网络优先出版，实际出版页码及内容以正式版为准

第 55 卷第 8 期 光 子 学 报 Vol.55 No.8
2026 年 8 月 ACTA PHOTONICA SINICA August 2026

引用格式：ZHAO Hongwei, XIA Wentao, LIU Wenlong, et al. Research Progress on LiDAR-Based 3D Object Detection Algorithms[J]. Acta Photonica Sinica, 2026, 55(8):0814001
赵洪伟，夏文涛，刘文龙，等. 激光雷达三维目标检测算法研究进展[J]. 光子学报，2026，55(8)：0814001

激光雷达三维目标检测算法研究进展

赵洪伟[1,2]，夏文涛[1,3]，刘文龙[1,2]，谢毅[1]，何超建[1,4,5,6]，上官明佳[7]，
于海娟[1,4,5,6]，杨盈莹[1,4,5,6]，杨松[1,4,5,6*]，林学春[1,4,5,6]

（1 中国科学院半导体研究所全固态光源实验室，北京 100083）
（2 中国科学院大学电子电气与通信工程学院，北京 101407）
（3 中国科学院大学现代产业学院，北京 101407）
（4 中国科学院大学材料科学与光电技术学院，北京 101407）
（5 中国科学院大学材料科学与光电工程中心，北京 100049）
（6 北京市全固态激光先进制造工程技术研究中心，北京 100083）
（7 厦门大学海洋与地球学院，海洋生物地球化学全国重点实验室，福建 厦门 361102）

摘 要：激光雷达三维目标检测可为自动驾驶和铁路异物入侵监测等任务提供目标三维位置和尺寸信息。受点云稀疏、遮挡造成局部缺失以及语义信息不足影响，现有方法在远距离目标、小目标和复杂环境下的检测结果仍不够稳定。本文从传统几何处理、深度学习检测和激光雷达—相机融合三类方法展开分析。传统几何方法流程清晰，适用于规则场景中的候选筛选和误差控制，但对阈值和环境条件较敏感；深度学习方法提升了检测精度，但在远距离稀疏点云和小目标检测中仍易漏检；多模态融合方法能够利用图像语义信息补充点云特征，但标定误差和时间同步误差会降低检测可靠性。后续研究仍需解决稀疏点云下检测不稳定、小目标易漏检和三维标注成本高等问题。

关键词：激光雷达；三维目标检测；点云；深度学习；多模态融合

中图分类号：TP183 **文献标识码**：A **doi**：10.3788/gzxb20265508.0814001

0 引言

在自动驾驶、铁路异物入侵监测等应用中，系统不仅需要识别目标类别，还需要获取目标在三维空间中的位置、尺寸和姿态信息，以支撑后续路径规划、风险预警和运动控制。三维目标检测因此成为环境感知中的基础任务[1]。

与基于图像的检测方法相比，视觉方法在纹理和语义表达方面具有一定优势，数据获取成本相对较低，但图像本身不直接提供准确的三维结构信息，目标距离和空间位置通常需要借助深度估计或多视几何等方法恢复，因此在复杂场景下容易受到光照变化、遮挡和尺度歧义的影响。激光雷达能够直接获取场景的三维几何信息和目标距离信息，在空间定位和结构表达方面更为直接，因此被广泛用于障碍物检测、场景建模和目标感知等任务[2]。

点云三维目标检测早期主要依赖传统几何方法[3]。这类方法通常围绕环境分割、空间聚类、规则拟合和手工特征设计展开，通过地面分割、候选提取和几何约束完成目标检测，流程较为清晰，但比较依赖场景选择和参数设置，在复杂背景、点云稀疏和局部遮挡条件下稳定性有限。随着公开数据集的完善和深度学习技术的发展[4]，研究逐渐转向以表征学习为核心的检测框架，输入表示、特征提取和检测结构不断丰富。与

基金项目：国家自然科学基金(No.62225507, 62475254)，中国科学院稳定支持基础研究领域青年团队计划 (No. YSBR－065)资助
第一作者：赵洪伟，男，硕士研究生，研究方向为激光雷达三维目标检测。E－mail: zhaohongwei24@semi.ac.cn
通讯作者：杨松，男，研究员，研究方向为超快激光与激光雷达。E－mail: yangsong@semi.ac.cn
收稿日期：2026-03-29；**录用日期**：2026-06-15

此同时，为弥补单一激光雷达在纹理和语义信息方面的不足，研究者开始将图像信息引入三维检测流程，激光雷达与相机融合方法也在逐渐发展[5]。但是不同技术路线在适用场景、检测性能和工程实现方面仍存在较大差异。对于远距离稀疏目标、小目标、强遮挡场景以及多模态配准误差影响下的检测问题，现有方法仍有进一步梳理和总结的必要。基于此，本文围绕激光雷达三维目标检测的主要技术路线展开，从传统几何方法、基于深度学习的点云检测方法以及激光雷达与相机融合方法三个方面，重点比较不同路线在检测精度、场景适应性和工程实现方面的差异，并讨论后续研究中需要关注的问题。

1 基于几何方法的三维目标检测

传统激光雷达三维目标检测通常先从原始点云中去除地面等背景点，再依据空间邻近关系或局部密度特征将非地面点划分为候选团簇，最后结合几何尺寸、形状特征完成目标判别。其处理流程一般包括背景分割、空间聚类和障碍物团簇分类三个步骤。本章将围绕这三个步骤，对相关算法原理、适用条件和局限性进行梳理。

1.1 背景分割

在道路、铁路和城市交通等场景中，地面点通常占据原始点云的较大比例。若不先去除地面点，地面点可能通过空间连通关系将相邻障碍物错误连接，同时也会增加后续聚类和候选提取的搜索范围，降低处理效率。因此，地面分割是传统三维目标检测前处理中的重要步骤之一。

早期的地面分割广泛采用随机采样一致性(Random Sample Consensus, RANSAC)算法。该算法建立在几何先验假设之上，即局部路面在数学上近似符合平面方程 $ax+by+cz+d=0$。RANSAC通过迭代采样策略在数据集中寻找最优参数模型。其核心优化目标是最大化属于平面模型的内点集基数，其代价函数可表述为：

$$\hat{\theta}=\arg max_{\theta}\sum_{p_i\in P}I\left(\frac{\left|ax_i+by_i+cz_i+d\right|}{\sqrt{a^2+b^2+c^2}}<\delta\right) \tag{1}$$

式中，P为原始点云集合，$I(\cdot)$为指示函数；当点到候选平面的距离小于阈值δ时，$I=1$，否则$I=0$。其中δ为距离容差阈值。虽然RANSAC具有较强的抗噪能力，但其迭代次数k与理论置信度p呈非线性关系$k=\frac{\log(1-p)}{\log(1-w^n)}$($w$为内点比例)，导致计算耗时具有不确定性。此外，单一平面假设无法描述带有坡度或起伏的真实道路，常导致远距离地面的欠分割。为解决上述问题，ZERMAS D等[6]提出了确定性的地面拟合(Ground Plane Fitting, GPF)算法。该研究引入了初始种子点(initial seed points)的概念，通过启发式规则选取局部区域高度最低的N_{LPR}个点作为初始集合，计算其均值μ与协方差矩阵Σ，并利用主成分分析将Σ最小特征值对应的特征向量作为平面法向量。GPF通过迭代加权最小二乘法快速收敛，在保证精度的同时将处理速度提升至50Hz以上，成为目前工程应用的经典范式。

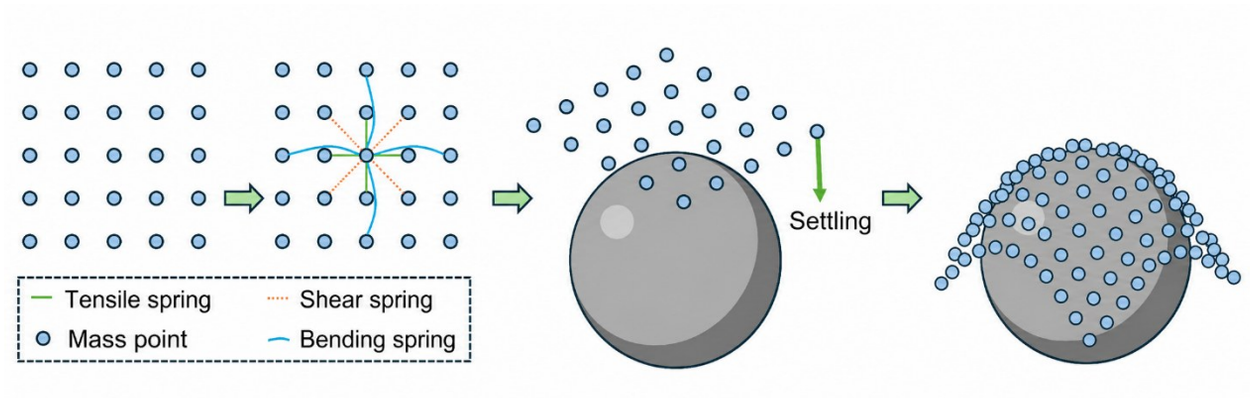

图1 布料模拟滤波(Cloth Simulation Filtering, CSF)算法的物理模型与弹簧质点配置

Fig.1 Physical Model of the Cloth Simulation Filtering(CSF) Algorithm and Spring-Mass Configuration

不过，GPF本质上仍然依赖平面模型，对非结构化地形的适应性有限。针对这一物理局限，HIMMELSBACH M等[7]提出了基于扫描线（Scanline-based）的分割策略。该方法利用机械式激光雷达线束在垂直方向上的有序分布特性，将三维点云映射至柱坐标系或极坐标网格。对于同一条扫描线上的相邻点p_i与p_{i-1}，算法计算其斜率梯度$m=\Delta z/\Delta\rho$。当局部斜率与高度差均小于特定阈值时，判定该点为地面。这种方法将3D分割问题降维为2D曲线分析，极大地降低了计算复杂度$O(N)$，且对微小坡度具有天然的适应性。BOGOSLAVSKYI I等人[8]在此基础上进一步提出了基于深度图（Range Image）的快速分割方法，通过在图像空间中使用广度优先搜索进行地面标记，避免了稀疏点云上复杂的邻域检索。

为了在数学层面实现对野外复杂起伏地形（如矿山、越野环境）的连续建模，CHEN Tongtong等人[9]引入了高斯过程回归（Gaussian Process Regression，GPR）。GPR摒弃了刚性几何假设，将地面高程z建模为关于水平坐标(x,y)的随机过程$z(x,y)\sim\mathcal{GP}\left(m(x,y),k\left((x,y),(x',y')\right)\right)$，通过径向基核函数（Radial Basis Function，RBF）捕捉地形的空间相关性，GPR能够生成平滑且连续的概率曲面[10]。尽管其计算复杂度较高（$O(N^3)$），但在处理非线性地形时表现出卓越的鲁棒性。另一类较有代表性的方法是由ZHANG Wuming等人[11]提出的布料模拟滤波（Cloth Simulation Filter，CSF），如图1所示，将地面估计问题转化为物理模拟问题：一块具有一定刚度的虚拟布料覆盖在翻转后的点云表面，在重力作用下最终贴合出的曲面即被视为地表。该方法通过求解质点—弹簧系统的运动方程，自适应拟合复杂地形，且参数较少，因此在测绘与自动驾驶领域都得到了广泛应用[12]。

1.2 空间聚类

完成背景分割后，剩余的点云通常呈现为离散的非地面点集。此时需要依据空间分布与局部密度特征，将属于同一物理实体的点归并为同一团簇。最基础且应用最广的方法是欧式聚类（Euclidean Clustering）。RUSU R B等[13]在点云库（Point Cloud Library，PCL）中给出了该算法的标准流程：首先利用kd-tree建立空间索引；随后以任意未标记点为起点，检索其邻域内距离小于阈值的点集，并通过队列递归扩展，直到簇闭合为止。然而，激光雷达点云具有明显的“近密远疏”特征，点云密度随距离衰减。这使得固定阈值的聚类方法难以同时兼顾近距离和远距离目标：阈值较小时，远处稀疏目标容易被错误切分为多个团簇；阈值较大时，近处相邻目标又可能被粘连在一起。针对这一问题，研究者提出了自适应聚类策略，其核心是让判别阈值随探测距离动态变化[14]。典型的阈值函数可写为线性模型：

$$d_{\mathrm{th}}(r)=\lambda\cdot r\cdot\Delta\theta+\sigma \tag{2}$$

式中，$\Delta\theta$为激光雷达的水平角分辨率，λ为调节系数，σ为传感器测距噪声底限。该机制使得近距离区域仍保持较强的区分能力，而远距离区域则具备更大的容差范围，本质上是在传感器几何分辨率约束下进行尺度一致化处理。

欧式聚类本质上依赖全局单一距离度量，隐含了目标在空间中呈现简单凸结构的假设。当目标呈现复杂拓扑或非凸几何，如护栏、树木等连续但不规则分布的目标时，这种方法容易因为局部密度微小变化而发

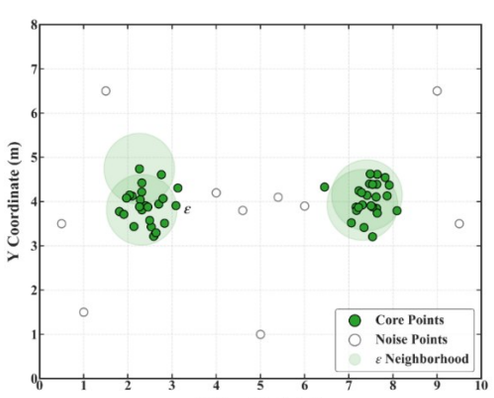

图2 结合核心点与密度可达性定义的DBSCAN聚类算法原理

Fig.2 Principle of the DBSCAN clustering algorithm defined by core points and density-reachability

生过分割。为此，具有噪声的基于密度的空间聚类算法（Density-Based Spatial Clustering of Applications with Noise，DBSCAN）被广泛应用于点云聚类与离群点剔除。该算法最早由ESTER M等人在1996年的KDD会议论文中提出[15]。其原理如图2所示，通过核心点、边界点和噪声点的定义，以密度可达关系构造簇，不需要预设簇数，并且对离群点具有鲁棒性。然而，原始DBSCAN在最坏情况下的时间复杂度为$O(N^2)$，难以满足在线处理需求。针对这一问题，WANG Heng等人[16]提出基于网格的改进方案，通过体素化加速密度查询。此外，针对粘连目标的分割，部分研究采用多层次聚类思路：先进行粗粒度欧式聚类，再对尺寸异常的候选簇分析其内部法向量分布直方图或反射强度梯度，若存在明显的双峰结构，则进一步做二次切分。

另一条路线是直接采用深度图分割，通过二维网格上的连通性和几何一致性判据替代三维邻域搜索，将聚类复杂度压缩到接近线性，从而实现毫秒级在线处理[7]。需要指出的是，当系统目标从单帧检测扩展到多帧一致的目标实例时，时序信息会成为提升稳定性的关键线索，例如速度一致性与轨迹平滑。HELD D等[17]在RSS 2016中提出了融合空间、时间与语义线索的实时3D分割概率框架，如图3所示，为利用时序减弱分割抖动提供了代表性思路，不过该工作更偏向实例分割，而非传统意义上的聚类。这类方法表明时序约束能够显著提升三维感知结果的稳定性，但其研究重点已经从单帧聚类转向时空联合建模[18]。

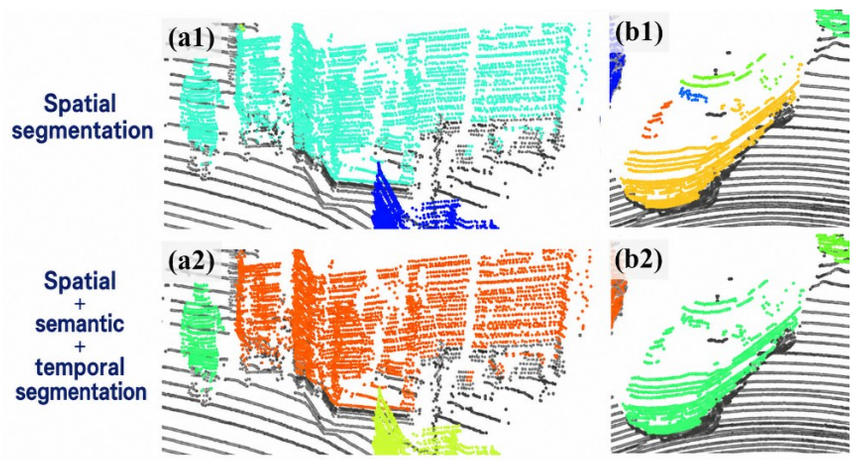

图3 融合空间、语义与时序多维线索的动态目标分割效果对比。(a1)仅空间特征：行人与建筑粘连（欠分割）；(a2)融合多维特征：成功剥离并独立分割行人；(b1)仅空间特征：车辆断裂成多个碎片（过分割）；(b2)融合多维特征：碎片正确合并为完整车辆

Fig.3 Comparison of dynamic object segmentation integrating spatial, semantic, and temporal multi-dimensional cues. (a1) Spatial features only: pedestrian merged with building (under-segmentation); (a2) Integrated multi-dimensional features: pedestrian successfully isolated; (b1) Spatial features only: vehicle fragmented into multiple pieces (over-segmentation); (b2) Integrated multi-dimensional features: fragments correctly merged into a complete vehicle

1.3 障碍物团簇分类

检测流程的最后一个环节是将无语义的点云簇转化为带有参数描述和类别标签的目标输出。这一过程通常包含两个部分：一是几何参数估计，即输出3D包围盒参数；二是类别判定，即输出目标类别。

在几何参数估计方面，受遮挡影响，激光雷达通常只能观测到物体朝向传感器的一侧，因而点云常表现为“L”型或“I”型[19]。由于只能获得部分轮廓，直接计算几何中心或最大外接边界的方法，如AABB[20]，往往会产生明显偏差。针对车辆等刚体目标，ZHANG Xiao等人[21]提出的L-shape fitting算法是基于几何先验的激光点云车辆目标二维轮廓拟合方向的代表性工作。该方法利用车辆目标水平俯视图近似矩形的几何先验，构建基于旋转角遍历的搜索式优化框架：先对目标点云提取凸包，再遍历所有可能的旋转角度，将点云投影到对应正交主轴坐标系中，最终依据面积最小化、投影方差最小化、点到矩形边缘距离最小化三大准则，筛选最优的凸包外接矩形。图4展示了L-shape拟合的典型结果与准则对比，该方法能够在车辆点云仅部分可见的场景下，恢复出合理的二维外接轮廓，验证了几何先验在目标轮廓参数估计中的有效性。

在类别判定方面，深度学习普及之前的主流方法主要依赖手工设计的特征描述子与传统机器学习分类

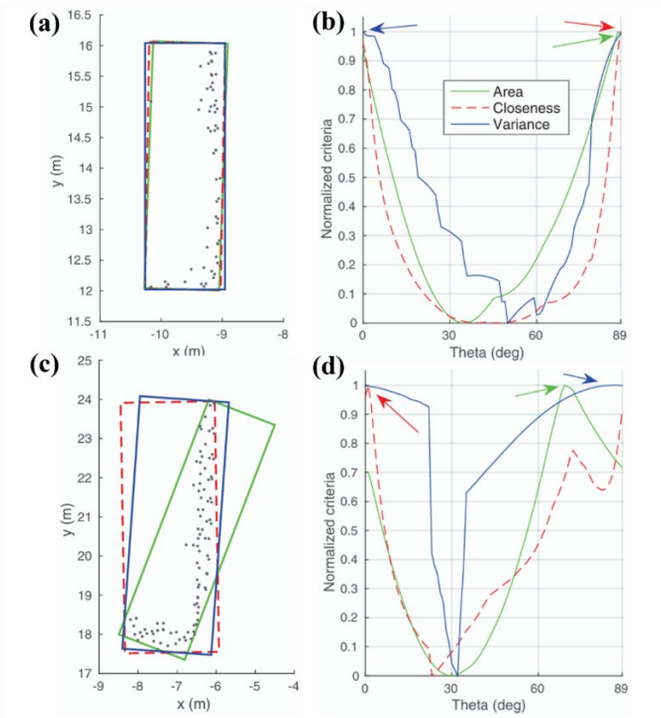

图4　针对刚体车辆目标的L-shape矩形拟合示例[21]。(a)三种准则拟合结果一致的车辆点云拟合结果；(b)示例(a)中三种准则随搜索角度变化的归一化得分曲线；(c)面积准则与另外两种准则拟合结果存在明显差异的车辆点云拟合结果；(d)示例(c)中三种准则随搜索角度变化的归一化得分曲线

Fig.4　L-shape rectangle fitting examples for rigid vehicle targets[21]. (a) Vehicle point cloud fitting result with consistent results under the three criteria; (b) normalized score curves of the three criteria with respect to the searched angle in example (a); (c) vehicle point cloud fitting result with a clear discrepancy between the area criterion and the other two criteria; (d) normalized score curves of the three criteria with respect to the searched angle in example (c)

器。特征工程的目标在于构造对平移和旋转相对稳定的几何描述。RUSU R B等人[22]提出的快速点特征直方图(Fast Point Feature Histograms，FPFH)通过统计查询点与其邻域点之间法向量夹角差异，构造多维直方图描述局部曲率特征[23]。视点特征直方图(Viewpoint Feature Histogram，VFH)在此基础上引入视点方向分量，使特征对位姿更敏感。形状函数集合(Ensemble of Shape Functions，ESF)则通过随机采样点对的距离、角度和面积分布，构造无需法向量计算的全局形状描述子。提取到的特征向量通常输入支持向量机(Support Vector Machine，SVM)或随机森林(Random Forest)进行分类。TEICHMAN A等人[24]提出的半监督学习框架利用跟踪轨迹中的时序信息(Boosting on Tracks)对单帧分类结果进行平滑，从而提升了分类准确率。虽然这些方法在复杂语义区分上不如深度神经网络，但其关于几何描述、特征设计与时序平滑的思想，为后续学习式方法提供了重要参照。表1对地面分割、空间聚类、轮廓拟合和特征描述等代表性算法进行了归纳。

综合来看，传统几何方法的优势在于流程清楚、结果可解释以及计算需求较低，适合规则场景或算力受限的工程系统。但这类方法通常需要预设地面模型、距离阈值、目标形状等先验条件，参数也往往依赖具体传感器和场景。当地面起伏、点云密度变化、遮挡和非凸目标同时存在时，固定规则难以保持稳定效果。此外，传统流程多采用串行结构，前端分割或聚类的误差会继续影响后续轮廓拟合和类别判断。为减轻人工规则和分阶段误差传递带来的影响，后续研究逐渐转向深度学习方法，通过可学习特征表达将点云组织、特征提取、候选生成和边界框回归纳入统一检测框架。第2章将围绕深度学习三维目标检测方法展开讨论。

2　基于深度学习的三维目标检测

三维目标检测的目标是从点云中估计目标的三维有向边界框及类别置信度。对于每个目标，其三维框可形式化表示为：

表1 典型三维点云处理几何算法性能对比

Table 1 Performance comparison of typical geometric algorithms for 3D point cloud processing

Method	Characteristics	Reference
GPF	Fast, but assumes planar ground.	[6]
Range Image	Low complexity; suitable for online processing.	[8]
GPR	Handles uneven terrain with higher computation.	[9]
CSF	Few parameters; adaptable to complex terrain.	[11]
Euclidean Clustering	Simple, but sensitive to thresholds.	[13]
DBSCAN	Robust to outliers, but inefficient online.	[15]
L-shape Fitting	Handles partial occlusion, but relies on priors.	[21]
FPFH	Describes local geometry with pose stability.	[22]

$$b=\left(x,y,z,l,w,h,\theta\right) \tag{3}$$

其中$\left(x,y,z\right)$为三维位置，(l,w,h)为尺寸，θ为航向角；检测输出为：

$$\{b_i,\ s_i,\ c_i\}_{i=1}^{N} \tag{4}$$

其中，b_i、s_i和c_i分别对应第i个目标的框参数、置信度与类别标签。评测时，预测框$\hat{b}$与真值框b的匹配通常以三维交并比为依据，其定义为：

$$IoU\left(b,\hat{b}\right)=\frac{V\left(b\cap\hat{b}\right)}{V\left(b\cup\hat{b}\right)} \tag{5}$$

式中，$V(\cdot)$表示三维体积。不同数据集在IoU阈值、类别设置和统计方式上不完全一致，因此比较不同方法的性能时，需要结合具体评测协议加以说明[25]。同时，受点云稀疏、密度分布不均以及遮挡造成局部信息缺失等因素影响，三维目标检测既需要恢复目标的几何边界，又需要保证类别判断的准确性，因此深度学习方法逐渐成为该领域的主要研究方向。现有研究可从输入表示、特征提取、模型架构和检测策略四个方面进行梳理[26]。其中，输入表示主要包括点表示、体素表示、柱状表示和距离视图表示；特征提取方法涵盖逐点特征映射、层次化邻域聚合、稀疏卷积、二维卷积骨干网络以及跨表示特征交互等方式；模型架构可分为点基、体素基、Pillar与BEV基、距离视图基和混合表示方法；检测策略则主要包括单阶段检测和两阶段检测，其中中心点方法通常归入单阶段检测范式。图5对上述关键模块与分类关系进行了归纳，并对应后文关于表示路线与检测范式的展开顺序。

2.1 传统方法到深度学习方法的过渡

传统点云处理通常采用滤波、地面分割、聚类、几何拟合和规则分类等串行步骤。其中，聚类的作用是根据空间邻近关系和局部密度，将非地面点划分为不同候选目标。以密度聚类DBSCAN为代表的方法能够在噪声存在时识别不同形状的点簇，但其性能对参数设置和点云密度变化较为敏感。尤其在自动驾驶场景中，点云常呈现近处稠密、远处稀疏、遮挡严重以及类别分布不均等特点，容易使前一阶段的误差继续传递到后续处理过程。且传统方法往往依赖多级阈值和经验规则完成候选提取与类别判断，当场景条件发生变化时，检测结果容易出现召回率下降和定位偏差增大的问题。深度学习方法通过可学习的特征表达和检测结构，降低了对经验参数的依赖。优势在于能够将点云组织、特征提取、候选生成和边界框回归放在同一框架下进行训练，在一定程度上减轻分割和聚类误差不断累积带来的影响。在实际应用中，激光雷达点云通常面临数据规模大、噪声来源复杂、跨帧配准困难以及实时计算压力较大等问题，这也推动了相关方法向更高效的计算结构发展，例如利用稀疏卷积减少无效计算、利用鸟瞰表示复用二维卷积网络，以及采用计算链路更短的单阶段和中心点检测方法。

2.2 表示路线：Point、Voxel稀疏卷积、Pillar与BEV、Range-view与Hybrid

输入表示直接影响三维检测模型如何组织点云数据与提取特征，以及后续计算过程的复杂程度，是模型设计中的关键环节。现有方法主要围绕几何细节保留、稀疏数据利用和实际部署效率等方面展开。点表示（Point-based）能够较好保留原始点云的细粒度几何信息，但邻域构建和特征聚合的计算开销较大；Voxel

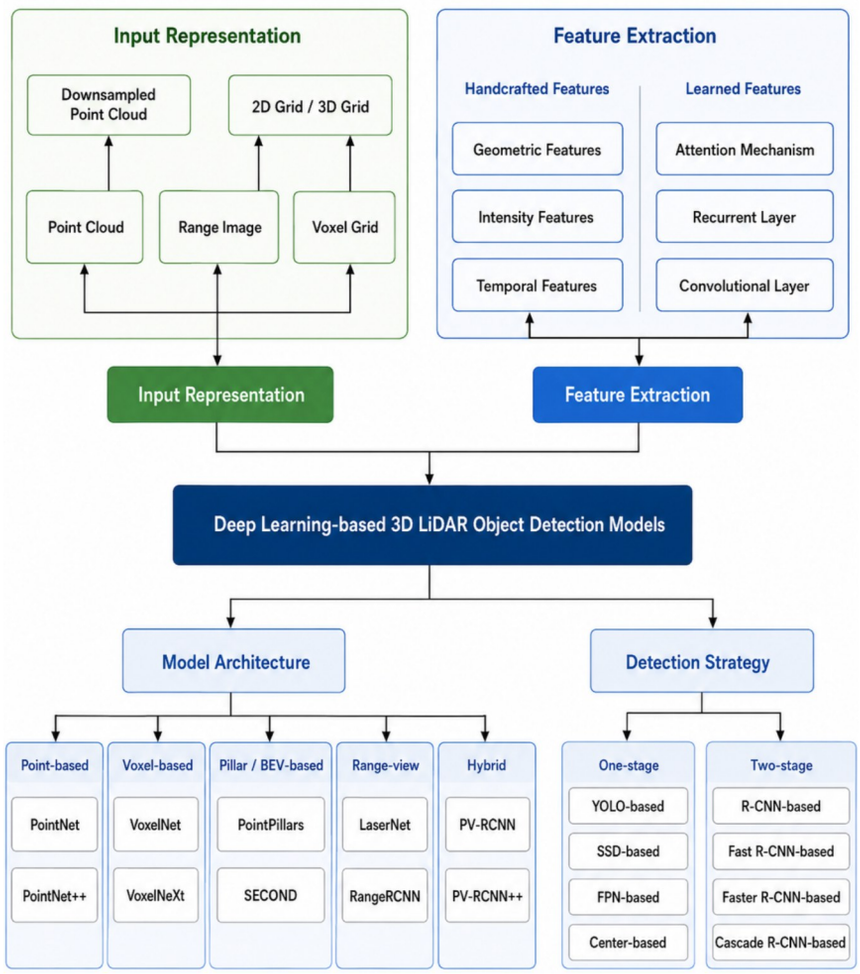

图5　基于深度学习的三维激光雷达目标检测关键模块与分类框架

Fig.5　Key modules and taxonomy of deep learning-based 3D LiDAR object detection methods

稀疏卷积方法通过结构化表示提高了计算效率；Pillar和鸟瞰图（Bird′s-Eye-View，BEV）方法将三维点云压缩到二维平面，便于复用成熟的二维卷积网络，从而更适合实时处理；Range-view方法利用激光雷达的天然投影形式进行特征提取，计算效率较高，但对遮挡和投影畸变更为敏感；Hybrid方法则结合体素表示的上下文信息与点表示的局部细节，以进一步提升检测精度。下文将围绕这些表示方式的基本思路、适用场景以及在精度和实时性方面的特点展开介绍[27]。

2.2.1　点表示三维检测：PointNet系列与代表性检测框架

点表示方法直接在原始点集上学习特征，避免体素化与投影带来的量化损失，因此对几何细节的保真度较高。其核心难点在于点集的无序性与邻域结构建模：模型既要对点的排列顺序保持不变性，又要在局部邻域内聚合几何上下文。PointNet通过逐点共享映射与对称聚合实现对无序点集的置换不变表征[28]，可将其视为对集合函数$f(\cdot)$的对称近似：

$$f(x_1,x_2,\cdots,x_n)=\gamma\left(\underset{i=1,\cdots,n}{\mathrm{MAX}}\{h(x_i)\}\right) \tag{6}$$

式中，$\gamma(\cdot)$为后续非线性映射，MAX为对称聚合算子，$h(\cdot)$为逐点共享映射。该结构保证网络对输入点排列的置换不变性，并将逐点特征聚合为全局表示，为后续点云分类、分割与检测任务提供基础表征能力。PointNet++在此基础上引入分层采样与局部邻域抽象，增强对局部结构和非均匀采样密度的建模能力，为后续检测中的前景点提取、关键点抽象和候选区域特征汇聚提供了基础算子[29]。

为避免特征变换网络学习到退化映射，PointNet在训练中对变换矩阵引入近似正交约束：

$$L_{\mathrm{reg}}=\|\mathrm{I}-\mathrm{A}A^{T}\|_F^2 \tag{7}$$

式中，L_{reg}为特征变换网络的正则化损失，A为特征变换矩阵，$\|\cdot\|_F$为Frobenius范数，该正则项促使A接近正交矩阵，从而稳定特征对齐过程并提升可靠性。

在三维检测中，基于点表示的检测方法主要分为两阶段和单阶段两类。两阶段方法先通过点级前景分割或候选框生成缩小搜索范围，再在候选区域内聚合特征并精炼边界框；单阶段方法则省去候选框精炼环节，直接由点特征回归目标类别和位置参数，以降低推理时延。PointRCNN是两阶段点表示检测方法的代表。第一阶段通过前景点分割筛选候选点，并由前景点生成三维候选框。由于前景点数量远少于背景点，训练中容易出现类别不均衡，因此该阶段引入焦点损失[30]，对易分样本降低权重，对困难样本提高权重，其形式为：

$$L_{\mathrm{focal}}(p_t)=-\alpha_t(1-p_t)^{\gamma}\log(p_t) \tag{8}$$

式中，$L_{\mathrm{focal}}(p_t)$表示焦点损失，p_t表示对真实类别的预测概率，α_t与γ控制类别权重与难样本强调程度，该损失有助于提升前景点召回并稳定后续候选生成。第二阶段中，PointRCNN将候选框内点集变换到规范坐标系，结合语义特征进行边界框精炼，从而提高定位质量和候选排序稳定性。3DSSD是单阶段点表示检测方法的代表[31]。该方法通过检测导向采样减少冗余点，在精度和效率之间取得较好平衡。总体而言，点表示路线能够较好保留原始点云的几何细节，但邻域查询、采样和局部聚合会带来较高计算开销，对点数规模和实现效率要求较高。因此，在大场景和强实时约束下，点表示方法通常需要通过控制候选数量、简化局部聚合过程，并结合BEV表示或稀疏体素表示，以降低计算代价。

2.2.2 Voxel稀疏卷积：SparseConv主干与可学习稀疏性

体素表示通过将不规则点云划分为规则三维网格，使点云数据能够以更适合卷积运算的方式输入网络。与直接在原始点集上进行邻域查询和特征聚合相比，体素化表示在保持空间结构信息的同时，更便于构建层次化特征提取过程。VoxelNet、SECOND、CenterPoint和Voxel R-CNN等方法表明，体素路线在检测精度与工程可实现性之间能够取得较好的平衡，因此已成为三维目标检测中的重要技术路线。

稀疏嵌入卷积检测（Sparsely Embedded Convolutional Detection，SECOND）是体素路线中的代表性方法，其基本流程由体素特征编码、稀疏卷积主干和检测头组成[32]。与普通三维卷积在整个规则网格上统一计算不同，SECOND仅在非空体素位置执行卷积运算，从而有效减少空体素带来的无效计算，显著提高训练和推理效率。在SECOND的整体检测流程中，点云首先被划分为体素并编码为体素特征，随后经过稀疏卷积主干提取空间特征，再送入检测头完成类别预测、边界框回归和方向估计。

为说明SECOND中稀疏卷积的计算方式，可将其写成基于输入—输出索引规则的求和形式：

$$Y'_{j,m}=\sum_{k}\sum_{l}W_{k,l,m}\tilde{D}'_{R_{k,j},k,l} \tag{9}$$

式中$R_{k,j}$表示在卷积核偏移k下，输出位置j对应的有效输入位置索引，$\tilde{D}'$为按规则收集的稀疏输入特征，$W_{k,l,m}$为卷积核参数。该表达式仅对规则集合中存在的有效对应关系累加，从而跳过空体素带来的空计算，提高稀疏卷积的吞吐效率。

在此基础上，体素路线的研究重点逐渐由局部稀疏卷积特征提取，转向稀疏表示条件下的全局上下文建模。动态聚类Transformer（Dynamic Clustering Transformer，DCT）认为，自动驾驶场景中的激光雷达点云具有较强的空间可分离性，固定窗口内的局部注意力难以充分描述目标之间的空间关系，因此通过动态聚类以及聚类token（cluster token）与体素token（voxel token）之间的交互，增强全局特征与局部体素特征的信息传播[33]。RVSA-3D（Voxel-based Fully Sparse Attention 3D Object Detection for Rail Transit Obstacle Perception）则面向铁路远距离稀疏障碍物检测，提出全稀疏注意力（fully sparse attention）框架，通过稀疏注意力驱动的下采样和跨尺度特征融合，提高远距离小目标的感知能力[34]。这些研究表明，体素检测方法的改进已不再局限于减少空体素计算，而是进一步关注在稀疏计算框架下增强长程依赖建模能力。

除了增强全局建模能力，另一类改进思路是进一步选择优化活跃体素。焦点稀疏卷积（Focal Sparse Convolution）指出，常规稀疏卷积若对所有激活体素统一扩展感受野，容易引入大量背景冗余并增加计算开销；若始终严格限制活跃位置，又可能削弱跨体素的信息传递。为减小这类无效扩张带来的负担，该方法首先学习各空间位置的重要性，再将计算集中到更有信息量的区域，并据此从输入活跃位置集合P_{in}中筛选“重

要输入”子集P_{im}：

$$P_{im}=\{p\,|\,I_0^p\geqslant\tau,p\in P_{in}\}\tag{10}$$

式中，I_0^p表示位置p处重要性立方体的中心响应，τ为阈值。P_{in}为输入活跃位置集合，P_{im}为筛选后的重要输入子集，该方法优先选出更值得保留的位置，再围绕这些位置构造动态输出形状，从而把计算资源优先分配给更关键的区域[35]。图6展示了焦点稀疏卷积的关键模块。图中先根据输入特征预测空间位置的重要性，再生成动态输出形状，并据此进行后续稀疏卷积计算。与SECOND相比，这种做法不只是跳过空体素，而是进一步对活跃区域进行筛选，因此更强调把计算集中到前景目标和周围的有效空间位置。

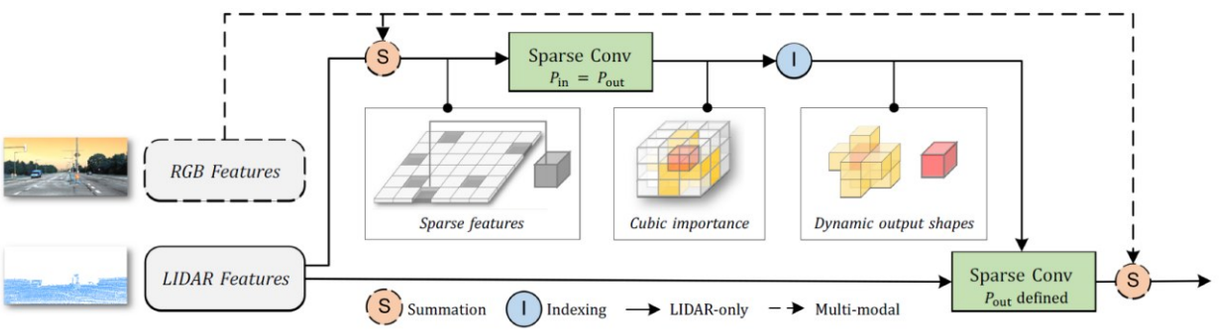

图6　焦点稀疏卷积关键模块示意[35]

Fig.6　Schematic diagram of the key modules of Focal Sparse Convolution[35]

体素路线的效果和效率主要受体素分辨率、有效体素数量和稀疏特征分布变化的共同影响。更细的体素有助于保留小目标边界和远距离几何细节，但也会显著增加有效体素数量，提高计算负担。可学习稀疏性则在一定程度上缓解了这一矛盾，使计算更多集中到对检测更有帮助的区域。因此体素化与稀疏卷积方法能够在保持较强空间建模能力并兼顾一定的工程效率。

2.2.3　Pillar与BEV表示：二维化压缩与密集预测

BEV路线的基本思路是沿高度方向对点云进行压缩，将原本的三维稀疏点云表示转化为鸟瞰平面上的二维伪图像，再利用成熟的二维卷积网络完成特征提取与目标预测。与直接在三维空间中进行卷积相比，这种表示方式更容易控制计算量，因此更有利于实时检测。PIXOR直接以BEV表示作为输入进行单阶段密集预测，体现了将三维问题转到二维平面后在计算效率上的优势。而PointPillars[36]通过柱状体替代体素，在鸟瞰平面离散后对每个Pillar内的点进行特征编码，再将编码结果回填到BEV伪图像中，形成伪图像并输入二维卷积网络完成检测。

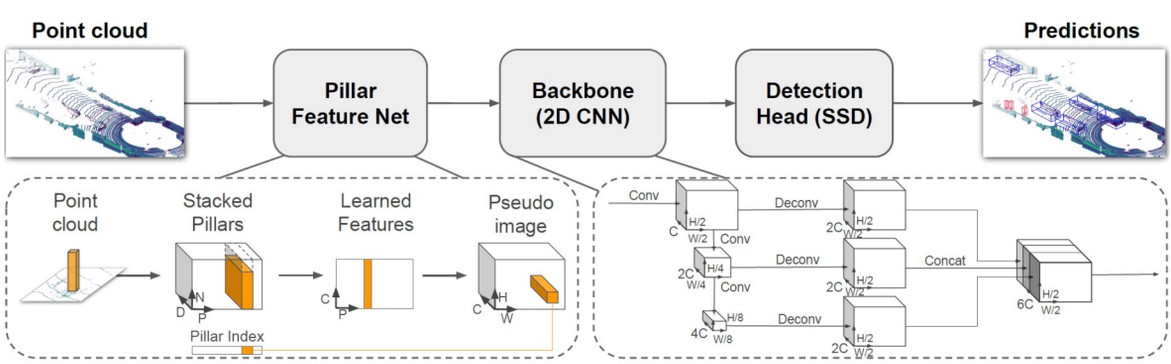

图7　PointPillars从Pillar编码到BEV伪图像的整体结构[36]

Fig.7　Overall architecture of PointPillars from pillar encoding to BEV pseudo-image[36]

PointPillars的整体流程如图7所示，该方法先将点云离散为柱状单元，对非空pillar内的点进行9维特征编码，再将柱特征回填到BEV平面形成伪图像，并输入二维卷积网络完成检测。实现时，模型限制每帧非空pillar数量和每个pillar内采样点数，从而构造固定尺寸的柱特征张量。编码后的柱特征再被回填到鸟瞰平

面的二维栅格中，形成可供二维卷积网络处理的伪图像。其点特征装饰方式和张量形状可写为：

$$\begin{cases} \hat{l}=\left[x, y, z, r, x_c, y_c, z_c, x_p, y_p\right] \in R^D, D=9 \\ T \in R^{D \times P \times N}, I \in R^{C \times H \times W} \end{cases} \tag{11}$$

式中，D为单点装饰维度，P与N分别为每帧非空Pillar数与每Pillar采样点数的上限，编码后的Pillar特征被回填到二维画布形成$(C|H|W)$的伪图像输入，从而将三维稀疏计算转化为高效的二维检测流程。

Pillar与BEV路线在垂向信息保留和细粒度特征表达方面存在不足，近年来相关研究开始向垂向信息补偿、特征增强和结构扩展等方面持续改进。分层垂向先验（Stratified Vertical Priors，SVP）从垂向先验恢复的角度出发，利用高度图分支与伪图像分支并行建模，并通过共享信息融合模块在多尺度上将vertical priors注入水平特征，从而缓解pillar表示中过度压缩垂向信息的问题[37]。SCNet3D则指出，pillar编码中简单的最大池化和浅层伪图像特征提取容易导致关键信息损失，尤其不利于小目标检测，因此通过特征增强模块和双分支SCNet提升柱级三维特征与BEV特征的表达能力[38]。此外，3DPillars进一步指出，传统PointPillars类方法不仅存在三维结构表达不足的问题，而且难以自然引入基于三维候选框的两阶段精炼流程，因此提出多轴伪图像堆叠与稀疏场景上下文特征模块，使pillar路线也能够较好支持两阶段检测[39]。

Pillar与BEV路线的发展重点由单纯追求高效率，逐渐转向在保持二维卷积高效性的同时补偿三维结构损失。因此这一路线适合实时检测和工程部署，但其检测性能在很大程度上取决于垂向信息恢复能力、柱级特征表达质量以及远距离小目标感知能力。

2.2.4 Range－view与Hybrid表示：投影处理与多表示融合

Range-view将点云投影为球面距离图或深度图，使三维点云能够在二维网格上进行卷积处理，具有输入紧凑、前向高效的特点。LaserNet是该路线中的代表方法之一，其在Range-view表示的基础上进行概率式三维目标检测，通过预测分布并结合后处理得到最终检测框[40]，此类方法利用二维投影降低了特征提取难度，在计算效率方面具有一定优势。

但Range投影也会带来一定的信息损失，由于三维点云被压缩到二维平面，不同距离处目标的尺度差异、遮挡关系和投影扭曲问题会更加突出，容易导致局部几何信息保留不足和空间一致性下降。因此，该类方法通常需要结合更强的几何建模、后处理优化，以缓解单一投影表示的局限。

混合表示同时利用体素表示较强的空间上下文建模能力和点表示较好的局部细节保留能力，因此逐渐成为兼顾检测精度与计算效率的重要方向。点-体素区域卷积神经网络（Point-Voxel Region-based Convolutional Neural Network，PV-RCNN）是其中较有代表性的方法，先通过稀疏体素卷积提取全局空间特征并生成候选区域（Region of Interest，RoI），再通过关键点集合抽象与RoI池化策略（RoI-grid pooling）进行点级精炼，从而兼顾检测精度和计算效率[41]。在PV-RCNN的结构里Hybrid路线是把体素表示用于候选生成，把点级特征用于候选区域精炼，发挥二者各自的优势。但也带来了网络结构复杂、关键点采样和RoI特征聚合调参成本较高等问题。

在此基础上，混合表示方法的改进逐渐从点体素特征融合扩展到体素全局建模、关键点特征学习和后处理优化。PV-RCNN、PV-RCNN＋＋等方法表明，体素特征适合用于候选框生成，点级特征适合用于候选区域精细回归，二者结合能够兼顾场景上下文和局部几何细节。HP-PV-RCNN在PV-RCNN＋＋基础上引入线性角注意力机制和Mamba2状态空间模型，以增强非空体素的局部建模和全局特征提取能力；同时利用Kolmogorov-Arnold网络（Kolmogorov-Arnold Network，KAN）改进关键点特征学习，并采用模糊非极大值抑制（Fuzzy Non-Maximum Suppression，Fuzzy-NMS）优化后处理结果[42]。这说明混合表示方法已不再停留于特征拼接，而是开始向特征建模、关键点学习和检测后处理的协同优化发展。

2.3 检测结构与训练机制扩展

2.3.1 两阶段、单阶段与中心点方法

上文已从输入表示角度梳理了点表示、体素稀疏卷积、Pillar与BEV、距离视图和混合表示等路线。本节进一步从候选组织方式和推理流程出发，比较两阶段、单阶段和中心点方法在检测精度、计算代价与实时性之间的差异。

两阶段方法通常先从全场景点云中生成少量候选区域，再在候选区域内进行特征聚合与边界框精炼，

因此在遮挡较强、目标边界复杂或候选之间干扰明显的场景中往往具有较好的定位能力[30,43]。不过，这类方法一般需要经历候选生成、区域特征提取和边界框精炼等多个环节，整体流程较长，训练和实现也相对复杂，其最终性能还容易受到候选质量、特征提取方式以及后处理步骤的影响。

与两阶段方法相比，单阶段方法直接在特征图上完成目标分类和边界框预测，不再显式区分候选生成和后续精炼，因此整体流程更紧凑，也更适合实时检测任务。其优势在于检测链路较短、并行效率较高。但与此同时，这类方法对特征表达质量、标签分配方式和后处理策略也更为依赖。在目标密集、遮挡明显或远距离点云稀疏等复杂场景下，若特征编码能力不足，模型更容易出现误检、漏检以及目标区分不清等问题。3DSSD、PointPillars 和 PIXOR 分别代表了点表示、柱状表示和鸟瞰表示下的单阶段检测的常见实现方式[31,36,44]。

中心点范式（Center-based）可视为两阶段与单阶段之间的中间路线。其不再依赖大规模锚框枚举，而是以目标中心点热力图组织候选，再回归尺寸、偏移、朝向等属性，从而在降低匹配与枚举成本的同时保持较好的检测精度。CenterPoint是这一类方法的代表，其整体流程如图8所示。

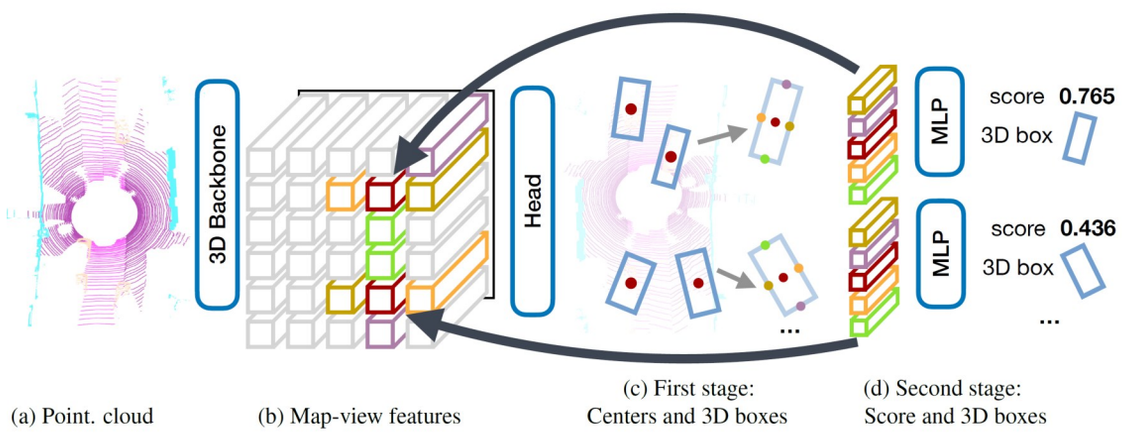

图8　CenterPoint中心点检测与二阶段评分[45]

Fig.8　CenterPoint center-based detection and two-stage scoring[45]

首先，点云经过三维主干网络提取特征，并转换为鸟瞰图特征。第一阶段通过中心点热力图定位候选，并在中心位置回归目标属性。第二阶段进一步引入*IoU*引导的置信度学习，以缓解分类得分与定位质量不一致的问题[45]。其中，候选框与真值框的三维*IoU*被映射为可学习的置信度目标，其定义如式(12)所示：

$$I_t = \min\left(1, \max\left(0, 2 \cdot \mathrm{IoU}_t - 0.5\right)\right) \tag{12}$$

式中，I_t表示由候选框与真值框三维*IoU*映射得到的置信度目标。这样做的目的，是使第二阶段学习到与定位质量更接近的评分信号，而不只依赖第一阶段的分类热力图得分。进一步地，将第一阶段热力图得分与第二阶段*IoU*置信度进行融合得到最终分数，如式(13)所示：

$$\begin{cases} \widehat{Q_t} = \sqrt{\widehat{Y_t} \cdot \widehat{I_t}} \\ \widehat{Y_t} = \max_{0 \leqslant k \leqslant K} \widehat{Y_{p,k}} \end{cases} \tag{13}$$

式中，$\hat{Y}_t$为第一阶段中心点热力图在候选位置的得分，$\hat{I}_t$为第二阶段预测的*IoU*引导置信度，几何平均用于兼顾分类置信与定位质量[45]。这种候选组织与评分方式降低了传统锚框枚举与匹配成本，但较为依赖鸟瞰表示质量及中心点监督的可靠性。

2.3.2　锚框样本分配与训练目标优化

除输入表示和检测骨干外，训练样本的划分也会影响三维目标检测效果，在锚框式检测器中这一问题更明显。现有许多锚框式方法通常借用二维检测中的做法，根据锚框与真实框之间的*IoU*值，把样本划分为正样本、负样本和忽略样本。但激光雷达点云本身存在稀疏、分布不均和局部缺失等问题，即使两个锚框的*IoU*接近，它们覆盖到的目标点数量和目标信息完整程度也可能不同。因此，只依靠*IoU*进行样本划分，容易造成训练目标不够明确。针对这一问题，点辅助样本选择方法（Point Assisted Sample Selection，PASS）

引入了基于点云分布的点级交并比 IoU_{point}，用来补充衡量锚框样本质量，并改进正负样本划分。在不增加额外模型参数和推理时间的情况下，该方法提高了锚框式检测器的训练样本质量[46]。这表明，合理的样本划分和明确的训练目标同样会影响三维目标检测性能。

2.3.3 低标注成本检测

随着大规模三维数据集的广泛应用，三维边界框高成本标注逐渐成为制约三维目标检测实际部署的重要问题。与二维检测相比，三维框标注不仅需要确定目标中心位置，还需同时标注尺寸、朝向和空间范围，因此人工成本高、耗时长。为缓解这一问题，近年来研究逐渐从全监督训练扩展到半监督、稀疏监督和点击监督等低标注成本设定，试图在显著减少人工标注负担的同时保持较高检测性能。SC3D是这一方向的代表性工作之一，该方法仅要求在每帧点云的BEV平面上提供一个粗点击标注，通过时序点云聚合判断被点击实例的运动状态，并分别为静态目标构建框级伪标签、为动态目标构建点级语义伪标签，随后再利用混合监督教师—学生框架将有限点击提示扩展到未点击实例，从而在极低标注成本下实现有竞争力的三维目标检测性能[47]。这说明当前三维目标检测研究的关注点已不再局限于如何设计更强的检测网络，还包括如何利用更经济、更可扩展的监督方式构建有效训练约束。

表2 典型深度学习三维目标检测方法性能对比

Table 2 Performance comparison of typical deep learning methods for 3D object detection

Category	Method	Main result	Reference
Point－based two－stage	PointRCNN	KITTI val Car 3D AP: 88.88/78.63/77.38	[30]
Point－based one－stage	3DSSD	KITTI test Car 3D AP: 88.36/79.57/74.55; 25+ FPS	[31]
Sparse voxel convolution	SECOND	KITTI test Car 3D AP: 83.13/73.66/66.20; 20 Hz	[32]
SparseConv enhancement	Focals Conv	KITTI val Car 3D AP(R40): 92.32/85.19/82.62; 112 ms	[35]
Pillar and BEV	PointPillars	KITTI test Car 3D AP: 79.05/74.99/68.30; 62 Hz	[36]
Range－view	LaserNet	ATG4D Vehicle BEV AP: 85.34; KITTI runtime: 30 ms	[40]
Point－voxel hybrid	PV－RCNN	KITTI test Car 3D AP: 90.25/81.43/76.82	[41]
Center－based	CenterPoint	nuScenes test: 58.0 mAP/65.5 NDS	[45]
Sample assignment	PASS	KITTI val BEV mAP gain: 0.30 - 1.38	[46]
Low－cost annotation	SCNet3D	KITTI test mAP(Mod.): 64.85; single－click labels	[47]

表2归纳了本章讨论的典型深度学习三维目标检测方法。从表示方式看，点表示、体素表示、Pillar与BEV表示、Range-view表示和Hybrid表示并不是简单的性能递进关系，而是在几何细节保留、计算效率和工程部署复杂度方面形成不同取舍。点表示保留细粒度结构，但邻域搜索和局部聚合代价较高；体素稀疏卷积增强了空间结构建模，但体素尺度选择会影响远距离小目标；Pillar与BEV方法在实时性上更具优势，但高度压缩会削弱垂向几何表达；Range-view方法输入紧凑，但投影畸变和遮挡敏感性较强；Hybrid方法通常精度较高，但结构复杂、复现和部署成本更高。因此，方法选择应结合目标尺度、场景稀疏性和算力约束要求综合判断。

尽管深度学习方法减少了传统几何方法对人工阈值和规则流程的依赖，但单一激光雷达仍存在信息不足的问题。远距离目标点数少，小目标和遮挡目标轮廓不完整，点云本身也缺少颜色、纹理等语义信息，因此在复杂场景中仍容易出现漏检和定位偏差。为补充点云几何信息的不足，后续研究开始引入相机图像，将图像语义信息与激光雷达空间信息结合。第3章将围绕激光雷达与相机融合方法展开讨论。

3 激光雷达与相机信息融合技术研究进展

3.1 单传感器局限与多传感器融合动机

目标感知方法的发展经历了由单传感器驱动向多传感器融合的演进过程。早期方法通常围绕单一传感器展开：基于相机的视觉感知方法依托丰富的纹理与语义信息，结合YOLO、R-CNN等神经网络方法在目标识别、场景理解等任务中取得了广泛应用[48]，基于激光雷达的感知方法则凭借直接三维测量能力，在距离估计、空间建模与几何结构恢复方面表现突出[49]。然而，单传感器方案在复杂环境下往往存在难以避免的性能瓶颈[50]。视觉方法易受光照变化、遮挡等外部变化影响[51]，激光雷达方法则受限于点云稀疏性、语义表达不足以及远距离细粒度目标辨识能力有限等问题[52]。随着感知任务从单一特殊环境扩展到室外开放场景，并进一步覆盖检测、分割、定位建图等复杂任务，依赖单一传感器已难以同时满足精度与实时性的综合要求。

在此背景下，激光雷达与相机融合作为多传感器感知中的一条重要技术路线，已逐渐成为相关研究的重点[53-54]。二者在几何信息与语义信息上的天然互补，使融合方法能够在语义理解、距离估计与空间位姿之间建立更有效的协同关系，并在自动驾驶[55]、机器人定位重构[56]、智能交通[57]等场景中体现出应用价值。现有研究围绕三个关键问题逐步展开：一是跨模态几何关系如何可靠建立（标定与对齐问题）；二是如何在统一空间参照下实现面向检测、分割等任务的稳定融合感知（检测与分割融合问题）；三是融合能力如何从单一感知任务扩展为全图定位建图的大规模检测任务。基于这一问题链条，相关研究可概括为“融合基础层（标定与自标定）—任务感知层（检测与分割融合）—能力支撑层（定位建图）”的依次递进的技术流程。

3.2 融合基础：跨模态对齐标定

激光雷达与相机的不同模态信息融合首先取决于跨模态几何关系的准确性。无论是将点云投影至图像平面获取语义特征，还是在统一表示空间进行跨模态特征交互，其前提均是传感器内外参已知并且时空对齐误差可控。因此，标定与对齐问题是决定融合性能上限的基础性步骤，围绕这一基础环节，相关研究路径也逐渐清晰：早期主要采用基于标定物的几何标定方法（target-based），随后发展出利用场景结构先验信息的无标定物自标定方法（targetless），并进一步演进到基于学习的端到端自标定方法（learning-based）。

3.2.1 基于标定物的外参标定方法

早期激光雷达与相机标定多采用基于标定物几何标定方法，即借助黑白棋盘格等人工标定物作为跨模态公共观测目标，如图9所示，通过在图像与点云中提取目标几何结构并构造约束求解外参。该类方法几何过程清晰、可解释性强，在受控条件下精度较高，因此长期作为多传感器融合系统的前置工程基础。

图9 黑白棋盘外参标定

Fig.9 Black-and-white checkerboard extrinsic parameter calibration

根据标定过程中采用的特征类型与几何约束形式，基于棋盘格的激光雷达与相机外参标定方法大致可以分为两类：一类是基于棋盘格角点的透视n点问题（Perspective-n-Point，PnP）类位姿求解方法，另一类是基于棋盘格平面几何约束的3D - 3D配准方法。

一方面，基于棋盘格角点的PnP类方法延续了传统相机标定的范式，即通过检测棋盘格角点建立标定

板三维点与图像二维像素点之间的对应关系，并在已知相机内参条件下求解相机位姿或投影关系。NAN Zongliang等[58]在融合检测系统的前置步骤中采用棋盘格（12×9）结合张正友标定法完成相机内参与畸变参数估计，并进一步结合P3P方法（n=3）求解激光雷达点云与图像之间的投影关系，用于后续点云投影、图像裁剪与联合目标检测流程，标定过程见图10。P3P本质上属于PnP问题的特例，因此该文的标定环节在方法学上仍可归为“基于棋盘格角点的PnP类标定路线”；不过其研究重点并不在标定算法本身，而在于为后续融合检测提供可靠的前置条件。

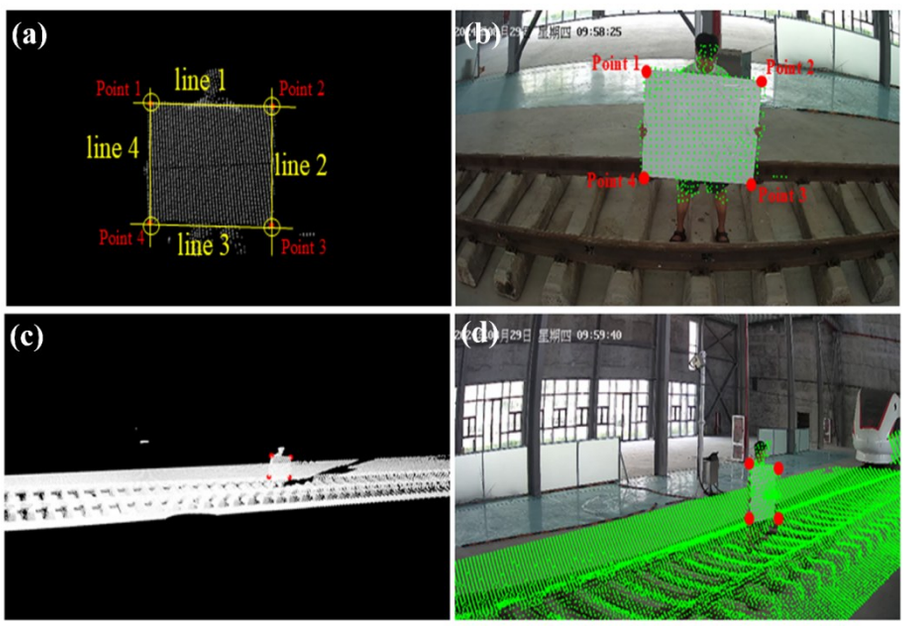

图10 基于标定板的图像与激光点云投影融合效果[58]。(a)和(b)代表局部点云投影情况；(c)和(d)代表全局点云投影情况，其中红色点代表四个投影角点

Fig.10 Calibration board-based projection fusion of image and LiDAR point cloud [58]. (a) and (b) represent the local point cloud projection situation; (c) and (d) represent the overall point cloud projection situation, where the red points represent the four projection corners

另一方面，考虑到激光雷达某些情况下的点云稀疏，难以稳定提取棋盘格精确角点，也有研究转而采用棋盘格平面几何约束构建3D-3D配准方法。SINGANDHUPE A等人[59]提出了一种基于平面配准的单帧激光雷达-双目相机外参标定方法。该方法先在点云和双目重建点云中分别提取棋盘格区域，并通过RANSAC拟合标定板平面；随后依据两侧平面方程分别构造固定数量的结构化三维点集，以减弱原始点云密度差异和点数不一致带来的影响；最后采用CoSM-ICP对两组结构化平面点集进行配准，直接估计激光雷达与双目相机之间的外参变换。与传统“角点检测+PnP”的2D-3D位姿求解范式相比，这类方法不依赖图像角点与点云角点的一一对应，而是将标定问题转换为平面几何驱动的3D-3D配准。

此外，在基于标定物方法内部，研究也在从纯几何配准逐步转向传感器机理建模和几何优化的联合建模思路。JEONG S等人[60]提出O^3激光雷达与相机外参标定方法：在单次观测、单个棋盘格标定板条件下，利用反光胶带强度自动定位标定板并用RANSAC拟合平面$w^{\top}x+w_0=0$，针对激光雷达沿光线方向测距噪声问题（由$t_i=\dfrac{w^{\top}\tilde{p}_i+w_0}{w^{\top}\left(\tilde{p}_i-c\right)}$得到投影点），引入沿光线的透视投影和虚拟增强点（如$q_i^l=p_i^l+\left(p_i^l-p_i^{l-1}\right)$）来增强局部几何结构，在此基础上，为将边界与角点等几何信息统一纳入配准求解过程，构建如下代价函数E：

$$E=E_{edge}+\lambda E_{position} \tag{14}$$

其中，E_{edge}为边界约束项，$E_{position}$为位置约束项，λ为权重系数。随后采用由粗到细的方式优化矩形边界与角点，建立棋盘格角点的3D-2D对应关系，再通过平移无关的旋转约束求得粗位姿并恢复平移尺度ρ：

$$\rho=\frac{\Sigma_{ii=1}^{m}\left(C_{f_i}\times R_L^C L_{f_i}\right)^T\left(C_{i_{CL}}\times C_{f_i}\right)}{\Sigma_{ii=1}^{m}\left(C_{i_{CL}}\times C_{f_i}\right)^T\left(C_{i_{CL}}\times C_{f_i}\right)} \tag{15}$$

其中C_{f_i}表示第i个相机图像中的归一化特征点，L_{f_i}表示第i个点云中的三维特征点，$C_{i_{CL}}$表示单位平移方向向量，R_L^C表示从LiDAR坐标系到相机坐标系的旋转矩阵。最终用PnP最小化重投影误差$min\sum_{i=1}^{m}//C_{f_i}-\pi\left(R_L^C L_{f_i}+C_{t_{CL}}\right)//^2$精化外参，从而在“一帧一板”条件下获得更高精度。具体流程见图11。

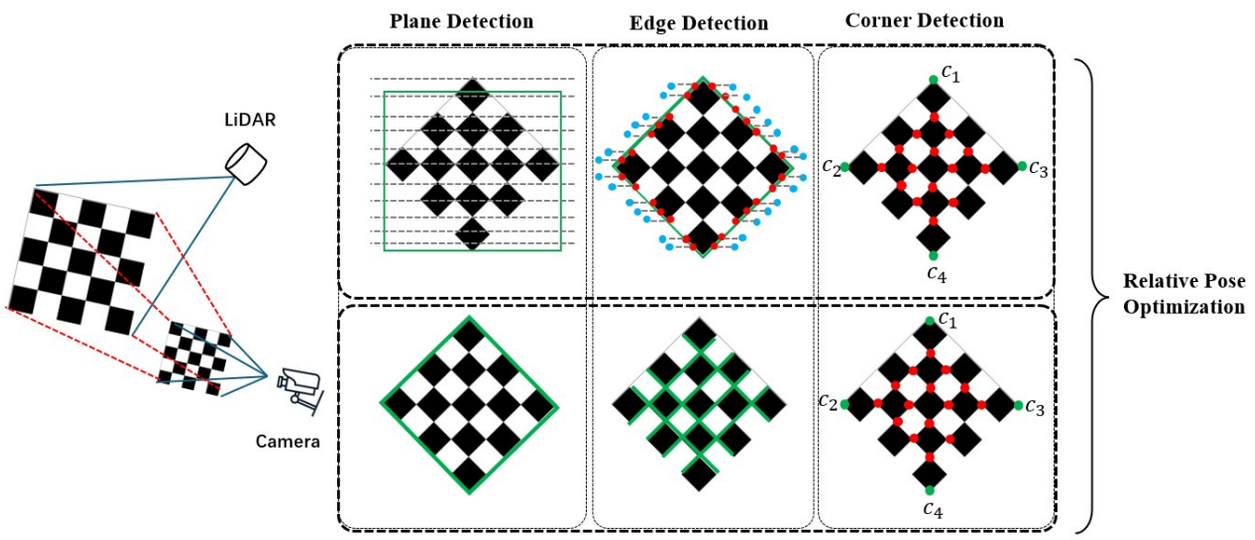

图11　O^3激光雷达-相机外参标定方法标定过程

Fig.11　O^3 LiDAR-Camera extrinsic calibration method calibration process

总体来看，基于黑白棋盘格的激光雷达与相机标定方法具有实现简便、几何约束明确和精度较高等优点，因此在多传感器融合系统中仍被广泛用作前置标定手段。但这类方法通常要求人工放置标定板并保证公共视场重叠，部分方法仍需要人工圈定标定板区域，或依赖噪声较小的图像与点云数据。在动态部署、现场快速维护与在线校正场景下，其自动化程度与适应性仍存在局限，这也推动后续研究转向基于场景先验的无标定物自标定方法和学习式在线标定。

3.2.2　基于场景先验的自标定方法

随着融合系统从实验环境走向真实部署场景，基于标定物方法在操作成本、现场适配性和重复标定便利性方面的不足逐渐显现，无标定物自标定方法因此成为重要的发展方向。场景先验自标定方法不依赖人工标定物，而是利用特定应用场景中稳定存在的几何结构，如线、面、边缘及其组合，构建跨模态约束，并通过优化求解激光雷达与相机之间的外参关系。其核心思路，是以场景固有结构替代人工标定目标所提供的几何先验。

HE Zhanyuan等人[61]提出了一种基于铁路场景固有几何特征的激光雷达与相机无标定板自标定方法。该方法首先将点云按中心投影生成反射强度图，并尝试使用尺度不变特征变换（Scale-Invariant Feature Transform，SIFT）与可见光图像进行点特征匹配，但由于跨模态差异与点云稀疏性，误匹配严重。图12展示了点特征匹配下的错误对齐。

因此，该方法转而提取线特征：在图像侧通过候选区域、边缘检测和霍夫直线检测提取待检测区域的二维线集合Q_c，在点云侧通过多帧累积、体素划分和RANSAC平面拟合，利用相邻平面的交线得到三维线集合Q_l，随后基于初始外参将3D线点按刚体变换与投影关系：

$$p_i=f(KP_i) \tag{16}$$

其中p_i表示第ii个3D线点投影到图像平面后的像素坐标，P_i表示第ii个三维点，K表示相机内参矩阵。随后重投影到图像平面并与Q_c做最近邻匹配，并定义重投影残差代价函数J：

$$J=\sum_{i=1}^{k}//f(KRTp_i)-q_i// \tag{17}$$

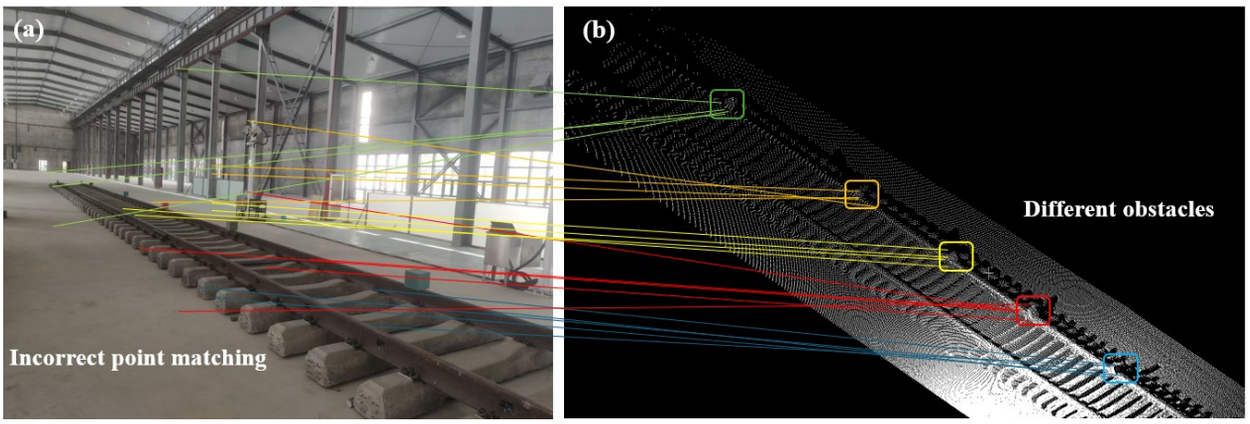

图 12 点特征匹配中的错误匹配示例。(a)真实场景图像;(b)点云场景图像;不同颜色方框表示需要对齐的障碍物区域

Fig.12 Examples of incorrect matches in point-feature matching. (a) Real-scene image; (b) point cloud scene; the differently colored boxes indicate obstacle regions to be aligned

其中k表示参与匹配的点的数量,通过迭代优化更新外参旋转矩阵与外参平移向量R,T得到最优外参矩阵,从而在无需人工标定板条件下完成激光雷达与相机外参自标定,匹配结果见图 13。

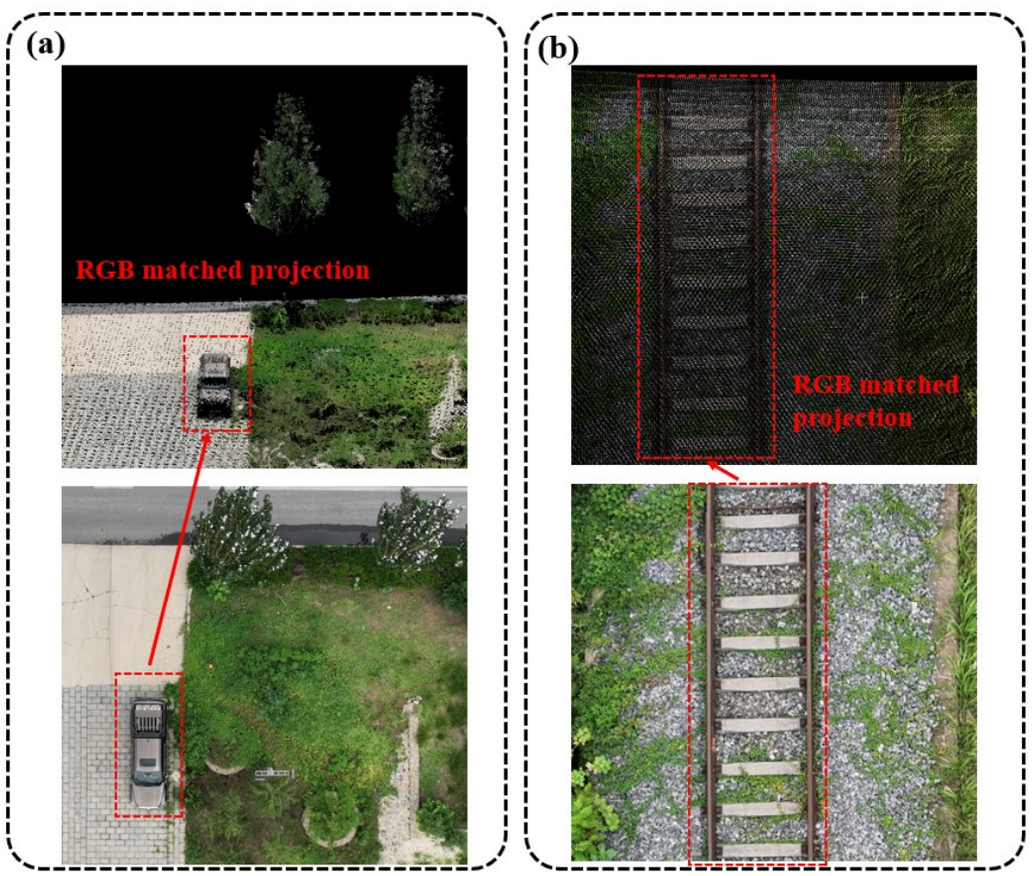

图 13 不同场景下的点云图像匹配结果。(a)正常路面环境下的匹配结果;(b)复杂山区铁路环境下的匹配结果

Fig.13 Point cloud and image matching results in different scenes. (a) Matching result in a normal road environment; (b) matching result in a complex mountainous railway environment

3.2.3 深度学习式自标定

随着人工智能方法的普及,以深度学习为核心的端到端标定方法逐渐发展起来。此类方法通常以初始外参为起点,将点云投影到图像平面,构造跨模态联合输入,并直接回归外参偏差,实现自动标定或在线校正。其优势在于能够通过数据驱动方式学习复杂的跨模态映射关系,在视场重叠不足、纹理变化大或几何

结构不规则的场景中具有潜在适应性。

XU Zejing等人[62]提出KFCalibNet学习式自标定网络：在给定初始外参的条件下，先将3D点云投影到图像平面生成“失配点云投影图”，再利用对称的KansFormer分支分别对RGB图像和点云投影图进行特征编码，从而提取更细粒度的跨模态表示。其中非线性映射$\phi(x)$采用可学习的一维函数，并通过网格细化机制增强对局部变化的表达能力：

$$\phi(x)=\omega_b b(x)+\omega_s spline(x) \tag{18}$$

其中x为当前层输入的特征值，ω_b为基础函数部分的可学习权重，$b(x)$为基础函数即Sigmoid线性单元（Sigmoid Linear Unit，SiLU）激活函数，ω_s样条函数部分的可学习权重，$spline(x)$为样条基函数项。在此基础上，进一步引入多头注意力机制$Attention(Q,K,E)$对跨模态特征间的相关性进行建模，其表达式为：

$$Attention(Q,K,E)=\mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) \tag{19}$$

其中Q为查询矩阵，K为键矩阵，d_k为K的特征维度。最后得到相关特征图X_{corr}。随后，在特征聚合阶段引入FastKAN模块，其核心激活函数采用高斯径向基函数形式，表示为：

$$\phi(x)=\sum_i \omega_i exp\left(-\frac{(x-c_i)^2}{2h^2}\right) \tag{20}$$

其中，c_i表示第i个高斯核的中心位置，h表示高斯核的宽度参数。通过以上公式降低计算复杂度，最终回归外参偏差T_{pred}并按$\hat{T}_{LC}=T_{pred}^{-1}T_{init}$输出标定结果，同时采用组合损失$L=\lambda_t L_t+\lambda_R L_R+\lambda_P L_P$（含点云一致性项）以提升收敛与几何一致性，从而在视场角（Field of View，FOV）重叠不足等工况下实现精确的激光雷达-相机外参自标定。

总体来看，融合基础层已经呈现出较清晰的演进趋势：由基于标定物的高精度几何标定，发展到基于场景结构先验的无标定物自标定方法，再进一步发展到学习式端到端标定与在线校正。这一演进为后续任务层融合中的投影关联、特征对齐与鲁棒性设计提供了方法基础。

3.3 任务感知层：从几何投影到深层交互的检测与分割融合方法

在标定与对齐关系建立之后，研究者开始关注图像语义信息与点云几何信息如何在检测或分割网络中有效结合。从已有研究看，早期方法主要依赖几何投影完成图像语义区域与点云空间信息的对应；随着复杂场景下标定误差、遮挡和视角差异的影响逐渐显现，后续研究开始在网络结构中引入错位补偿和特征对齐机制。近年方法则进一步关注图像语义与点云几何在深层网络中的交互，以提高检测与分割任务在复杂环境下的稳定性。

3.3.1 几何投影驱动的模块化融合

早期任务级融合方法通常以几何投影为纽带，采用模块化串行流程（pipeline）实现视觉语义约束和激光雷达几何检测与测距的协同。这类方法的共同特点是功能分工明确，视觉分支负责生成语义区域或候选目标，激光雷达分支负责空间几何判定与距离估计，因此工程可解释性较强，实现也相对直接。

WANG Zhangyu等人[63]提出了一种相机—激光雷达融合的铁路场景目标检测框架。具体流程见图14。该方法采用两阶段流程，首先设计多尺度预测（Multi-scale Prediction，MSP）语义分割网络对图像进行像素级分割，提取轨道区域与前方列车区域（候选区域），并利用已标定的内外参将激光雷达点投影到图像平面，筛选出轨道候选区域点云与列车候选区域点云；随后在检测阶段，基于轨道候选区域点云通过栅格化和RANSAC拟合轨道高度曲线：

$$z=ay^2+by+c \tag{21}$$

其中，y代表激光雷达点的纵向坐标信息，z代表轨道平面在纵向位置y处的高度值。并计算高度残差$r_i=z_i-\left(ay_i^2+by_i+c\right)$来检测高于轨面的小障碍物，其中，$r_i$表示第$i$个点相对于拟合轨道高度曲线的高度偏差。对于列车候选区域点云，则通过自适应聚类估计前方列车距离，从而实现小障碍物与前方列车的融合检测。

这类模块化投影融合方法在早期研究中具有重要价值，但局限同样明显：方法性能高度依赖标定质量

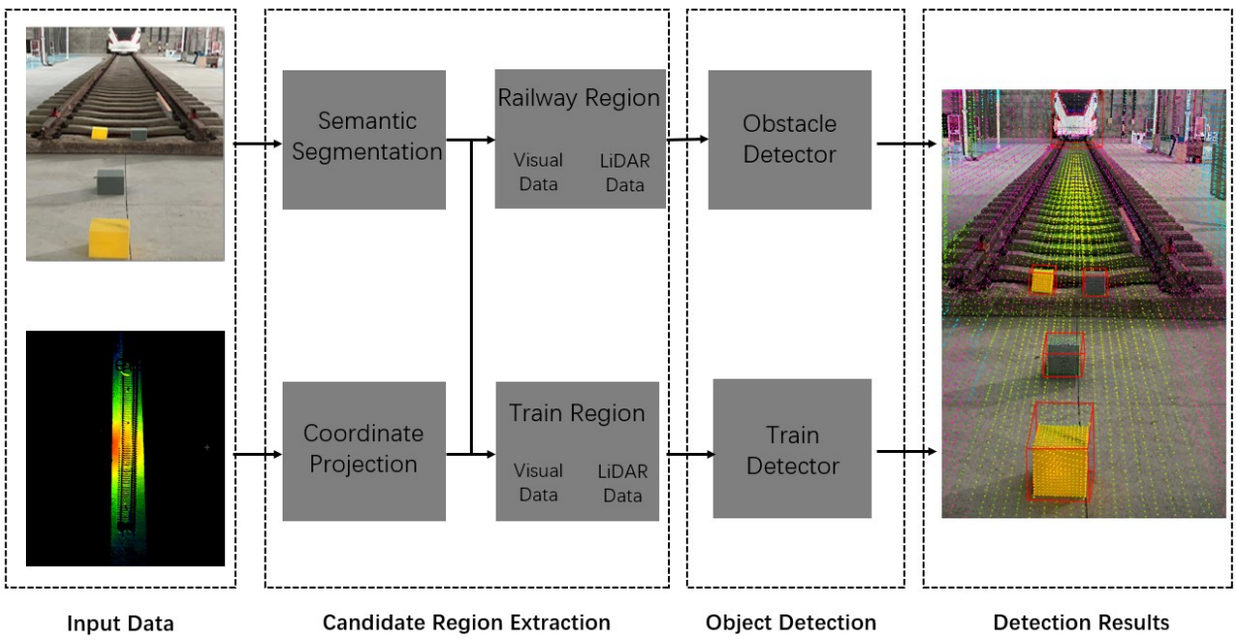

图 14 激光雷达与相机信息融合框架

Fig.14 LiDAR and camera information fusion framework

和投影精度，一旦外参存在偏差或运行中出现振动与时间漂移，候选区域筛选质量会明显下降。而且串行流程容易产生误差传播，视觉分割误差会直接影响后续点云几何处理。因此后续研究开始在任务网络内部引入可学习对齐补偿与更深层次的跨模态特征交互。

3.3.2 面向错位补偿的融合方法

随着融合系统在动态环境中的部署需求增加，研究者逐渐意识到，仅依赖静态外参建立的几何对应关系并不足以消除实际运行中的错位问题，如弱同步、微小安装偏移和振动扰动等。因此，任务级融合方法开始引入可学习的对齐补偿机制，使模型具备一定的错位容错能力。

ZHAO Lin 等人[64]提出了面向分割的激光雷达与图像融合（LiDAR-Image Fusion for Segmentation，LIF-Seg）粗到细的激光雷达-相机融合语义分割框架。其在粗融合阶段先将激光雷达点投影至多相机图像平面，再提取各投影点邻域 $w \times w$（如 3×3）局部图像上下文，并与点云强度信息共同构成早期融合特征。该投影过程可表示为：

$$\left[p_{ix}, p_{iy}, 1 \right]^T = K_i T_i L_{xyz}^h \tag{22}$$

其中，转置向量代表第 i 个相机中该激光雷达点投影后的齐次像素坐标，L_{xyz}^h 表示 LiDAR 点的齐次坐标。此外为解决弱时空同步导致的跨模态错位，在中间阶段用图像语义网络提取 F_{image}，将粗特征投影成伪图 F_{points} 并预测像素级偏移来更新投影位置实现偏移校正，再将对齐后的图像语义特征回投影到点云与粗特征融合，偏移矫正过程见图 15。

最后在细化阶段将融合特征输入激光雷达分割子网得到最终点级语义输出，并以联合损失函数 $L = L_{\text{sem}} + \alpha L_{\text{aux}}$ 对网络进行优化，其中 L_{sem} 负责监督点云语义分割，L_{aux} 负责监督跨模态投影偏移学习，二者通过权重 α 加权求和以联合优化分割性能与对齐精度。

3.3.3 深层双向交互特征提取融合方法

在更高层级的任务融合中，研究重点进一步从单向信息注入转向跨模态表示的双向协同构建，并显式考虑复杂工况和退化场景下的适用性。LIU Wentao 等人[65]提出 RailFusion 交互融合网络用于 3D 铁路目标检测，以 ResNet50 和 VoxelNet 分别提取图像与点云特征后，引入跨域特征提取（Cross-domain Feature Extraction，CDFE）与多模态融合（Multi-modal Fusion，MMF）两个核心模块。其中，CDFE 利用激光雷达生成的伪深度图与语义占据图，对图像特征的前视到俯视视角变换过程进行监督，以增强图像分支对空间深度信息和前景语义信息的感知能力。同时，再将预测得到的图像 BEV 语义信息投影回激光雷达体素空间，

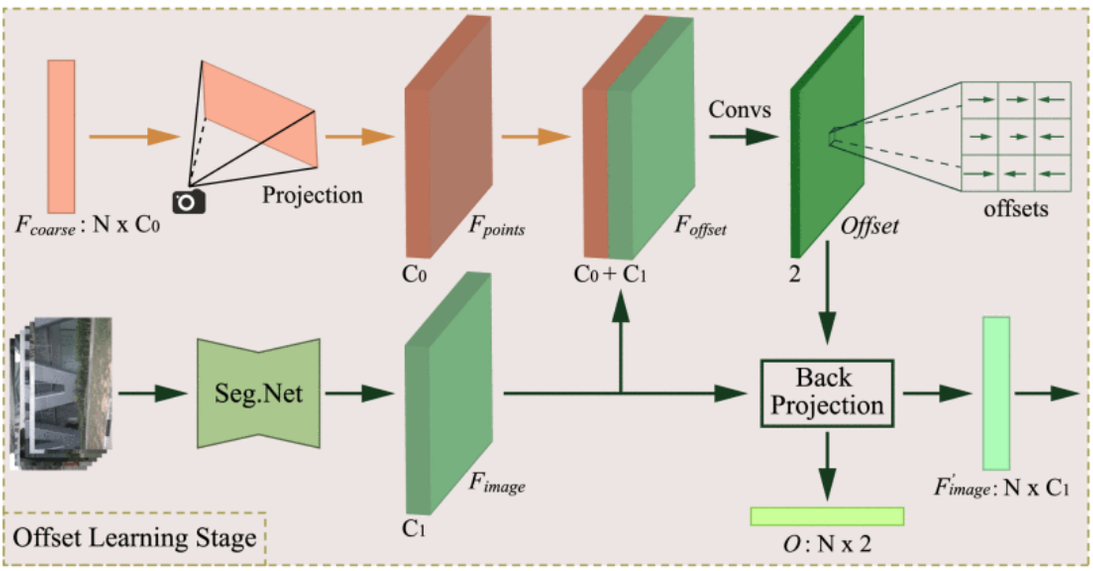

图15　偏移学习阶段示意图[64]

Fig.15　Schematic diagram of the offset learning stage[64]

对点云特征进行语义增强。其LiDAR语义增强过程可表示为：

$$F_{\text{lidar}} = F_{\text{lidar}} \times \left(O_{cc} \cdot \alpha_{\text{fore}} + (1 - O_{cc}) \cdot \alpha_{\text{back}} \right) \tag{23}$$

其中，F_{lidar}表示激光雷达语义特征，O_{cc}为占据布尔变量，用于判断该体素是否属于前景目标，α_{fore}与α_{back}分别为前景和背景体素权重。该式的作用在于通过区分前景与背景体素的重要性，对点云特征进行加权增强，从而提升稀疏点云对目标语义信息的表达能力。MMF则用可变形注意力实现像素级跨模态对齐以缓解外参不准导致的错位，其融合形式为：

$$DAttn\left(Q, P, F_{\text{image}}\right) = \sum_{m=1}^{M} W_m \left[\sum_{k=1}^{K} A_{mqk}(Q) \cdot \mathrm{W}'_m \mathrm{F}_{\text{image}}\left(P + \Delta P_{mqk}\right) \right] \tag{24}$$

其中，Q为查询特征，P为参考点的位置，M为注意力头数，A_{mqk}由多层感知机生成的注意力权重分数，ΔP_{mqk}动态生成的采样偏移量。随后进一步采用卷积块注意力模块(Convolutional Block Attention Module，CBAM)进行通道和空间维度的动态加权，使融合权重能够随场景和目标变化自适应调整，从而提升融合特征的有效性与鲁棒性。

这说明RailFusion并非简单拼接图像与点云特征，而是通过视角转换监督、前景语义加权和动态跨模态对齐，增强复杂铁路场景下图像与激光雷达特征的互补性。尽管该方法是在铁路场景中验证的，但其核心思想是在统一表示空间中构建双向跨模态交互，并显式面向错位与退化设计鲁棒机制，是当前多传感器融合检测方法的重要发展方向。这类方法表明，激光雷达—相机融合的目标已经不再只是提升平均检测精度，而是更多地指向复杂工况下的检测可靠性。

3.4　定位建图中的多传感器紧耦合

对三维目标检测而言，稳定的定位建图结果能够为多帧点云配准和跨帧检测结果融合提供空间基准。激光雷达与相机信息融合的应用外延并不限于检测与分割等任务感知层，也体现在即时定位与地图构建(Simultaneous Localization and Mapping，SLAM)中。此类研究关注的核心问题在于：如何利用不同模态在几何、纹理与运动信息上的互补性，构建稳定、低漂移的状态估计，为后续感知与任务执行提供统一的时空基准。

XIA Yu等人[66]提出了一种LiDAR - Camera - IMU - GNSS紧耦合即时定位与地图构建(Simultaneous Localization and Mapping，SLAM)融合框架：以前端点线特征激光-视觉-惯性里程计为基础，在视觉-惯性子系统中引入线段检测器(Line Segment Detector，LSD)线段并做点线融合，同时将激光点云投影到相机坐标系建立深度关联，并利用滑动窗口光束法平差(Bundle Adjustment，BA)优化相机位姿。在激光-惯性子系统中，则通过惯性测量单元(Inertial Measurement Unit，IMU)预积分消除点云运动畸变并提取边缘，再

结合视觉-激光双回环检测构造回环因子，并在后端用因子图和增量平滑与建图（incremental Smoothing and Mapping，iSAM2）统一优化里程计因子、IMU预积分因子、回环因子与全球导航卫星系统（Global Navigation Satellite System，GNSS）全局约束因子（如$r_G = p_L^k - p_G^k$），从而获得全局一致的位姿估计与高精度点云地图。表3对上述典型激光雷达与相机融合方法进行了归纳。

表3 典型激光雷达与相机融合方法总结

Table 3 Summary of typical LiDAR-camera joint perception methods

Category	Main result	Reference
Target-based calibration	High calibration accuracy, but relies on manual targets.	[58]
Plane-constrained calibration	Translation RMSE about 0.01 m and rotation RMSE about 0.05°.	[59]
Model-based calibration	Accurate with few observations, but the workflow is complex.	[60]
Targetless calibration	Uses road-line features instead of artificial targets, but depends on scene structure.	[61]
Learning-based calibration	Regresses extrinsic errors automatically, but depends on training data.	[62]
Projection-based fusion	Uses image proposals and LiDAR geometry for detection and ranging.	[63]
Misalignment-aware fusion	Improves mIoU from 76.1 to 78.2, with higher model complexity.	[64]
Railway feature fusion	RailFusion reaches 57.2 mAP and 68.8 NDS.	[65]
Tightly coupled mapping	Reduces trajectory RMSE compared with LIO-SAM.	[66]

多模态融合将激光雷达的空间测量能力与相机的图像语义信息结合起来，在一定程度上弥补单一传感器的不足。对于远距离目标、遮挡目标和点云较稀疏的目标，图像中的纹理和类别能够为点云检测提供补充。但融合效果不只取决于传感器数量，图像和点云在成像方式、空间分布和特征表达上存在差异，浅层融合实现较直接，但对外参标定和时间同步要求较高。深层融合能够加强特征交互，但若权重分配不合理，网络可能偏向某一模态，导致另一模态的信息没有得到充分利用。

实际应用中，对齐误差和系统复杂度是限制融合方法稳定性的主要因素。车辆振动、安装误差与雨雾、强光等环境因素，都会影响图像与点云之间的对应关系。一旦语义信息被错误地分配到点云位置上，融合不仅不能提升检测效果，还可能引入噪声，造成误检或定位偏差。同时，多传感器系统涉及采集、同步、标定、坐标变换、特征关联和后处理等多个环节，任何环节出现偏差都会影响最终结果。因此，多模态融合不应只看作信息叠加，更需要关注信息可靠性、融合位置、错误模态干扰和工程部署中的稳定性。

4 结论

激光雷达三维目标检测已从传统几何处理发展到深度学习检测框架，并进一步向多模态融合延伸。不同方法在精度、效率和工程实现之间各有侧重，但在复杂场景下仍面临共同问题：有效观测不足、目标结构不完整以及检测结果稳定性不够。远距离稀疏目标、小目标和遮挡目标仍是现有方法的弱项，同时，体素离散化、BEV压缩和投影变换等处理方式虽然提高了计算效率，也可能带来局部结构信息损失。多模态融合能够补充图像语义信息，但其效果依赖传感器标定、时间同步和跨模态对齐，一旦出现外参漂移、图像退化或模态异常，融合信息可能反而影响检测可靠性[67]。因此，三维目标检测的评价不应只依赖公开数据集上的平均精度，还应结合远距离目标、检测环境和算力约束等因素综合分析。

未来研究需要从单纯追求检测指标提升，进一步转向复杂场景适应性和工程可部署性。一方面，可围绕远距离稀疏目标和小目标检测，引入多帧时序信息、点云补全、尺度自适应特征增强等方法，提高有限点云条件下的目标可分性；另一方面，应继续探索低标注成本学习方式，利用半监督、弱监督和跨域自训练降低三维框标注依赖[68]。在多模态融合方面，后续研究不应只增加融合结构，还需重点考虑标定误差、时间不

同步和模态质量下降对检测结果的影响。面向自动驾驶、铁路异物入侵监测等应用,应进一步降低模型复杂度,提高实时检测能力,并完善系统维护机制,形成兼顾精度、效率和可靠性的三维目标检测方案。

参考文献

[1] ZHAO Zengxu, HU Lianqing, REN Bin, et al. PointPillars-S 3D object detection algorithm based on LiDAR[J]. Acta Photonica Sinica, 2025, 54(6): 0614002.
赵增旭, 胡连庆, 任彬, 等. 基于激光雷达的PointPillars-S三维目标检测算法[J]. 光子学报, 2025, 54(6): 0614002.

[2] TANG Jie, CHEN Wenwu, ZHOU Xinran, et al. Deep Learning-based Point Cloud Registration via Neighborhood Geometric Centroids (Invited)[J]. Acta Photonica Sinica, 2025, 54(9): 0954211.
汤洁, 陈文武, 周昕然, 等. 基于邻域几何质心的深度学习点云配准(特邀)[J]. 光子学报, 2025, 54(9): 0954211.

[3] VALVERDE M, MOUTINHO A, ZACCHI J V. A survey of deep learning-based 3D object detection methods for autonomous driving across different sensor modalities[J]. Sensors, 2025, 25(17): 5264.

[4] GEIGER A, LENZ P, URTASUN R. Are we ready for autonomous driving? The KITTI vision benchmark suite[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2012: 3354-3361.

[5] 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. Piscataway: IEEE, 2017: 1907-1915.

[6] ZERMAS D, IZZAT I, PAPANIKOLOPOULOS N. Fast segmentation of 3D point clouds: A paradigm on LiDAR data for autonomous vehicle applications[C]//2017 IEEE International Conference on Robotics and Automation (ICRA). Piscataway: IEEE, 2017: 5067-5073.

[7] HIMMELSBACH M, VON HUNDELSHAUSEN F, WUENSCHE H J. Fast segmentation of 3D point clouds for ground vehicles[C]//2010 IEEE Intelligent Vehicles Symposium. Piscataway: IEEE, 2010: 560-565.

[8] BOGOSLAVSKYI I, STACHNISS C. Fast range image-based segmentation of sparse 3D laser scans for online operation [C]//2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Piscataway: IEEE, 2016: 163-169.

[9] CHEN Tongtong, DAI Bin, LIU Daxue, et al. Sparse Gaussian process regression based ground segmentation for autonomous land vehicles[C]//The 27th Chinese Control and Decision Conference (2015 CCDC). Piscataway: IEEE, 2015: 3993-3998.

[10] FU Huijin, MA Zhen, YANG Qi, et al. Lidar-based foreign object detection for railway intrusion under adverse weather conditions[C]//International Conference on Optical Communication, Signal Processing, and Optical Engineering (OCSPOE 2025). SPIE, 2025, 13797: 153-159.

[11] ZHANG Wuming, QI Jianbo, WAN Peng, et al. An easy-to-use airborne LiDAR data filtering method based on cloth simulation[J]. Remote Sensing, 2016, 8(6): 501.

[12] ZEYBEK M. Spatio-Temporal Detection and Filtering of Dynamic Objects in Mobile LiDAR Point Clouds[J]. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 2026, 48: 371-378.

[13] RUSU R B, COUSINS S. 3D is here: Point Cloud Library (PCL)[C]//2011 IEEE International Conference on Robotics and Automation. Piscataway: IEEE, 2011: 1-4.

[14] YANG Yuxing, ZHANG Bowen, YANG Boyu, et al. End-to-end railway obstacle detection enhanced by point cloud segmentation[J]. Engineering Applications of Artificial Intelligence, 2026, 168: 114008.

[15] ESTER M, KRIEGEL H P, SANDER J, et al. A density-based algorithm for discovering clusters in large spatial databases with noise[C]//Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining. Menlo Park: AAAI Press, 1996: 226-231.

[16] WANG Heng, WANG Bin, LIU Bingbing, et al. Pedestrian recognition and tracking using 3D LiDAR for autonomous vehicle[J]. Robotics and Autonomous Systems, 2017, 88: 71-78.

[17] HELD D, GUILLORY D, REBSAMEN B, et al. A probabilistic framework for real-time 3D segmentation using spatial, temporal, and semantic cues[C]//Robotics: Science and Systems. 2016.

[18] IVARSEN M F, ST-MAURICE J P, HUSSEY G C, et al. Point-cloud clustering and tracking algorithm for radar interferometry[J]. Physical Review E, 2024, 110(4): 045207.

[19] GU Bo, LIU Jianxu, XIONG Huiyuan, et al. ECPC-ICP: A 6D vehicle pose estimation method by fusing the roadside lidar point cloud and road feature[J]. Sensors, 2021, 21(10): 3489.

[20] WANG C C, WANG Mudan, SUN Jun, et al. A safety warning algorithm based on axis aligned bounding box method to prevent onsite accidents of mobile construction machineries[J]. Sensors, 2021, 21(21): 7075.

[21] ZHANG Xiao, XU Wenda, DONG Chiyu, et al. Efficient L-shape fitting for vehicle detection using laser scanners[C]// 2017 IEEE Intelligent Vehicles Symposium (IV). Piscataway: IEEE, 2017: 54-59.

[22] RUSU R B, BLODOW N, BEETZ M. Fast point feature histograms (FPFH) for 3D registration[C]//2009 IEEE International Conference on Robotics and Automation. Piscataway: IEEE, 2009: 3212-3217.

[23] ZongliangNAN, ZHU Guoan, ZHANG Xu, et al. A novel high-precision railway obstacle detection algorithm based on 3D LiDAR[J]. Sensors, 2024, 24(10): 3148.
[24] TEICHMAN A, LEVINSON J, THRUN S. Towards 3D object recognition via classification of arbitrary object tracks [C]//2011 IEEE International Conference on Robotics and Automation. Piscataway: IEEE, 2011: 4034-4041.
[25] GOSWAMI P, VAISHNAV R, ANAND T, et al. A comprehensive review on LiDAR based 3D deep learning object detection algorithms [C]//2025 International Conference on Computer, Electrical & Communication Engineering (ICCECE). Piscataway: IEEE, 2025: 1-6.
[26] ZHANG Xiang, WANG Hai, DONG Haoran. A survey of deep learning-driven 3D object detection: Sensor modalities, technical architectures, and applications[J]. Sensors, 2025, 25(12): 3668.
[27] WU Shuwen, LI Yanxi, ZHANG Shaochen, et al. Review of deep learning-based 3D point cloud object detection[J]. Journal of Telemetry, Tracking and Command, 2024, 45(5): 1-18.
武淑文, 李燕烯, 张少琛, 等. 基于深度学习的3D点云目标检测研究综述[J]. 遥测遥控, 2024, 45(5): 1-18.
[28] 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. Piscataway: IEEE, 2017: 652-660.
[29] 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.
[30] SHI Shaoshuai, WANG Xiaogang, LI Hongsheng. PointRCNN: 3D object proposal generation and detection from point cloud[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2019: 770-779.
[31] YANG Zetong, SUN Yanan, LIU Shu, et al. 3DSSD: Point-based 3D single stage object detector[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2020: 11040-11048.
[32] YAN Yan, MAO Yuxing, LI Bo. SECOND: Sparsely embedded convolutional detection[J]. Sensors, 2018, 18(10): 3337.
[33] CUI Yubo, LI Zhiheng, FANG Zheng, et al. Dynamic clustering transformer for LiDAR-based 3D object detection[J]. Pattern Recognition, 2026, 172: 112444.
[34] LIAN Lirong, QIN Yong, CAO Zhiwei, et al. RVSA-3D: Voxel-based fully sparse attention 3D object detection for rail transit obstacle perception[J]. Pattern Recognition, 2026, 171: 112324.
[35] CHEN Yukang, LI Yanwei, ZHANG Xiangyu, et al. Focal sparse convolutional networks for 3D object detection[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2022: 5428-5437.
[36] 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. Piscataway: IEEE, 2019: 12697-12705.
[37] AO Lei, WAN Wenkang, OuyangNAN, et al. SVP: Stratified vertical priors for LiDAR-based 3D object detection[J]. Neurocomputing, 2026, 659: 131737.
[38] LI Junru, WANG Zhiling, GONG Diancheng, et al. SCNet3D: Rethinking the feature extraction process of pillar-based 3D object detection[J]. IEEE Transactions on Intelligent Transportation Systems, 2025, 26(1): 770-784.
[39] NOH J, LEE J, PARK H, et al. 3DPillars: Pillar-based two-stage 3D object detection[J]. Expert Systems With Applications, 2025, 289: 128349.
[40] 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. Piscataway: IEEE, 2019: 12677-12686.
[41] 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. Piscataway: IEEE, 2020: 10529-10538.
[42] REN Junfen, WEN Changji, ZHANG Long, et al. High performance point-voxel feature set abstraction with mamba for 3D object detection[J]. Expert Systems With Applications, 2025, 286: 128127.
[43] SHI Shaoshuai, WANG Zhe, SHI Jianping, et al. From points to parts: 3D object detection from point cloud with part-aware and part-aggregation network[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 43(8): 2647-2664.
[44] 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. Piscataway: IEEE, 2018: 7652-7660.
[45] 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. Piscataway: IEEE, 2021: 11784-11793.
[46] CHEN Shitao, ZHANG Haolin, ZHENG Nanning. Leveraging anchor-based LiDAR 3D object detection via point assist-

ed sample selection[J]. IEEE Transactions on Intelligent Transportation Systems, 2025, 26(6): 7939-7952.
[47] XIA Qiming, LIN Hongwei, YE Wei, et al. Label-efficient outdoor 3D object detection via single click annotation from LiDAR point cloud[J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2026, 234: 151-166.
[48] MI Zeng, LIAN Zhe. Review of YOLO methods for general object detection[J]. Computer Engineering and Applications, 2024, 60(21): 38-54.
米增, 连哲. 面向通用目标检测的YOLO方法研究综述[J]. 计算机工程与应用, 2024, 60(21): 38-54.
[49] HUANG Zhe, WANG Yongcai, LI Deying. Review of 3D object detection methods[J]. Chinese Journal of Intelligent Science and Technology, 2023, 5(1): 7-31.
黄哲, 王永才, 李德英. 3D目标检测方法研究综述[J]. 智能科学与技术学报, 2023, 5(1): 7-31.
[50] YANG Boquan, LI Jixiong, ZENG Ting. A review of environmental perception technology based on multi-sensor information fusion in autonomous driving[J]. World Electric Vehicle Journal, 2025, 16(1): 20.
[51] ZHAO Xuanyi, DOU Xiaohan, ZHENG Jihong, et al. A Lightweight Multi-Stage Visual Detection Approach for Complex Traffic Scenes[J]. Sensors, 2025, 25(16): 5014.
[52] YU Shuainan, WANG Xichao, YanCI, et al. MSPE-Fusion: A multimodal 3D object detection method with multi-sensor perception enhanced fusion[J]. Neurocomputing, 2025, 645: 130486.
[53] ZHANG Yongsheng, TU Chen, GAO Kun, et al. Multisensor information fusion: Future of environmental perception in intelligent vehicles[J]. Journal of Intelligent and Connected Vehicles, 2024, 7(3): 163-176.
[54] FAN Zheng, ZHANG Lele, WANG Xueyi, et al. LiDAR, IMU, and camera fusion for simultaneous localization and mapping: A systematic review[J]. Artificial Intelligence Review, 2025, 58(6): 174.
[55] SHI Pengtao, WEI Kangle. Extrinsic calibration method of LiDAR and camera based on semantic segmentation for autonomous driving environment[J]. Laser & Optoelectronics Progress, 2024, 61(24): 2428011.
史鹏涛, 危康乐. 自动驾驶环境下基于语义分割的激光雷达与相机外参标定方法[J]. 激光与光电子学进展, 2024, 61(24): 2428011.
[56] LIU Yang, JI Jie, PAN Deng, et al. Agricultural robot localization method based on LiDAR and IMU fusion[J]. Smart Agriculture, 2024, 6(3): 94-106.
刘洋, 冀杰, 潘登, 赵立军, 李明生. 基于激光雷达与IMU融合的农业机器人定位方法[J]. 智慧农业(中英文), 2024, 6(3): 94-106.
[57] FU Ting, XIE Shuke, HUI Weichao, et al. LiDAR-camera fusion: Dual-scale correction for vehicle multi-object detection and trajectory extraction[J]. Journal of Intelligent Transportation Systems, 2024: 1-14.
[58] ZongliangNAN, LIU Wenlong, ZHU Guoan, et al. LiDAR-camera joint obstacle detection algorithm for railway track area[J]. Expert Systems with Applications, 2025, 275: 127089.
[59] SINGANDHUPE A, LA H M. Single frame lidar and stereo camera calibration using registration of 3D planes[C]//2021 Fifth IEEE International Conference on Robotic Computing (IRC). Piscataway: IEEE, 2021: 115-118.
[60] JEONG S, KIM S. O3 LiDAR-camera calibration: One-shot, one-target and overcoming LiDAR limitations[J]. IEEE Sensors Journal, 2024, 24(11): 18659-18671.
[61] HE Zhanyuan, CHENG Guangzhi, ZHANG Yongbo, et al. Design of a LiDAR-camera self-calibration algorithm based on railway geometric features[C]//2025 IEEE 8th International Conference on Signal Processing and Machine Learning (SPML). Piscataway: IEEE, 2025: 454-459.
[62] XU Zejing, LIU Yiqing, GAO Ruipeng, et al. KFCalibNet: A KansFormer-based self-calibration network for camera and LiDAR[C]//2025 IEEE International Conference on Robotics and Automation (ICRA). Piscataway: IEEE, 2025: 627-633.
[63] WANG Zhangyu, YU Guizhen, WU Xinkai, et al. A camera and LiDAR data fusion method for railway object detection [J]. IEEE Sensors Journal, 2021, 21(12): 13442-13454.
[64] ZHAO Lin, ZHOU Hui, ZHU Xinge, et al. Lif-Seg: LiDAR and camera image fusion for 3D LiDAR semantic segmentation[J]. IEEE Transactions on Multimedia, 2023, 26: 1158-1168.
[65] LIU Wentao, WANG Yunpeng, YU Guizhen, et al. RailFusion: A LiDAR-camera data interaction network for 3-D railway object detection[J]. IEEE Transactions on Intelligent Transportation Systems, 2025, 26(8): 12761-12773.
[66] XIA Yu, WU Hongwei, ZHU Liucun, et al. A multi-sensor fusion framework with tight coupling for precise positioning and optimization[J]. Signal Processing, 2024, 217: 109343.
[67] FAWOLE O A, RAWAT D B. Recent advances in 3D object detection for self-driving vehicles: A survey[J]. AI, 2024, 5(3): 1255-1285.
[68] ZHU Minling, GONG Yadong, TIAN Chunwei, et al. A systematic survey of transformer-based 3D object detection for autonomous driving: Methods, challenges and trends[J]. Drones, 2024, 8(8): 41.

Research Progress on LiDAR-Based 3D Object Detection Algorithms

ZHAO Hongwei[1,2], XIA Wentao[1,3], LIU Wenlong[1,2], XIE Yi[1], HE Chaojian[1,4,5], SHANGGUAN Mingjia, YU Haijuan[1,4,5], YANG Yingying[1,4,5], YANG Song[1,4,5*], LIN Xuechun[1,4,5]

(1 *Laboratory of All-Solid-State Light Sources, Institute of Semiconductors, Chinese Academy of Sciences, Beijing* 100083, *China*)

(2 *School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing* 101407, *China*)

(3 *School of Industry-Education Integration, University of Chinese Academy of Sciences, Beijing* 101407, *China*)

(4 *College of Materials Science and Opto-Electronic Technology, University of Chinese Academy of Sciences, Beijing* 101407, *China*)(5 *Center of Materials Science and Optoelectronics Engineering, University of Chinese Academy of Sciences, Beijing* 100049, *China*)(6 *Beijing Engineering Technology Research Center of All-Solid-State Lasers Advanced Manufacturing, Beijing* 100083, *China*)

(5 *State Key Laboratory of Marine Biogeochemistry, College of Ocean and Earth Sciences, Xiamen University, Xiamen, Fujian* 361102, *China*)

Abstract: LiDAR-based three-dimensional object detection is a key perception task for autonomous driving, railway intrusion monitoring, and other safety-critical applications. This review aims to clarify the main technical routes, representative methods, applicable scenarios, and remaining limitations of LiDAR-based 3D object detection. Instead of treating existing algorithms as a simple chronological sequence, this paper reorganizes them according to their processing mechanisms and practical constraints. Particular attention is paid to how different methods handle point cloud sparsity, incomplete object observations, semantic insufficiency, computational cost, and multimodal alignment errors. The purpose is to provide a structured reference for method selection, performance interpretation, and future research on reliable 3D object detection.The reviewed studies are organized into three parts: traditional geometric processing, deep learning-based detection, and LiDAR-camera fusion. For traditional methods, the review discusses ground segmentation, range-image segmentation, Euclidean clustering, density-based clustering, geometric fitting, and handcrafted descriptors, with emphasis on their roles in candidate generation, noise suppression, and error control. For deep learning methods, the review analyzes two closely related aspects: input representation and detection structure. The input representations include point-based methods, voxel-based sparse convolution, pillar and bird's-eye-view representations, range-view projection, and hybrid point-voxel schemes. The detection structures include two-stage, one-stage, and center-based paradigms. For multimodal fusion, this review covers calibration-based alignment, targetless and learning-based self-calibration, projection-guided fusion, misalignment compensation, cross-modal feature interaction, and tightly coupled mapping. These methods are compared in terms of geometric detail preservation, feature representation ability, detection accuracy, inference efficiency, robustness, and deployment complexity. The review shows that traditional geometric methods remain useful when interpretability, simple implementation, and low computational cost are required. Ground fitting, clustering, L-shape fitting, and local geometric descriptors can reduce the search space and provide explicit shape constraints, but their performance is strongly affected by threshold settings, ground assumptions, target priors, and point density variation. When the scene contains uneven terrain, sparse distant objects, occlusion, or non-convex structures, errors in early segmentation and clustering are likely to propagate to bounding-box fitting and classification. Deep learning methods reduce this dependence on handcrafted rules by integrating point organization, feature extraction, candidate generation, and box regression into trainable frameworks. Point-based methods preserve fine-grained geometry and avoid discretization loss, but neighborhood search and local aggregation increase computational burden. Voxel-based sparse convolution improves spatial regularity and reduces invalid computation in empty regions, but voxel resolution still affects small-object representation and memory cost. Pillar and BEV methods convert 3D detection into efficient 2D dense prediction, making them suitable for real-time deployment, although height compression weakens vertical geometric expression. Range-view methods are compact and fast because they follow the native scanning form of LiDAR, but they are sensitive to occlusion, scale variation, and projection distortion. Hybrid methods combine voxel-level context with point-level detail and often achieve high accuracy, but their pipelines are more complex and harder to deploy. From the detection-structure perspective, two-stage methods usually provide stronger candidate refinement and

localization accuracy, one-stage methods offer shorter inference paths and higher efficiency, and center-based methods reduce anchor matching complexity while maintaining a balance between accuracy and speed. LiDAR-camera fusion can improve recognition in sparse, distant, or semantically ambiguous scenes by introducing image texture and semantic cues. However, its effectiveness depends on reliable extrinsic calibration, temporal synchronization, spatial alignment, and appropriate cross-modal feature weighting. Misalignment, sensor degradation, or inaccurate semantic projection may introduce additional noise rather than improve detection. Reliable LiDAR-based 3D object detection cannot be achieved by improving network accuracy alone. Method selection should be guided by target scale, point cloud density, scene regularity, computing resources, and system maintenance requirements. For distant sparse targets, small objects, and occluded objects, future work should strengthen multi-frame information use, scale-adaptive feature learning, point cloud completion, and uncertainty estimation. For data annotation, low-cost supervision such as weak supervision, semi-supervised learning, click-level annotation, and cross-domain self-training deserves further attention. For multimodal systems, future research should focus less on simply adding fusion modules and more on calibration drift, time synchronization error, modality degradation, and failure detection. A practical 3D detection system should finally balance accuracy, real-time performance, robustness, annotation cost, and deployability under real operating conditions.

Key words: LiDAR; point cloud; three-dimensional object detection; deep learning; LiDAR-camera fusion

OCIS Codes: 280.3640; 150.6910; 150.4232; 100.3008

CSTR: 32255.14.gzxb20265508.0814001

Foundation item: National Natural Science Foundation of China (No.62225507, 62475254); CAS Project for Young Scientists in Basic Research (No. YSBR-065).