：2025年3月31日

摘　要

基于深度学习的3D目标检测技术在自动驾驶、机器人导航等众多前沿领域意义重大。然而，当前该技术仍面临诸多挑战，如处理大规模3D数据时计算复杂度高、小目标检测困难等，这些问题严重制约了其进一步发展与广泛应用。为突破困境，文章详细介绍了KITTI、NuScenes等常用数据集，并对基于图像、点云及多传感器融合的3D目标检测方法进行了分类分析。基于图像的方法受限于深度信息不足，检测精度较低；基于点云的方法借助深度信息，精度优势明显；多传感器融合方法则展现出更强的检测性能；而基于Transformer和图神经网络(GNN)的方法通过全局上下文建模与空间关系推理推动技术突破。通过对主流模型在KITTI和NuScenes数据集上的性能评估与对比，分析了不同模型在检测精度及复杂场景适应性上的表现差异。结论表明，未来可进一步探索轻量级网络架构、多模态动态融合策略及基于物理感知的小目标增强技术，结合Transformer全局建模与GNN关系推理，推动3D目标检测在实时性、复杂场景适应性及小目标检测精度上的突破。

关键词

目标检测，3D车辆检测，深度学习，计算机视觉

Survey on 3D Object Detection Based on Deep Learning

Bo Shao[1], Shiqi Xu[1], Ziqi Zhang[1], Linqian Yang[1], Weiqing Gao[2], Jiayi He[1], Aoxue Qin[1], Xiyin Wu[1]

[1]College of Big Data and Information Engineering, Guizhou University, Guiyang Guizhou
[2]College of Mathematics and Statistics, Guizhou University, Guiyang Guizhou

Received: Feb. 27th, 2025; accepted: Mar. 18th, 2025; published: Mar. 31st, 2025

Abstract

3D object detection based on deep learning holds significant importance in cutting-edge fields such

文章引用: 邵博, 徐仕琦, 张子琦, 杨琳倩, 高炜晴, 何嘉懿, 秦傲雪, 吴茜茵. 基于深度学习的 3D 目标检测技术综述[J]. 图像与信号处理, 2025, 14(2): 173-184. DOI: 10.12677/jisp.2025.142017

as autonomous driving and robotic navigation. However, the technology still faces multiple challenges, including high computational complexity when processing large-scale 3D data and difficulties in small object detection, which severely restrict its further development and widespread application. To address these limitations, this article provides a detailed introduction to commonly used datasets like KITTI and NuScenes, along with a categorized analysis of 3D object detection methods based on images, point clouds, and multi-sensor fusion. Image-based methods suffer from limited depth information and lower detection accuracy, while point cloud-based approaches demonstrate clear precision advantages by leveraging depth data. Multi-sensor fusion methods exhibit superior detection performance, whereas Transformer-based and Graph Neural Network (GNN) approaches drive technological breakthroughs through global context modeling and spatial relationship reasoning. Through performance evaluation and comparison of mainstream models on KITTI and NuScenes datasets, the study analyzes differences in detection accuracy and adaptability to complex scenarios. The conclusion suggests that future research could focus on exploring lightweight network architectures, dynamic multi-modal fusion strategies, and physics-aware enhancement techniques for small objects. By combining Transformer's global modeling with GNN's relational reasoning, breakthroughs may be achieved in real-time performance, complex scenario adaptability, and small object detection accuracy for 3D object detection.

Keywords

Object Detections, 3D Vehicle Detection, Deep Learning, Computer Vision

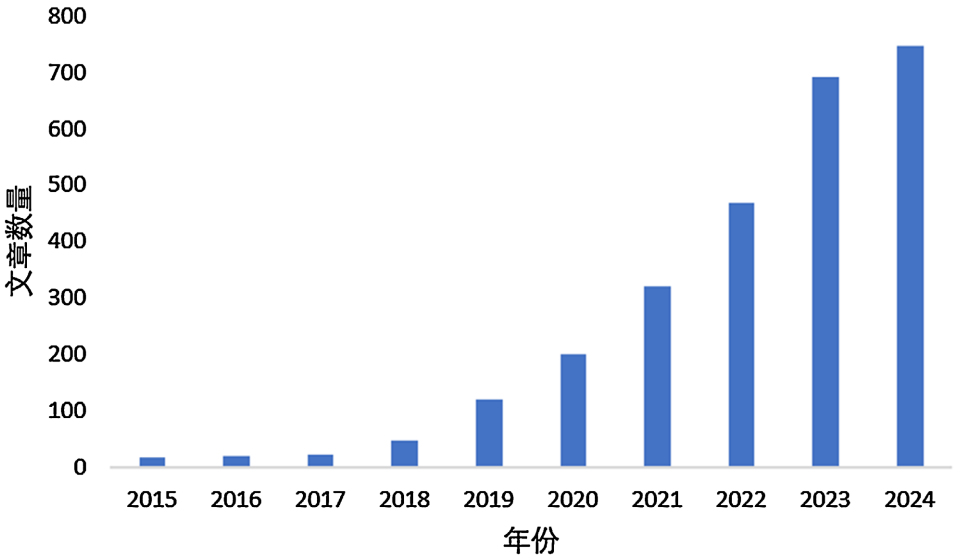
1. 引言

3D 目标检测是计算机视觉领域的关键技术，在诸多前沿科技领域扮演着重要角色。在自动驾驶场景中，车辆依靠 3D 目标检测技术实时精确地获取周围物体的三维位置、尺寸和类别信息，进而做出安全可靠的决策，其准确性和实时性直接决定了自动驾驶的安全性与可靠性[1]。随着《国家创新驱动发展战略纲要》等政策的推进[2]，对科技创新的重视程度日益提升，为 3D 目标检测技术的发展创造了良好的政策环境，促使其在多领域的应用研究不断深入。在过去的 10 年里，3D 目标检测在各领域受到了热烈的关注，出现了越来越多的有关 3D 目标检测的论文发表(如图 1)，其中包含了对 3D 目标检测方法的理论解释和对已有的 3D 目标检测模型的改进和推广应用。

早期，目标检测主要基于手工特征和机器学习算法[3]。在传统模式下，人们通过手工设计特征，再借助机器学习算法进行目标检测。但面对复杂多变的 3D 场景，这些方法存在诸多弊端。例如，在交通场景中，光照变化、物体遮挡等情况频繁出现，传统方法的检测精度和效率较低。手工设计的特征难以全面、准确地描述目标物体，机器学习算法在处理复杂特征时也力不从心，导致在实际应用中无法满足需求。从 2010 年起，基于传统机器学习的 3D 目标检测算法逐渐被应用，虽取得一定成果，但由于自然场景下 3D 目标的形态、位置、背景等因素的不确定性，这些算法的鲁棒性和泛化能力较差，难以在实际场景中广泛应用。

近年来，深度学习的蓬勃发展为 3D 目标检测带来了新的契机。深度学习具有强大的自动特征提取能力，能够从海量数据中学习到复杂的特征表示[4]。基于深度学习的 3D 目标检测方法，在检测精度、速度和适应性等方面都有显著提升。学者们利用深度学习的优势，不断优化算法和模型结构，使其在复杂

背景和多变环境下的检测效果更加理想[5]。然而，当前技术仍面临不少挑战，如处理大规模 3D 数据时计算资源消耗巨大，小目标检测精度难以保证等。本文对基于深度学习的 3D 目标检测技术进行全面地总结与分析，旨在梳理其发展脉络、剖析关键技术、探讨面临的问题，为后续研究提供参考与借鉴。



数据来源自 Google 学术检索关键字“3D object detection”

Figure 1. Number of papers related to object detection from 2015~2024
图 1. 2015~2024 年 3D 目标检测相关论文的数量

2. 数据集

为更全面、清晰地了解 3D 目标检测技术常用数据集的特点，以便在算法训练、评估时合理选择和运用，现将常用于智能驾驶领域 3D 目标检测的主要数据集关键信息汇总如下表 1。

Table 1. Summary table of 3D object detection datasets
表 1. 3D 目标检测数据集汇总表

数据集	场景数量	类别数量	标注帧数	3D 框数量	发布年份	传感器类型
KITTI	22	8	15K	200K	2012	RGB + 激光雷达
nuScenes	1K	23	40K	1.4M	2020	RGB + 激光雷达
H3D	160	8	27K	1.1M	2019	RGB + 激光雷达
Waymo	1K	4	200K	12M	2020	RGB + 激光雷达
Lyft Level 5	366	9	46K	1.3M	2019	RGB + 激光雷达
A*3D	-	7	39K	230K	2019	RGB + 激光雷达
ApolloScape	-	35	140K	70K	2019	RGB + 激光雷达

2.1. KITTI 数据集

KITTI 数据集(图 2)是自动驾驶场景下计算机视觉算法评估的重要数据集。它由配备 64 通道 LiDAR、4 个摄像头和 GPS/IMU 组合系统的车辆采集，包含城市、居民区和道路等 20 个场景的 RGB 图像、3D 激光雷达点云以及 GPS 坐标[6]。该数据集的 3D 目标检测基准包含 7481 张训练图像、7518 张测试图像及

相应点云，共标注了 80256 个物体，并根据检测难度将物体标注分为“easy”“moderate”和“hard”三类。KITTI 数据集的出现推动了 3D 目标检测算法的发展，许多研究成果都在该数据集上进行评估和比较。

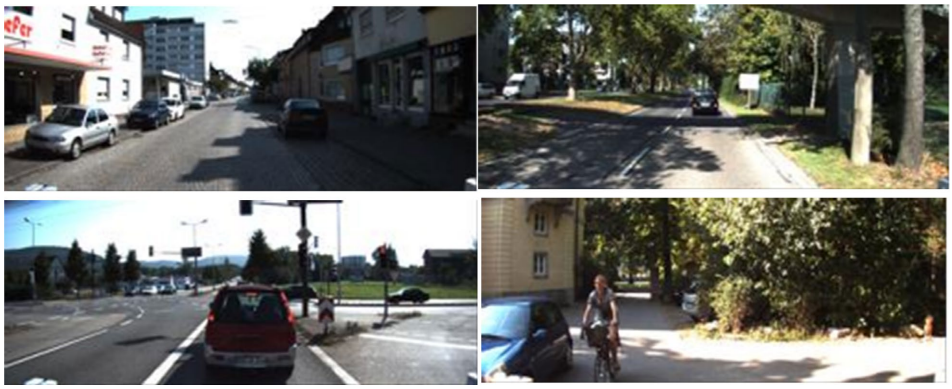


Figure 2. KITTI dataset
图 2. KITTI 数据集

2.2. NuScenes 数据集

NuScenes 数据集(图 3)规模比 KITTI 更大，由 6 个摄像头和 32 线束激光雷达在波士顿和新加坡采集，包含 700 个训练场景、150 个验证场景和 150 个测试场景[7]。其 3D 标注涵盖 360 度视野内的 23 个类别，在 3D 目标检测任务中，通常会去除样本较少的稀有类别，保留 10 个类别。该数据集包含 1000 个交通密集、驾驶场景极具挑战性的驾驶场景，且物体标注包含可见性、活动状态、姿态等属性，为 3D 目标检测算法的研究提供了更丰富的数据。

Figure 3. NuScenes dataset
图 3. NuScenes 数据集

2.3. Waymo 数据集

Waymo Open Dataset (图 4)是由谷歌旗下的 Waymo 公司发布的自动驾驶数据集[8]。它包含 3000 个驾驶记录，约 600,000 帧数据，有大约 2500 万个 3D 边界框和 2200 万个 2D 边界框。Waymo 数据集具有大量不同自动驾驶场景的数据，为 3D 目标检测算法在复杂场景下的训练和评估提供了有力支持，有助于推动算法在实际应用中的性能提升。

Figure 4. Waymo dataset
图 4. Waymo 数据集

2.4. 其他数据集

除上述数据集外，还有 ApolloScape、H3D、Lyft Level 5 等数据集。ApolloScape 由百度提供，包含复杂交通流的点云和高质量标签[9]；H3D 是本田提供的自动驾驶场景点云数据集，包含大规模自动驾驶数据和完整 360 度激光雷达数据集；Lyft Level 5 数据集通过 64 线雷达和多个摄像头采集，包含超过 55,000 个人工标注的 3D 注释帧、表面地图和高清空间语义地图。这些数据集在场景、类别、标注等方面各有特点，共同推动了 3D 目标检测技术的发展。

3. 基于深度学习的 3D 目标检测技术

随着深度学习在计算机视觉领域的广泛应用，3D 目标检测技术取得了显著进展。当前的基于深度学习的 3D 目标检测方法主要可归纳为三大类：基于图像的方法、基于点云的方法以及多传感器融合的方法。此外，近年来基于 Transformer 和图神经网络的方法也逐渐受到关注。本部分对各类方法的代表性研究进行了详细综述。现有 3D 目标检测首先训练消极样本和积极样本，经特征表示后进行分类；同时，传感器数据通过提案生成器产生提案，再经特征表示后进行回归；最后，分类和回归的结果共同形成检测结果，如图 5 所示。表 2 则对这些类别及其局限性进行了简要概括。

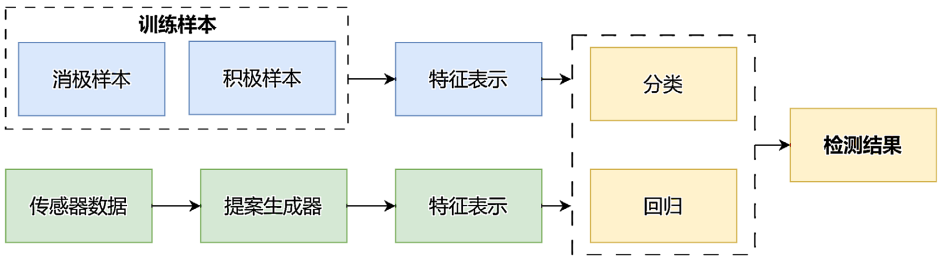


Figure 5. The overall framework of existing 3D object detection methods
图 5. 现有 3D 目标检测方法的总体框架

Table 2. Comparison of 3D object detection methods in different modes
表 2. 不同模式下 3D 目标检测方法的比较

模式		方法	局限性
图像		将图像用于预测 3D 物体的边界框。先预测 2D 边界框，然后通过重投影约束或回归模型外推到 3D。	深度信息不足，检测结果的准确性较低。
点云	2D 投影	将点云投影到 2D 平面，并利用 2D 检测框架回归投影到图像上的 3D 边界框。	投影过程中存在信息丢失。
	分割	进行体素化以获得 3D 体素，并通过对体素使用卷积操作生成表示，以预测物体的 3D 边界框。	昂贵的 3D 卷积操作增加了推理时间。计算量很大。
	3D 直接处理	应用原始点云直接预测物体的 3D 边界框。	大规模点云增加了运行时间。
多传感器融合		融合图像和点云以生成对 3D 边界框的预测。具有鲁棒性且能相互补充。	融合方法计算成本高且不够成熟。

3.1. 基于图像的 3D 目标检测方法

基于图像的 3D 目标检测方法主要依赖 RGB 图像的丰富纹理和颜色信息，以推理目标的 3D 边界框。早期的单目图像方法，如 Mono3D [10]，首先生成 3D 候选框，并借助语义信息、上下文特征和手工设计的形状特征，通过能量模型对候选框进行评分，最后利用 Fast RCNN 进行位置回归以优化边界框。然而，该方法依赖物体类别，并且为了提高召回率往往需要大量候选框，导致计算开销较高。为克服上述问题，研究人员提出了一系列改进方案。DeepStereoOP 通过融合 RGB 图像与深度信息提升 3D 目标检测性能[11]；3DVP 方法采用 3D 体素模式，结合 RGB 值、3D 形状和遮挡掩码，以最小化 3D 框投影到 2D 平面与 2D 检测结果之间的误差来优化检测效果[12]；SubCNN 引入子类别信息生成区域提案，并结合多尺度图像金字塔模型提升小目标的检测能力[13]。

近年来，许多研究人员开始关注直接从 2D 视角进行 3D 目标检测。例如，2D-driven 3D 目标检测方法基于手工特征训练多层感知器以预测 3D 边界框；MonoDIS 则利用 2D 和 3D 检测损失的解耦变换，并引入自监督置信度评分以提高检测精度[14]。此外，Pseudo-LiDAR 方法通过预测深度图并反投影生成 3D 点云，提出深度传播算法以扩散深度测量信息。然而，由于深度信息估计存在较高的不确定性，这类方法在遮挡场景和弱光环境下的检测精度仍受较大影响。

3.2. 基于点云的 3D 目标检测方法

LiDAR 传感器使用激光束来测量环境中障碍物的距离。该传感器输出一组 3D 点。与基于图像的方法相比，点云提供了可靠的深度信息，可以准确地定位目标。与图像中包含的结构信息不同，激光雷达点云具有无序性、稀疏性和信息有限的特征。此类方法可分为 2D 投影处理、3D 直接处理以及基于分割的方法。下面分别讲述这三部分。

3.2.1. 基于 2D 投影的方法

该类方法首先将点云投影到 2D 视图，如前视图、鸟瞰图(BEV)或范围视图(RV)，然后利用 2D 卷积网络进行目标检测。例如，LMNet 采用前视图投影，并结合带扩展卷积的 FCN 进行单阶段检测，该方法可达到实时检测性能，但检测精度较低[15]；BirdNet 通过 BEV 投影后应用 Faster R-CNN 估计 3D 边界框；受 YOLO 启发，Complex-YOLO 将点云投影到 BEV 中，然后使用单阶段策略估计物体的三维边界框，显著提高了检测效率。PointPillars 通过将点云转换为柱状结构(pillars)，利用 2D CNN 进行特征提取，有效降低了计算开销。然而，由于点云的稀疏性和不均匀分布，该类方法在检测小尺度目标时存在较大

挑战。

3.2.2. 基于 3D 直接处理的方法

此类方法直接利用原始点云进行特征提取，并采用 3D 神经网络进行检测。3D FCN 和 Vote3Deep 直接利用 3D 卷积神经网络进行 3D 目标检测，但受限于点云的稀疏性和 3D 卷积的计算成本，难以高效学习多尺度特征。PointNet 及其改进版 PointNet++通过对点云进行直接处理，利用 MaxPooling 对称函数提取点云特征，解决点的无序问题。基于此，PointRCNN 采用 PointNet++架构，通过 3D 提案生成模块获取高质量候选框，从而更好地学习局部空间特征，并结合全局语义特征提升检测精度[16]。此外，基于球形锚点的提案生成方法使用 PointNet++作为骨干网络，提取每个点的语义上下文特征。同时，在边框预测的第二阶段，增加了一个 IoU 估计分支进行后处理，进一步提高了目标检测的精度。然而，基于 3D 卷积的方法计算代价较高，处理大规模场景时存在效率瓶颈。

3.2.3. 基于分割的方法

此类方法通过空间分割策略对点云进行体素化处理，并利用 3D 卷积提取分组体素的特征。例如，VoxelNet 将点云划分为一定数量的体素，并通过随机采样和归一化提取每个非空体素的局部特征[17]。PV-RCNN 结合三维体素 CNN 与 PointNet 架构，实现更具判别力的点云特征学习。然而，这些方法的检测性能依赖于体素划分策略，分割质量直接影响最终检测结果。

3.3. 多传感器融合的 3D 目标检测方法

考虑到基于图像的方法和基于点云的方法的优缺点，多传感器融合的方法试图将这两种模式与不同的策略相融合。将激光雷达点云与图像进行融合，对点云进行投影变换，然后通过不同特征融合方案将多视图投影平面与图像进行集成。融合方式主要包括早期融合、晚期融合和深度融合。早期融合是在特征提取层或数据预处理阶段将不同传感器的特征融合，如 AVOD 在提案阶段融合各模态特征，通过 FC 层输出 3D 框的类别和坐标[18]。晚期融合则在检测结果阶段融合，如一些方法先分别进行图像和点云的检测，再融合两者结果。深度融合是让特征图在不同层次上进行交互，如 MV3D 利用深度融合方案聚合特征。但这些基于投影变换的融合方法在投影过程中会丢失空间信息[19]，小目标检测性能较差，且 LiDAR 点云的稀疏性限制了融合效果。

近年来，一些方法尝试直接处理原始点云并进行更深层次的特征融合。F-PointNet 结合 2D 检测结果获取物体的视锥空间，并利用 PointNet++进行 3D 目标检测[20]；PointFusion 融合 RGB 图像块和相应点云特征，提高了检测性能。此外，3D-CVF 提出跨视图空间特征融合策略，以增强多模态信息的互补性；EPNet 则利用一致性增强损失，提高目标定位与分类的精确度。这些方法从不同角度探索了点云处理和特征融合的新途径，为 3D 目标检测技术的发展注入了新的活力，推动着该领域不断向前发展。

3.4. 基于 Transformer 的 3D 目标检测方法

近年来，基于 Transformer 的 3D 目标检测方法通过注意力机制的全局上下文建模能力，显著突破了传统方法的性能瓶颈。例如，BEVFormer 通过学习统一的 BEV 表示，利用时空 Transformer 有效聚合多摄像头图像的时空信息，在复杂场景感知中表现出色；BEVFusion 创新性地将多模态特征统一在共享的 BEV 表示空间中，通过优化 BEV 池化操作，解决了视图转换中的效率瓶颈问题[21]；而 IS-Fusion 提出了实例--场景协作融合的概念，通过分层场景融合(HSF)模块和实例引导融合(IGF)模块，分别捕捉不同粒度的场景上下文信息和实例级的多模态信息，促进实例与场景特征之间的交互，从而获得更适合实例感知任务的增强 BEV 表示[22]。这些方法构建了从时空信息聚合到多模态特征精调的完整技术链，共同推

动基于 Transformer 的 3D 检测迈向场景适应性更强、特征粒度更细的新阶段。

3.5. 基于图神经网络的 3D 目标检测方法

近期，基于图神经网络的 3D 目标检测方法通过图结构建模点云或多模态数据的空间关系，显著提升了复杂场景下的检测性能。例如，Point-GNN 提出固定半径近邻图和自动配准机制，避免点云采样冗余，增强平移不变性，设计框合并与评分操作，优化多顶点检测结果[23]；DCGNN 引入密度聚类球查询优化点云分组，结合层次化 GNN 架构，同时捕捉局部点集细节与全局跨点集关系，实现单阶段高效检测[24]；HetGNN-3D 构建异构图融合激光雷达与图像数据，通过动态消息传递和跨模态边关系预测实现传感器解耦，子图读出优化检测结果，提升鲁棒性[25]。这些方法构建了从点云优化到多模态协同的完整技术路径，通过图结构的关系建模能力，突破传统点云处理的局限性，为自动驾驶感知提供了鲁棒性更强的解决方案。

4. 主流 3D 目标检测模型性能评估与对比

4.1. 评估指标体系

以下是 3D 目标检测中广泛应用的评估指标：

4.1.1. 交并比(IoU)

主要用于衡量预测框与真实框的重叠程度。若预测框和真实框完美重合，IoU 值为 1；若两者无重叠，IoU 为 0。IoU 越接近 1，表明模型预测的目标位置和尺寸越准确，能直观反映模型对目标的定位精度。

4.1.2. 平均精度(AP)

综合考量模型在不同召回率下的精度表现。召回率体现模型找到所有真实目标的能力，精度表示模型判断为目标的检测结果中真正属于目标的比例。AP 通过对不同召回率下精度的综合计算，全面评估模型在不同检测难度下的整体性能，AP 值越高，模型在检测各类目标时表现越出色。

4.1.3. nuScenes 数据集特殊指标

①平均平移误差(ATE)。评估预测的 3D 物体中心位置与真实物体中心位置的偏差程度，ATE 值越小，模型对目标位置的预测越精准。

②平均尺度误差(ASE)。衡量预测物体的尺寸与真实尺寸的差异，该值越小，模型对目标大小的判断越准确。

③平均方向误差(AVE)。反映预测物体的方向与真实方向的差异，用于评估模型对目标方向预测的准确性。

④平均属性误差(AAE)。针对预测物体的属性(如可见性、活动状态等)与真实属性之间的误差进行计算，AAE 越小，模型对物体属性的判断越接近真实情况。

⑤NuScenes 检测分数(NDS)：这是一个综合指标，通过综合考虑各类别目标的检测精度、召回率等因素，全面衡量模型在多种目标检测任务上的整体表现，NDS 值越高，意味着模型 在 nuScenes 数据集上的检测性能越出色。

这些指标从不同维度评估模型性能，有助于更细致地分析模型在 3D 目标检测中的表现，为模型改进和优化提供精准方向。

4.2. 不同模型性能对比分析

我们对比了一系列 3D 目标检测方法在多类目标上的检测成果。表 3 展示 KITTI 测试集前沿方法对

比，含不同模式下各方法针对不同目标类别在三种难度的准确性；另一表 4 呈现 nuScenes 测试集前沿方法结果，包括 mAP、NDS 等及各目标精度。

分析可知，KITTI 数据集上，基于图像的模型如 Mono3D、Deep3DBox 因缺深度信息，检测精度低，“easy”难度下汽车检测平均精度多为个位数。基于点云的模型如 VoxelNet、PointPillars，凭深度信息提升精度，“easy”难度下超 70%。

nuScenes 数据集评估中，模型性能差异显著。基于激光雷达的 PointPillar，mAP 约 30.5%，NDS 为 45.3%。CenterPoint 优化后，mAP 升至 60.3%，NDS 达 67.3%。RGB + 激光雷达融合的 BEVFusion 和 IS-Fusion 表现突出，前者 mAP 达 70.2%，NDS 为 72.9%；后者 mAP 为 73.0%，NDS 达 75.2%。可见多传感器融合对自动驾驶 3D 目标检测至关重要，能整合优势提升复杂场景检测性能。

Table 3. Performance comparison of 3D object detection models on the KITTI test set (average accuracy %)
表 3. KITTI 测试集上 3D 目标检测模型性能比较(平均精度%)

方法	模式	汽车			行人			骑自行车的人		
		易	中	难	易	中	难	易	中	难
Mono3D	RGB	2.53	2.31	2.31	-	-	-	-	-	-
Deep3DBox	RGB	5.84	4.09	3.83	-	-	-	-	-	-
OFT-Net	RGB	3.28	2.50	2.27	1.06	1.11	1.06	0.43	0.43	0.43
MonoPair	RGB	13.04	9.99	8.75	-	-	-	-	-	-
VoxelNet	激光雷达	77.47	65.11	57.73	-	-	-	-	-	-
PointPillars	激光雷达	82.58	74.32	68.99	51.45	41.92	38.89	77.10	58.65	51.92
PointRCNN	激光雷达	86.96	75.64	70.70	47.98	39.37	36.01	74.96	58.82	52.53
MV3D	RGB + 激光雷达	74.97	63.63	54.00	-	-	-	-	-	-
AVOD	RGB + 激光雷达	83.07	71.76	65.73	50.46	42.27	39.04	63.76	50.55	44.93
F-PointNet	RGB + 激光雷达	82.19	69.79	60.59	50.53	42.15	38.08	72.27	56.12	49.01
3D-CVF	RGB + 激光雷达	89.20	80.05	73.11	-	-	-	-	-	-

Table 4. Performance comparison of 3D object detection models on the NuScenes test set (%)
表 4. NuScenes 测试集上 3D 目标检测模型性能比较(%)

方法	模式	mAP	NDS	汽车	卡车	施工车辆	公共汽车	拖车	障碍物	摩托车	自行车	行人	交通锥
PointPillar	激光雷达	30.5	45.3	68.4	23.0	4.1	28.2	23.4	38.9	27.4	1.1	59.7	30.8
CenterPoint	激光雷达	60.3	67.3	85.2	53.5	20.0	63.6	56.0	71.1	59.5	30.7	84.6	78.4
FusionPainting	RGB + 激光雷达	66.3	70.4	86.3	58.5	27.7	66.8	59.4	70.2	71.2	51.7	87.5	84.2
PointAugment-ing	RGB + 激光雷达	66.8	71.0	87.5	57.3	28.0	65.2	60.7	72.6	74.3	50.9	87.9	83.6
BEVFusion	RGB + 激光雷达	70.2	72.9	88.6	60.1	39.3	69.8	63.8	80.0	74.1	51.0	89.2	86.5
IS-Fusion	RGB + 激光雷达	73.0	75.2	88.3	62.7	38.4	74.9	67.3	78.1	82.4	59.5	89.3	89.2

5. 现有挑战

5.1. 处理大规模 3D 数据的计算复杂性

3D 目标检测需要处理大量的 3D 数据，如激光雷达点云数据。这些数据具有数据量大、维度高的特点，给计算带来了巨大挑战。3D 卷积操作计算成本高，会增加推理时间，使得一些基于 3D 卷积的方法难以满足实时性要求[26]。即使采用一些优化策略，如体素化、下采样等，仍然无法完全解决计算复杂性的问题。在实际应用中，如自动驾驶场景，需要实时处理传感器采集的大量数据，对计算资源的需求极高，这限制了一些复杂模型的应用。

5.2. 小目标的检测问题

小目标在 3D 场景中普遍存在，由于其在数据中的占比小、特征不明显，检测难度较大。在稀疏数据和噪声背景下，小目标的检测更加困难[27]。点云数据的稀疏性使得小目标的点云数量较少，难以提取有效的特征；噪声会干扰对小目标的检测，导致误检或漏检。此外，现有模型在感受野、特征提取能力等方面可能存在不足，对小目标的检测精度和召回率较低，无法满足实际应用的需求。

6. 总结与未来趋势

6.1. 总结

本文全面且深入地综述了基于深度学习的 3D 目标检测技术，详细介绍了 KITTI、NuScenes、Waymo 等常用数据集。文中深入探讨基于图像、点云及多传感器融合的 3D 目标检测方法，基于图像的方法受限于深度信息不足，检测精度受限；基于点云的方法借助深度信息，在检测精度上优势明显；多传感器融合方法融合多种传感器优势，展现出提升检测性能的潜力。通过在 KITTI 和 NuScenes 等数据集上对主流模型的性能评估与对比可知，不同模型在检测精度、召回率等指标以及复杂和遮挡场景适应性上表现不同。尽管该技术已取得显著进展，在自动驾驶、机器人视觉等实际领域广泛应用，为智能系统提供关键环境感知能力，但仍面临处理大规模 3D 数据计算复杂和小目标检测困难等挑战。

6.2. 未来趋势

展望未来，3D 目标检测技术将在多方面取得显著进展。为提升实时检测能力，可融合 MobileNet 这类轻量级模型架构，结合层级化动态网络设计(如基于 MoE 的稀疏激活机制)，通过模型剪枝、量化技术及神经架构搜索(NAS)实现硬件适配的推理加速，同时利用稀疏卷积处理 3D 点云数据，满足自动驾驶等场景的实时性需求。在复杂环境下，采用多尺度特征融合与注意力机制，结合生成对抗网络(GAN)合成极端条件下的小目标数据，可增强模型鲁棒性；引入物理感知模型(如光线追踪补偿稀疏点云)与几何约束，进一步提升小目标检测精度。

多模态数据融合方面，除数据层预处理对齐、特征层融合模块设计及决策层结果综合外，可开发自适应融合权重机制，根据场景复杂度动态调整图像与点云的融合策略，并探索时空序列融合技术捕捉多帧运动轨迹，解决快速移动物体检测延迟问题。深度学习模型优化上，运用知识蒸馏等模型压缩技术降低部署成本，探索基于 Transformer 的全局上下文建模与量子神经网络(QNN)的交叉应用，同时通过元学习(Meta-Learning)实现小样本快速迁移，提升模型泛化能力。此外，构建包含小目标专项指标的多维度评估框架，将推动算法向实用化迈进。

这些关键技术的协同发展，不仅会推动 3D 目标检测技术在精度、速度和适应性上迈向新高度，还将促使其在自动驾驶、机器人、智能安防等领域实现更广泛、更深入的应用，为智能时代的感知技术变革

注入强大动力。

基金项目

2024 年贵州大学创新创业训练计划项目(项目编号 gzusc2024024)。

参考文献

[1] Li, Z., Du, Y., Zhu, M., Zhou, S. and Zhang, L. (2021) A Survey of 3D Object Detection Algorithms for Intelligent Vehicles Development. *Artificial Life and Robotics*, **27**, 115-122. https://doi.org/10.1007/s10015-021-00711-0

[2] 振兴发展靠人才(一)——《关于贯彻〈国家创新驱动发展战略纲要〉建设科技强省的实施意见》解读[J]. 共产党员, 2017(18): 46-47.

[3] 陈辉东, 丁小燕, 刘艳霞. 基于深度学习的目标检测算法综述[J]. 北京联合大学学报, 2021, 35(3): 39-46.

[4] 谢富, 朱定局. 深度学习目标检测方法综述[J]. 计算机系统应用, 2022, 31(2): 1-12.

[5] 戴德云, 陈宗海, 鲍鹏, 等. 电动汽车自动驾驶 3D 目标检测综述[J]. 世界电动汽车杂志, 2021, 12(3): 139.

[6] Geiger, A., Lenz, P. and Urtasun, R. (2012) Are We Ready for Autonomous Driving? The KITTI Vision Benchmark Suite. 2012 *IEEE Conference on Computer Vision and Pattern Recognition*, Providence, 16-21 June 2012, 3354-3361. https://doi.org/10.1109/cvpr.2012.6248074

[7] Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., *et al*. (2020) NuScenes: A Multimodal Dataset for Autonomous Driving. 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Seattle, 13-19 June 2020, 11618-11628. https://doi.org/10.1109/cvpr42600.2020.01164

[8] Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., *et al*. (2020) Scalability in Perception for Autonomous Driving: Waymo Open Dataset. 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Seattle, 13-19 June 2020, 2443-2451. https://doi.org/10.1109/cvpr42600.2020.00252

[9] Huang, X., Wang, P., Cheng, X., Zhou, D., Geng, Q. and Yang, R. (2020) The Apolloscape Open Dataset for Autonomous Driving and Its Application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **42**, 2702-2719. https://doi.org/10.1109/tpami.2019.2926463

[10] He, T. and Soatto, S. (2019) Mono3D++: Monocular 3D Vehicle Detection with Two-Scale 3D Hypotheses and Task Priors. *Proceedings of the AAAI Conference on Artificial Intelligence*, **33**, 8409-8416. https://doi.org/10.1609/aaai.v33i01.33018409

[11] Flynn, J., Neulander, I., Philbin, J. and Snavely, N. (2016) Deep Stereo: Learning to Predict New Views from the World's Imagery. 2016 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Las Vegas, 27-30 June 2016, 5515-5524. https://doi.org/10.1109/cvpr.2016.595

[12] Xiang, Y., Choi, W., Lin, Y. and Savarese, S. (2015) Data-Driven 3D Voxel Patterns for Object Category Recognition. 2015 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Boston, 7-12 June 2015, 1903-1911. https://doi.org/10.1109/cvpr.2015.7298800

[13] Choi, W., Lin, Y., Xiang, Y., *et al*. (2018) Subcategory-Aware Convolutional Neural Networks for Object Detection. U.S. Patent 9,965,719.

[14] Simonelli, A., Bulo, S.R., Porzi, L., Lopez-Antequera, M. and Kontschieder, P. (2019) Disentangling Monocular 3D Object Detection. 2019 *IEEE/CVF International Conference on Computer Vision* (*ICCV*), Seoul, 27 October-2 November 2019, 1991-1999. https://doi.org/10.1109/iccv.2019.00208

[15] Minemura, K., Liau, H., Monrroy, A. and Kato, S. (2018) LMNet: Real-Time Multiclass Object Detection on CPU Using 3D LiDAR. 2018 3*rd Asia-Pacific Conference on Intelligent Robot Systems* (*ACIRS*), Singapore, 21-23 July 2018, 28-34. https://doi.org/10.1109/acirs.2018.8467245

[16] Shi, S., Wang, X. and Li, H. (2019) PointRCNN: 3D Object Proposal Generation and Detection from Point Cloud. 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Long Beach, 15-20 June 2019, 770-779. https://doi.org/10.1109/cvpr.2019.00086

[17] Zhou, Y. and Tuzel, O. (2018) VoxelNet: End-to-End Learning for Point Cloud Based 3D Object Detection. 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-23 June 2018, 4490-4499. https://doi.org/10.1109/cvpr.2018.00472

[18] Ku, J., Mozifian, M., Lee, J., Harakeh, A. and Waslander, S.L. (2018) Joint 3D Proposal Generation and Object Detection from View Aggregation. 2018 *IEEE/RSJ International Conference on Intelligent Robots and Systems* (*IROS*), Madrid, 1-5 October 2018, 1-8. https://doi.org/10.1109/iros.2018.8594049

[19] Chen, X., Ma, H., Wan, J., Li, B. and Xia, T. (2017) Multi-View 3D Object Detection Network for Autonomous Driving.

2017 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Honolulu, 21-26 July 2017, 6526-6534. https://doi.org/10.1109/cvpr.2017.691

[20] Qi, C.R., Liu, W., Wu, C., Su, H. and Guibas, L.J. (2018) Frustum PointNets for 3D Object Detection from RGB-D Data. 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-23 June 2018, 918-927. https://doi.org/10.1109/cvpr.2018.00102

[21] Liu, Z., Tang, H., Amini, A., Yang, X., Mao, H., Rus, D.L., *et al*. (2023) BEVFusion: Multi-Task Multi-Sensor Fusion with Unified Bird's-Eye View Representation. 2023 *IEEE International Conference on Robotics and Automation* (*ICRA*), London, 29 May-2 June 2023, 2774-2781. https://doi.org/10.1109/icra48891.2023.10160968

[22] Yin, J., Shen, J., Chen, R., Li, W., Yang, R., Frossard, P., *et al*. (2024) IS-Fusion: Instance-Scene Collaborative Fusion for Multimodal 3D Object Detection. 2024 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Seattle, 16-22 June 2024, 14905-14915. https://doi.org/10.1109/cvpr52733.2024.01412

[23] Shi, W. and Rajkumar, R. (2020) Point-GNN: Graph Neural Network for 3D Object Detection in a Point Cloud. 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Seattle, 13-19 June 2020, 1708-1716. https://doi.org/10.1109/cvpr42600.2020.00178

[24] Xiong, S., Li, B. and Zhu, S. (2022) DCGNN: A Single-Stage 3D Object Detection Network Based on Density Clustering and Graph Neural Network. *Complex & Intelligent Systems*, **9**, 3399-3408. https://doi.org/10.1007/s40747-022-00926-z

[25] 汪明明, 陈庆奎, 付直兵. HetGNN-3D: 基于异构图神经网络的 3D 目标检测优化模型[J]. 小型微型计算机系统, 2024, 45(2): 438-445.

[26] Qi, C.R., Yi, L., Su, H. and Guibas, L.J. (2017) PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space. *Proceedings of the* 31*st International Conference on Neural Information Processing Systems*, Long Beach, 4-9 December 2017, 5105-5114.

[27] Yang, Z., Sun, Y., Liu, S., Shen, X. and Jia, J. (2019) STD: Sparse-to-Dense 3D Object Detector for Point Cloud. 2019 *IEEE/CVF International Conference on Computer Vision* (*ICCV*), Seoul, 27 October-2 November 2019, 1951-1960. https://doi.org/10.1109/iccv.2019.00204