Website：ycyk.brit.com.cn

基于深度学习的3D点云目标检测研究综述

武淑文，李燕烯，张少琛，杨金福
（北京工业大学 北京 100124）

摘要：近年来，3D目标检测作为自动驾驶、移动机器人、虚拟现实等应用产业的重要基础任务，受到了各领域研究人员的广泛关注。其旨在三维空间中对感兴趣目标进行定位与分类，给出相应的3D包围盒，包括目标的位置、大小和方向，为后续对三维场景的理解与感知、对车辆的规划与决策提供基础信息。激光雷达传感器捕获的点云因其具有准确的三维信息与深度信息，成为3D目标检测最为常用的输入数据。本文对基于深度学习的3D激光雷达点云目标检测进行综述，总结了点云的数据特点与处理方法，介绍了相应的几类检测方法以及点云和图像融合的多模态检测方法，对不同方法的性能进行对比分析，最后讨论3D点云目标检测未来面临的挑战和发展趋势。

关键词：3D目标检测；点云；激光雷达；自动驾驶；深度学习
中图分类号：TP391.4; TN958.98 **文献标志码：**A **文章编号：**2095-1000(2024)05-0001-18
DOI：10.12347/j.ycyk.20240604001
引用格式：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.

3D Object Detection Methods Based on Point Cloud with Deep Learning: A Survey

WU Shuwen，LI Yanxi，ZHANG Shaochen，YANG Jinfu
（Beijing University of Technology, Beijing 100124, China）

Abstract: In recent years, as a crucial and fundamental task in applications such as autonomous driving, mobile robotics, and virtual reality, 3D object detection has received extensive attention from researchers in various fields. It aims to localize and classify objects of interest in 3D space and give the corresponding 3D bounding boxes, including the position, size, and orientation of objects, which provides the basic information for the subsequent understanding and perception of the 3D scene as well as planning and decision-making. Point clouds captured by LiDAR have become the most commonly used input data for 3D object detection due to their accurate 3D information and depth information. In this paper, the 3D object detection methods based on LiDAR point cloud with deep learning are reviewed, the characteristics and processing methods of point cloud are summarized, and several corresponding types of detection methods and multimodal fusion methods of point cloud and image are introduced. At the same time, this paper compares the performance of different methods and discusses the challenges and development trends of 3D object detection based on point cloud in the future.

Keywords: 3D object detection; Point cloud; LiDAR; Autonomous driving; Deep learning
Citation: WU Shuwen, LI Yanxi, ZHANG Shaochen, et al. 3D Object Detection Methods Based on Point Cloud with Deep Learning:A Survey[J]. Journal of Telemetry, Tracking and Command, 2024, 45(5): 1–18.

0 引言

在计算机视觉领域中，作为许多其他视觉任务与应用的基础与核心问题，目标检测得到了众多研究人员的广泛关注[1,2]。其中，基于图像的2D目标检测任务是在图像中找出人类感兴趣的物体并进行分类与定位，具体标注的目标根据不同任务与不同应用场景的要求而确定[3]。经过多年来的发展，2D目标检测已经取得了显著进展，在人脸识别、工业检测、文本检测与遥感目标检测等任

基金项目：国家自然科学基金（61973009）
通信作者：杨金福（jfyang@bjut.edu.cn）
收稿日期：2024-06-04；**修回日期：**2024-07-22

务中得到充分应用[4]。近年来，随着自动驾驶、移动机器人、虚拟现实等新兴产业的蓬勃发展，仅在图像上生成2D包围框无法满足在现实世界中对距离的判断，以及后续对路线的规划。因此，许多学者与机构开展了3D目标检测方法的研究。3D目标检测旨在3D空间中识别与定位出感兴趣目标，用紧密的3D立体框包围目标，并给出目标的类别、大小和方向[5,6]。3D目标检测是3D空间场景理解的基础，在各类应用领域兼具重要性与挑战性。尤其是在场景复杂的自动驾驶任务中，3D目标检测对附近关键物体进行识别与定位，为下游任务提供关键信息[7−9]，在车辆环境感知系统中起着至关重要的作用。

3D目标检测方法包括基于图像的方法与基于点云的方法。尽管图像的获取成本较低，但图像缺乏准确的三维结构信息，从图像中估计深度误差较大，因此基于图像的3D目标检测方法的性能受到限制。随着激光扫描和深度获取硬件设备成本的降低以及软件技术的日益发展，研究人员开始研究基于激光雷达点云的3D目标检测方法，并应用于多个领域。激光雷达以其高精度、高频率及高分辨率的优势，成为自动驾驶系统中的主流传感器[10,11]，其生成的3D点云数据保留了3D空间中原始的几何信息，具有强大的三维表征能力。基于激光雷达点云的目标检测通过对道路的提取、障碍物的探测和周边目标的感知，能够为自动驾驶车辆提供准确的定位和导航信息。在移动机器人领域，基于激光雷达点云的3D目标检测通过获取周围环境的精确信息，实现自主避障、导航、建图等功能。在军事领域，得益于激光雷达良好的低空探测性能、隐蔽性与抗干扰性，基于激光雷达的目标检测可用于对低空飞行目标的识别与跟踪、参数测量和姿态调整。在水下，由于激光在海水中衰减小、穿透性好，可以利用激光雷达对水下目标进行搜索、定位、识别与跟踪等，将水下激光雷达与全球定位系统、惯性导航系统联合使用，可以测量水下目标的三维信息，构建水下目标的三维模型。

与2D目标检测已经得到较为充分研究相比，尽管点云数据能提供准确的位置信息以描述3D空间，但3D目标检测正处于快速发展阶段，仍面临诸多挑战。基于激光雷达点云的3D目标检测已逐渐成为各研究机构与汽车公司的研究热点。本文对以激光雷达点云为输入的3D目标检测方法进行综述，介绍了点云的特点与不同处理方法，以及每一类处理方法相对应的代表算法，包括基于原始点云、基于体素、基于点云与体素融合、基于投影视图和基于图的检测方法。此外，简要介绍了基于图像和点云融合的多模态检测方法。最后，在介绍相关数据集与评价指标的基础上，对有关方法的性能进行简要对比与分析，并对未来趋势予以展望。

1 点云数据的获取与特点

1.1 点云的获取

激光雷达获取点云的过程如下：首先由激光发射器发射出激光束，激光束遇到物体后被反射回激光接收器，再通过分析激光遇到目标对象的折返时间，计算出目标对象与发射器的相对距离，这样通过脉冲激光不断地扫描目标对象，经过一定处理后得到精确的三维立体数据[12,13]。当激光雷达传感器旋转一圈时，所有反射的点坐标形成一个点云。换言之，点云是在同一空间参考系下，表达场景中所有目标的空间分布和表面特性的点集合[14]。

点云中的每个点通常由四维向量组成，包括三维坐标(x,y,z)和激光反射强度r。其中，强度值r反映物体反射表面的相关信息，如物体表面的材质、粗糙程度和反射率等。由于点云拥有还原目标三维几何、表面和尺度信息的能力，能够提供物体在三维空间中精确的位置信息，在自动驾驶、三维重建、文物修护、医学与军事等领域发挥着重要作用。

1.2 点云的特点

3D点云数据具有稀疏性、无序性、非结构化等特点，且2D目标检测方法无法直接应用于3D目标检测任务，因此，基于点云的3D目标检测仍面临较大挑战。

1.2.1 稀疏性

2D图像中像素稠密且均匀分布，而激光雷达点云具有很强的稀疏性，且密度分布不均。如图1所示为被广泛使用的KITTI(自动驾驶任务数据集)中某一场景点云，可以看到，整个场景点云并不是完全覆盖，且检测框内的目标点云尤其稀疏。其根本原因是激光雷达传感器的工作原理使得采样点只分布于物体的表面。而目标形状的缺失主

要有以下几种原因：

① 外部遮挡。激光束到达首个物体表面就会发生反射，使得位于该物体之后的被遮挡的其他物体不会再接触到激光，造成被遮挡物体点云不完整甚至缺失。此外，激光雷达在扫描物体表面采集点云数据时，近处物体表面反射的点会比远处物体表面反射的多，导致采集到的点云密度不均。

② 自遮挡。即对某一物体，其靠近激光发射器的部分表面会自然地遮挡自身远离传感器的部分表面，造成同一物体点云数据的缺失。

③ 信号缺失。部分物体表面材料以及某些特殊的反射角度，会使得激光照射后无法被反射，导致点云不完整或空洞缺失[15]。

点云的稀疏性会导致目标检测时特征提取困难。由于前景物体的有效特征只占整个场景特征的一小部分，若直接借鉴2D目标检测中的卷积操作对所有场景点云进行处理，会存在较多的无效卷积操作，计算消耗和内存占用也非常庞大。为了降低计算量，减少冗余操作，在处理点云数据时通常会对其进行降采样操作，但降采样操作可能会造成感兴趣目标特征的丢失。

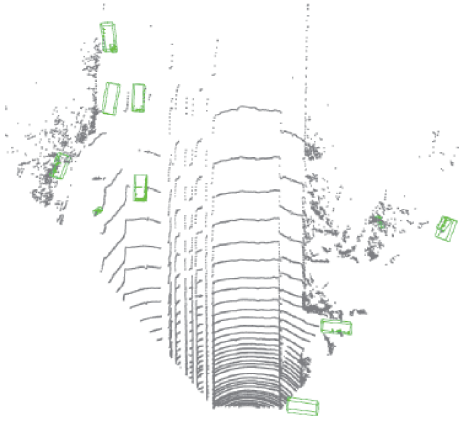

图 1　点云的稀疏性

Fig. 1　Sparsity of point cloud

1.2.2　无序性

区别于2D图像像素的有序排列，点云数据是空间中的无序点集合，其点的排列在空间上没有固定顺序。点云包含的特征与空间信息和点的储存顺序无关，即由不同设备或同一设备在不同位置采集到的点云所代表的含义是相同的。点云中的任意点互换位置或整个点云平移与旋转，其代表的物体的形状和大小不会改变。因此，要求处理点云的网络需要对不同的排列顺序以及变换具有不变性[16]。点云的无序性与变换不变性如图2所示。

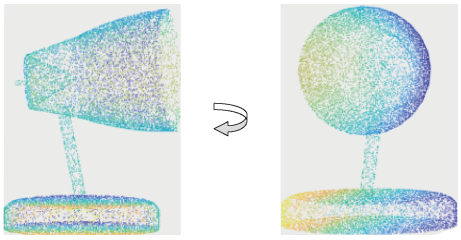

图 2　点云的无序性与变换不变性

Fig. 2　Unordered arrangement and invariance under transformations of point cloud

1.2.3　非结构化

区别于2D图像像素规则排列，点云中的每个点在空间中离散分布、独立存在，是一种不规则的数据。卷积算子难以利用数据中的空间局部相关性，导致无法直接在点云上应用2D目标检测网络处理点云数据来提取特征。因此，需要将不规则点云数据规则化后再进行卷积，或者利用多层感知机或图结构表示直接对点云进行处理[17]。

2　点云的处理方法

基于以上点云的特性分析，对点云数据进行恰当的处理以有效地提取特征是关乎3D目标检测模型性能的关键。目前，点云数据的处理方法主要分为原始点云、体素(柱)、投影视图、图结构等。

2.1　原始点云

原始点云包含精确的位置坐标(x,y,z)，其深度可由$\sqrt{x^2+y^2+z^2}$给出，将原始点云作为输入数据进行特征提取，可以尽可能多地保留点云的原始信息。例如PointNet[18]与PointNet++[19]等方法，使用直接对点操作的特征提取骨干网络进行点云处理，使原始的点云数据可以直接输入到网络进行预测。然而这种方法存在冗余，需要巨大的计算开销。因此，一些方法将点云转换为体素或者投影为2D图像后再送入后续处理模块。

2.2　体素

为了将原始点云规则化表示，一些方法借鉴2D图像的像素表示，将3D空间离散为3D体素，即将3D空间划分为多个堆叠的、大小相同的立方块，以便利用3D卷积操作提取特征。

给定激光雷达点云输入$P=\{p_1,p_2,\cdots,p_N\}$和点云沿z，y，x轴分别包含范围为(D,H,W)的3D空间，定义每个体素的大小为(v_D,v_H,v_W)，故生成的体素数目为$(D/v_D,H/v_H,W/v_W)$。体素化就是将点云P中的点分配给体素的过程。由于点云的稀疏性以及密度不均匀性，会产生包含不同数量点的体素。若一个体素内没有被分配到任何点，称其为空体素，反之若至少有一个点被分配到体素中，则称其为非空体素。体素化过程将点云转换为规则的体素网格，便于后续的高效处理[20]。但体素化过程将不可避免地带来信息丢失，信息丢失程度与分辨率(即体素大小)和其可容纳的最大点数密切相关[21]。

此外，与体素类似，柱的表示不考虑3D空间在z轴上的划分，即将3D空间划分为固定尺寸的柱体，再将点云分配到相应柱体中[22]。这种表示可被视为某种多通道的鸟瞰图(Bird's-Eye-View, BEV)，后续可采用2D卷积，大大提高了运行速度。

2.3 投影视图

将3D点云投影到2D平面上，就可利用2D目标检测网络进行检测。常用的投影视图包括距离视图和鸟瞰图两类。

2.3.1 距离视图

将点云投影到柱面或球面，再将柱面或球面的表面展开就形成了距离视图(Range-View，RV)[23]。给定一个3D点$p=(x,y,z)$，它投影到距离视图上的坐标(r,c)计算如式(1)所示：

$$\begin{aligned}\theta&=a\tan 2(y,x)\\ \varphi&=\arcsin(z/\sqrt{x^2+y^2+z^2})\\ r&=\lfloor\theta/\Delta\theta\rfloor\\ c&=\lfloor\varphi/\Delta\varphi\rfloor\end{aligned}\tag{1}$$

其中，θ和φ分别表示该点的水平观测角和垂直观测角，$\Delta\theta$和$\Delta\varphi$分别为激光束的水平和垂直分辨率。对于距离视图中每个像素坐标位置，通常将距离、坐标和强度编码为输入通道[24,25]。点云的距离视图表示将点云投影到2D图像上，起到了数据降维的作用，并且保留了丰富的原始信息，比点云表示更加紧凑。然而距离视图中不同距离的物体尺度相差较大，且存在物体相互遮挡的问题。

2.3.2 鸟瞰图

鸟瞰图是指将点云沿高度方向垂直投影到平面上所形成的视图。给定一个3D点$p=(x,y,z)$，将其离散到分辨率为λ米的2D网格中，其在BEV中的网格坐标$(r_{\mathrm{BEV}},c_{\mathrm{BEV}})$计算如式(2)所示。

$$\begin{aligned}r_{\mathrm{BEV}}&=\left\lfloor\frac{x}{\lambda}\right\rfloor\\ c_{\mathrm{BEV}}&=\left\lfloor\frac{y}{\lambda}\right\rfloor\end{aligned}\tag{2}$$

对于BEV中每个像素的坐标位置，通常将二进制占用编码、激光雷达密度、激光雷达点的统计值(如高度、强度的最大值)等作为输入通道[26]。鸟瞰图避免了前视图存在的尺度与遮挡问题，在鸟瞰图中物体会保持原有的尺度大小且彼此相互分离，互不干扰。鸟瞰图的不足之处是忽略了物体在z轴上的尺寸和位置，使得物体被限制在同一地平面上，而且难以表征行人等小目标。

2.4 图结构

图结构模型能够捕捉非结构化数据任意节点间的复杂关系，因此适合处理无序与非结构化的点云数据。给定激光雷达点云输入$P=\{p_1,p_2,...,p_N\}$，其中一个3D坐标为x_i，初始特征为s_i的点$p_i=(x_i,s_i)$，构造图表达式[27]为

$$\begin{aligned}G&=(P,E)\\ E&=\{(p_i,p_j)\,\big|\,\|x_i-x_j\|_2<r\}\end{aligned}\tag{3}$$

其中，r为固定的半径。图的构造过程是一种固定半径近邻搜索问题，即在距指定点的固定距离内找到欧几里得空间中的所有点。由于整个场景点云数量庞大，基于原始点云构造图的计算效率很低，因此，通常使用体素化后的下采样点云构造图。

3 基于点云的3D目标检测方法

基于以上对点云处理方法的分析，本文将基于点云的3D目标检测方法分为基于原始点云、基于体素、基于点云与体素融合、基于投影视图和基于图结构的方法。此外，介绍了几种具有代表性的基于点云与图像融合算法。图3为基于点云的3D目标检测的流程。

3.1 基于原始点云的方法

基于原始点云的方法主要思路为直接提取点云特征，最大限度地利用原始点云的几何信息。

2017年，Qi等人开创性地提出能够端到端地直接处理点云数据的神经网络PointNet[18]。网络的

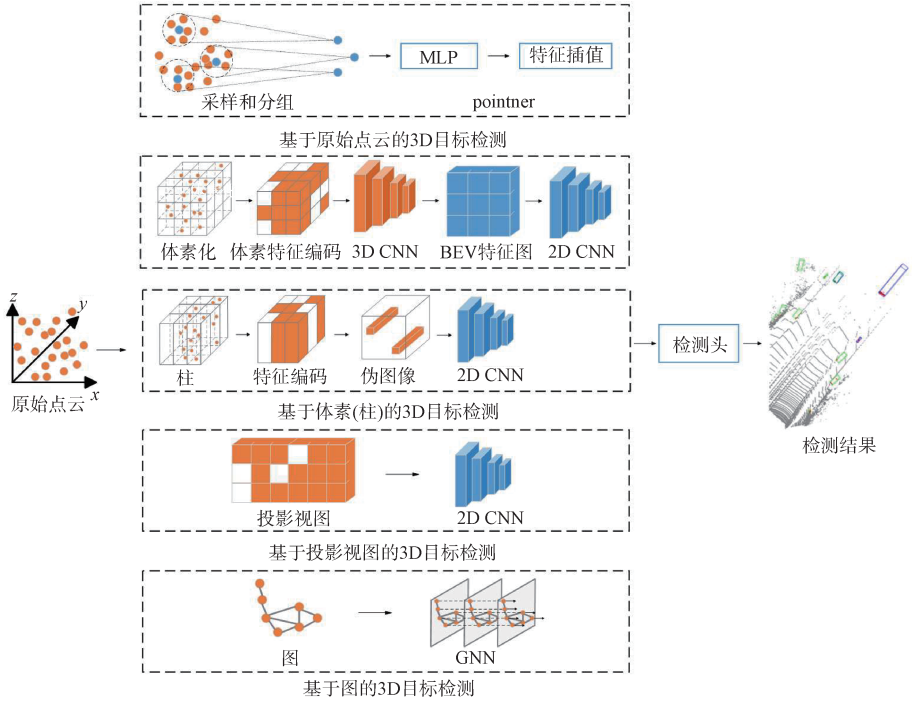

图 3　基于点云的 3D 目标检测流程

Fig. 3　The general pipelines of 3D object detection based on point cloud

输入为所有点云的坐标，经过一系列多层感知机(MLP)对所有点的特征进行编码，然后用 T-Net 预测一个变换矩阵来学习点云的变换不变性，并对特征进行对齐，进而使用最大池化对称函数(Max-pooling)聚合点的信息，生成全局特征向量，最后将其与每个点的特征向量进行融合。PointNet 解决了点云的无序性和变换不变性带来的问题，但忽略了点之间的相互关系，无法利用相邻点之间的局部特征，导致其不适用于复杂场景。随后，经过改进的 PointNet++[19]被提出，其中的集合提取(SA)层和特征传播(FP)层分别遵循编码器-解码器结构。在 SA 层提出多尺度分组(MSG)和多分辨率分组(MRG)将点层次分组，再利用 pointnet 结构提取点的局部特征，充分利用了上下文局部信息。在 FP 层通过插值将学习到的多尺度特征传播回所有点。

PointNet[18]和 PointNet++[19]简单有效，已成为后续基于原始点云方法的基础网络。如 PointRCNN[28]作为一种两阶段检测器，第一阶段将点云分割成前景点和背景点，在每个前景点上生成少量高质量的候选框。具体如下：首先使用 PointNet++[19]的编码器-解码器架构提取高维逐点特性，再进行点的前景与背景二值分割，同时利用基于 bin 的包围盒回归方法来预测三维感兴趣候选框 RoI，并采用非极大值抑制(NMS)剔除冗余 RoI(Region of Interest，感兴趣区域)，只保留高质量 RoI。第二阶段则利用池化等操作得到的点云特征和语义分割得到的分割掩码对候选框进行修正，得到预测结果。该方法虽得到了较好的性能，但推理速度较慢且资源占用较大。

而 3DSSD[29]是一种单阶段检测器，针对点云下采样造成前景点丢失的问题，提出一种新的融合采样方法，结合基于语义信息的特征最远点采样(Feature-FPS)和基于距离信息的距离最远点采样(Distance-FPS)，实现特征提取。为了提高计算效率，该方法去掉了之前基于点的方法中采用的特征传播(FP)层和 3D 包围盒细化的第二阶段，并设计了一个包含候选生成层、无锚回归头和三维中心度分配策略的包围盒预测网络，以弥补解码器部分的不足。该方法在达到与两阶段方法相当性能的同时极大地提升了运行速度。

基于点云的方法使用简单的对称函数聚合局部信息，缺乏对上下文信息交互的建模能力问题，对此，Pointformer[30]将 Transformer 架构引入 3D 点云的目标检测任务中，充分利用了 Transformer 模型学习上下文相关表示的能力和在处理集合结构

数据上的优势。该方法设计了 Local-Global Transformer(LGT，局部-全局转换器)，将局部特征与全局特征相结合。其中 Local Transformer(LT，局部转换器)对局部区域的点集提取特征，LGT 使用多尺度交叉注意机制整合两种分辨率的特征，Global Transformer(GT，全局转换器)学习上下文感知表示。

3.2 基于体素的方法

VoxelNet[20]是一种典型的基于体素的方法，其同时得益于稀疏的点结构与体素网格的高处理效率。在将点云划分为等间隔的体素后，提出体素特征编码(VFE)层对每个非空体素进行编码，将逐点特征与局部聚合特征相结合。具体如下：首先计算非空体素内所有点的均值将其作为质心，用每个点与质心的相对偏移量来增强点特征；然后将体素内的点送入全连接网络获得逐点特征，通过最大池化聚合后拼接到逐点特征上，得到此体素的输出特征。该网络通过堆叠 VFE 层实现了体素内点间的交互。通过 3D 卷积聚合体素特征，生成 3D 特征图，转换为 BEV 特征图后，送入区域候选生成网络(RPN)得到检测结果。但点云的稀疏性使得大部分体素为空体素，且空体素也参与运算，造成 3D 卷积的计算效率低，检测速度慢。

SECOND[21]对 VoxelNet[20]进行了改进，采用稀疏卷积来避免对无效的空体素进行运算，只对非空体素编码，大幅度提升了运行速率，成为后续广泛应用于 3D 目标检测的 3D 稀疏卷积范式。

为了进一步提升 3D 检测器的运行效率，PointPillars[22]设计了一种新的柱体划分方式来表示三维空间点云。该方法不对 z 轴进行划分，根据点云的 x 轴和 y 轴将点分配到一系列柱体中，对堆叠柱体提取特征并将其转化为 2D 伪图像，再利用 2D 卷积学习特征。由于避免使用高计算成本的 3D 卷积，PointPillars[22]实现了三倍于 SECOND[21]的运行速度，达到了 62 FPS。

Voxel R-CNN[31]是一种两阶段检测器。考虑到之前基于体素的方法在将 3D 体素特征转化为 BEV 特征表示后，并未恢复 3D 上下文结构信息，在第二阶段设计了体素 RoI 池化，从三维体素特征中提取感兴趣特征进行细化。具体来说，对于 3D 骨干网络的最后两个阶段的体素，利用体素查询与 pointnet 模块聚合相邻体素特征，将多阶段多尺度的聚合特征相连得到 RoI 特征。该方法在保持实时处理速度的同时提升了检测精度。

借鉴 CenterNet[32]在 2D 图像目标检测中的思想，CenterPoint[33]将关键点检测应用于 3D 目标检测中。第一阶段使用 VoxelNet[20]或 PointPillars[22]作为主干网络，利用关键点检测器查找对象中心，从中心位置的点特征回归到所有其他对象属性，如 3D 大小、方向和速度等。第二阶段利用预估的三维边界框的每个面三维中心的点特征进行细化。该方法更好地满足了点云检测网络的旋转不变性和等变性，提升了检测速度，简化了下游任务。

考虑到 3D 卷积骨干网络的感受野受限，无法有效利用上下文信息，Voxel Transformer(VoTr，体素转换器)[34]引入 Transformer 架构，提出用稀疏体素模块和子流形体素模块来分别处理空体素和非空体素，通过注意力机制构建体素之间的远程关系，提出一种快速体素查询算法，以加快非空体素的查询速度。

3.3 点云与体素融合的方法

基于体素的方法在计算上更加高效，但在体素化过程中会导致细粒度信息的丢失，无法充分利用点云的空间信息，一些方法将两种点云处理方式结合以兼顾两者的优点。

SA-SSD[35]是一种结构感知的单阶段检测器，在保持单阶段方法高效率的同时，显式地利用点云的细粒度结构信息以提高检测精度，同时设计了只在训练阶段使用的点监督辅助网络。具体如下：辅助网络首先将由 3D 卷积骨干网络得到的特征转换为逐点特征，再通过两个辅助网络学习点云结构信息，包括前景点分割和中心估计，在不增加推理成本的同时提升了单阶段网络的精度。

部分感知和部分聚合神经网络(Part-A^2 Net)[36]，是一种两阶段检测方法，是对 PointRCNN[28]的扩展。部分感知网络利用标签信息作为监督，进行前景点分割和前景点与对应检测框相对位置的预测，以减少检测框的冗余。针对之前方法的池化操作存在模糊性的问题，部分聚合网络利用 RoI-aware 池化操作对每个 3D 候选中的部分信息进行聚合，并利用第一阶段得到的部分感知信息进行细化。

PV-RCNN[37]将三维体素卷积神经网络和 PointNet[18]中的集合提取操作相结合，具有三维体素卷积的高效性和高质量候选，以及 PointNet[18]网络的精确位置信息和灵活感受野的优点。该方法提出

体素集合提取模块，将3D卷积得到的多尺度体素特征编码为一组关键点。具体如下：通过最远点采样(FPS)得到一组关键点后，以多个半径找出其相邻的非空体素，接着使用pointnet模块聚合体素特征，并融合原始点云特征和BEV特征，以弥补量化损失和扩大感受野。此外，该方法还提出了多尺度RoI特征提取层，将场景关键点特征聚合到RoI网格中，用于包围框置信度预测和定位细化。PV-RCNN++[38]对PV-RCNN[37]进行了改进，提出以候选区域为中心的分区关键点采样策略和Vector-Pool局部特征聚合模块，提高了检测速度。

3.4 基于投影视图的方法

3.4.1 基于距离视图的方法

VeloFCN[23]是基于距离视图方法的开创性方法，将激光雷达扫描点云投影为距离视图，引入2D目标检测中广泛使用的全卷积神经网络提取特征、预测目标置信度和边界框。

LaserNet[39]将激光雷达数据转换为距离视图后，使用全卷积网络预测每个点的类别概率，并在自顶向下视图中回归边界框的概率分布，再通过均值漂移聚类对边界框的分布进行组合，以降低单个预测的噪声。在训练时用边界框的四个角的参数约束训练，推理时则采用一种新的自适应非极大值抑制算法去掉重复的边界框分布。

RangeRCNN[25]是一种基于距离视图的两阶段方法。文献[25]认为距离视图是旋转激光雷达传感器的原生表示，保留了所有原始信息且具有致密紧凑的特性，因此，针对基于距离视图方法的几个缺点分别提出了相应策略。首先，对于距离视图的尺度变化问题，利用空洞卷积实现灵活的感受野。其次，由于距离视图存在尺度与遮挡问题，通过RV-PV-BEV模块将距离视图提取的特征转换为BEV特征，生成高质量候选区域。不同于基于BEV的方法，本方法已经从距离视图中提取无损特征后转化，减小了量化误差。然后，在第二阶段使用前一阶段得到的点特征进行细化。

RangeDet[40]针对基于距离视图的方法效果不佳的问题，分别提出相应改进策略。首先，针对距离视图中尺度随距离变化大的问题，设计了范围条件金字塔，将相近距离的目标交由同一层处理。其次，针对距离视图具有紧凑性，而当前方法未能有效利用这一特性生成高分辨率输出的问题，使用加权非极大值抑制得到最终边界框。再次，针对特征提取在2D空间而检测在3D空间的不一致导致几何信息丢失的问题，提出一种新的卷积操作，更好地利用距离视图中的3D几何信息。

RSN[41]则结合了基于距离视图和体素网格方法的优势。在第一阶段，将距离视图送入高效的2D卷积神经网络以提取特征、分割前景点。与传统的语义分割不同的是该网络重点关注查全率而非高精度。而在第二阶段，将前景点体素化后用3D稀疏卷积网络和改进的CenterNet[32]检测头作进一步处理。

3.4.2 基于鸟瞰图的方法

BirdNet[42]将激光雷达点云投影到BEV平面，使用改进的VGG-16架构作为特征提取网络，再经过RPN输出RoI。考虑到场景中部分目标在BEV中所占像素很小，因此，移除了第四个最大池化层，以减少降采样次数。此外，为了将2D边界框转化为3D边界框，采用后处理的方法，通过粗略估计地平面高度和提取BEV的高度通道来获得z值。BirdNet+[43]是BirdNet[42]的改进，是基于BEV表示的端到端3D目标检测器。在第一阶段，使用改进的ResNet-50架构提取BEV特征，经过RPN输出2D边界框。在第二阶段，通过两个完全连接层进行分类与3D边界框的回归，获得了明显优于之前版本的性能。

PIXOR[26]是一种专注于实时性的单阶段3D目标检测器。该方法用3D占用张量和激光雷达点反射率的组合得到BEV表示，同时将高度信息作为BEV特征的第三维度通道以防止丢失点云的空间信息。此外，利用特征金字塔网络架构进行特征提取，使用逐像素检测头预测边界框，该方法兼顾了运行速度和检测效果。

Complex-YOLO[44]将2D图像目标检测器YOLOv2扩展到3D目标检测领域，提出一种欧拉区域建议网络(E-RPN)估计每个框方向角的虚部和实部，目的是在回归角度时在数学空间中避免出现奇点。基于每个类别的预定义高度来完成从2D到3D的过渡。Complex-YOLO的速度达到了50 FPS，是VoxelNet[20]的10倍左右。

3.5 基于图结构的方法

PointRGCN[45]是首个利用图结构表示和图卷积网络的3D目标检测方法。首先，采用PointRCNN[28]中的RPN生成候选框。然后，利用图表示点云包含的几何信息，提出残差图卷积网络(R-

GCN)和上下文图卷积网络(C-GCN)。R-GCN聚合每个候选框中点云的特征信息，即通过扩展候选框，选取固定数量的点，对每个点用特征向量编码，并与相应的RPN特征连接，同时将每个图卷积层的输出特征经过最大池化后连接到每个点，作为此候选的特征。C-GCN聚合候选框之间的上下文信息，通过将R-GCN中的候选作为节点构造图，利用边缘卷积处理后将得到的全局特征与每个候选的局部特征连接。该方法取得了较好的效果，但由于图卷积网络本身的计算和内存消耗巨大，该方法运行实时性较差。

Point-GNN[27]是一种基于图神经网络的单阶段3D目标检测方法。考虑到用所有点作为节点构造图会带来巨大的计算负担，该方法使用体素下采样点云构造图，并对原始点云提取特征作为节点的初始状态值，以保留原始点云中的信息。为了增强图卷积网络的平移不变性，文献[27]提出一种自动配准机制，根据相邻节点的相对位置和状态值更新迭代节点的特征状态。为了得到更精确的检测结果，文献[27]设计了一种基于包围盒合并和评分的非极大值抑制(NMS)方法，以提高定位精度。Point-GNN展示了使用图神经网络进行3D检测的潜力。

PC-RGNN[46]针对Point-GNN[27]模型中没有考虑节点间的多尺度上下文信息问题，设计了一种注意力多尺度图神经网络模块，通过局部-全局注意力机制和多尺度图上下文信息聚合，以全面捕获点之间的几何关系。

针对之前方法无法有效处理密度不均和稀疏点云的问题，Graph R-CNN[47]提出了相应的策略。首先，通过动态最远体素采样适应点的不均匀分布；然后，提出感兴趣区域图池化，在每个3D目标候选中利用图神经网络迭代传递信息，更好地对上下文信息与挖掘点之间的关系进行建模，并将图节点投影到相机图像，通过双线性插值聚合节点对应于相机图像中像素处的特征向量，将其融合到节点特征。

3.6 点云与图像融合方法

3D点云数据能提供准确的深度信息和几何结构，但缺乏图像数据所具有的颜色、纹理等语义信息，使模型难以准确检测点云数目少的小物体与远距离物体，难以区分相近的物体[48−50]。因此，研究人员考虑将两种模态进行融合，同时利用点云数据与图像数据的优势，提高目标检测性能。根据融合阶段的不同，点云与图像融合的3D目标检测方法可分为数据级融合、特征级融合和决策级融合，如图4所示。

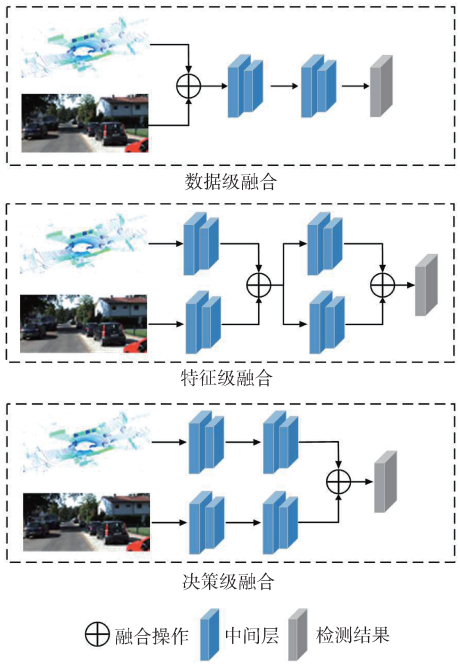

图4 多模态融合3D目标检测方法

Fig. 4 Multimodal fusion-based 3D object detection

3.6.1 数据级融合

数据级融合是在使用3D骨干网提取特征之前，通过空间对齐和投影直接融合点云和图像数据，或者将图像中的知识整合到点云中，使点云带有相应的语义信息或用图像语义作为先验信息选取点云。

Frustum PointNets(F-PointNet)[51]是一种顺序融合方法，通过2D目标检测结果截取需要的3D点云范围，缩小了搜索空间。具体如下：首先，利用2D目标检测器在图像上产生2D检测框，通过相机投影矩阵将其提升到3D截锥体，在截锥体中定义3D搜索空间；然后，通过基于PointNet[18]的网络对截锥体中的点云进行3D实例分割和边界框回归。通过使用2D图像先验知识，F-PointNet[51]在一些点稀疏和物体重叠的情况下也能取得不错的检测效果。然而3D搜索空间的确定过于依赖2D检测器，因此当2D检测结果出现问题时会影响3D目标检测。

MVX-Net[52](多模态体素网络)设计了两种融合路径。其一是点融合，将点云投影到图像平面，

用预训练的2D检测器提取特征，拼接到相应点上后送入VoxelNet[20]架构。该路径将图像中的知识附加到点云数据上，再用常规的体素网络进行检测。其二是体素融合，与点融合方法类似，只是先将点云体素化，将非空体素投影到图像上，用2D检测器提取特征后拼接到相应体素特征上，送入后续3D目标检测头。该路径将图像的感兴趣区域特征附加到体素特征上，属于特征级融合。

PointPainting[53]是一种顺序多模态融合方法。其思想是将激光雷达点云投影到相应图像的语义分割输出中，用图像的分割结果来“装饰”点云，即将图像分割分数作为新的维度与点云本身的坐标等维度结合，再送入3D目标检测网络进行检测。该方法适用于任何3D目标检测架构，但依赖于2D图像的分割效果。

针对小物体和远距离物体能够扫描到的激光雷达点云数量很少，但在相应的相机图像中通常可以被捕捉的情况，MVP[54](多模态虚拟点3D检测)利用图像语义分割结果补充虚拟点。具体如下：首先，对点云对应的2D图像进行实例分割得到分割掩码，并将点云投影到相机参考系作为实际点；然后，在每个2D实例分割掩码中进行随机采样作为虚拟点，其中虚拟点的深度根据与其距离最近的实际点的深度来确定；最后，将所有虚拟点投影回点云空间，作为原始点云的补充。这种方法一定程度上缓解了点云的稀疏性与密度不均带来的问题，但虚拟点的深度估计不够精确。

数据级融合可以更充分地利用不同模态数据的优势，但其大多数方法以顺序方式进行，时间成本较高。

3.6.2 特征级融合

特征级融合是在基于LiDAR的3D目标检测器的中间阶段，在特征空间中通过拼接、元素加法、平均等方式将点云特征与图像特征融合的方法。

MV3D[55](多视图3D目标检测网络)是首先使用多视图特征表示的多模态融合方法，使用点云的鸟瞰图(BEV)和前视图(FV)以及RGB(红绿蓝)图像作为输入。在第一阶段，BEV、FV和图像分别经过各自的骨干网络处理。其中，对BEV采用3D RPN生成3D候选框，并将其投影到三个视图以裁剪相应的建议区域。而在第二阶段，使用一个深度融合网络将经过RoI池化裁剪后的三种特征分层融合，对第一阶段的3D候选框进行修正。AVOD[56](聚合视图目标检测)针对MV3D[55]中提出的RPN架构检测小物体效果不佳的问题，提出将图像多分辨率特征与BEV特征图融合作为RPN的输入，可以为小物体生成高质量的候选。

3D-CVF[57](3D交叉视图融合)使用跨视图空间特征融合策略将相机特征和激光雷达特征融合。首先，采用自动校准投影，将2D相机特征转换为BEV域中的平滑空间特征图；然后，提出门控特征融合网络将相机特征和激光雷达特征融合。在后续的细化阶段，不仅使用融合特征，还使用第一阶段池化的多尺度相机特征和激光雷达特征，经过PointNet[18]编码，与融合后的特征一起用于产生最终的检测结果。相比于其他融合方法，该方法能确保以较快的速度获得较好的性能。

由于现有的融合方法是通过校准矩阵将激光雷达点和图像对齐的，因此易受到任一模态不良数据和传感器错位的影响。一些方法将Transformer机制引入到多模态融合的3D目标检测中，如TransFusion[58]由卷积主干网络和基于Transformer解码器的检测头组成。Transformer检测头使用稀疏的对象查询为激光雷达预测初始边界框，利用空间调制交叉注意力机制，将其与图像特征自适应关联融合。此外，提出一种图像引导的查询初始化策略，用于处理点云中难以检测到的目标。DeepFusion[59]提出InverseAug和LearnableAlign两个策略，其中，InverseAug反转几何相关的数据增强，以实现激光雷达点和图像像素之间的精确几何对准；LearnableAlign则利用交叉注意力在融合过程中动态捕捉图像和激光雷达特征之间的相关性。在流行的3D点云检测框架PointPillars[22]和CenterPoint[33]上提高了模型的识别和定位能力。

特征级融合能更加深入地融合多模态表示，但是相机和激光雷达的特征本质上是异构的，因此，如何有效对齐特征一直以来都是研究重点。

3.6.3 决策级融合

决策级融合是利用激光雷达点云分支和相机图像分支的输出结果进行最终预测。CLOCs[60]是决策级融合的典型代表。该方法分别使用2D检测器和3D检测器对图像和点云处理后生成相应的候选框，再将两种模态的候选框进行编码得到稀疏张量，然后利用2D卷积对稀疏输入张量中的非空元

素进行融合处理，通过关联两种模态的特征对3D候选框进行细化。这种融合方式避免了不同模态输入或中间特征的复杂交互，不会受到不同模态特征不对齐的限制，但未充分受益于不同模态的优势。

4 数据集与评价指标

4.1 数据集

3D目标检测常用的数据集包括室内数据集与室外数据集。这里主要介绍比较常用的几个室外数据集。

① KITTI

KITTI[61]数据集是使用最广泛的自动驾驶场景下的计算机视觉算法评测数据集之一，由德国卡尔斯鲁厄理工学院和丰田美国技术研究院联合创建。自动驾驶平台配备了两台高分辨率彩色和灰度摄像头、Velodyne HDL-64E激光雷达传感器和定位系统，采集于白天德国卡尔斯鲁厄的农村地区与高速公路。其中3D目标检测数据集包含7 481个训练集样本数据和7 518个测试样本数据，每个样本数据包括图像、点云、标注文件与标定文件，关注的对象为汽车(Car)、行人(Pedestrian)和骑自行车的人(Cyclist)，共包含80 256个标注对象。对于每个类别，检测结果基于三个难度级别进行评估，分别是容易(easy)、中等(moderate)和困难(hard)，这三个等级根据目标的大小、受遮挡状态和被截断级别确定。

② nuScenes

nuScenes[62]数据集是首个提供自动驾驶汽车全套传感器设备的大型数据集，由Motional团队在各种照明和天气条件下于波士顿和新加坡收集，传感器包括6个摄像头、1个激光雷达、5个毫米波雷达、GPS和IMU(惯性测量单元)。nuScenes数据集包含1 000个场景，每个场景20秒，以2 Hz的频率标注了23个类别目标的3D框、类别信息以及其他的一些属性(可见度、状态)，包含140万张相机图像，39万个激光雷达扫描，140万个毫米波雷达扫描，以及140万个3D框注释。在检测任务中，nuScenes数据集需要检测10个对象类别的3D边界框和属性。

③ Waymo

Waymo[63]数据集是一个用于自动驾驶任务的大规模激光雷达点云数据集，其采集设备由5个激光雷达和5个高分辨率RGB相机组成，在一天中的不同时间在旧金山等城市采集。整个数据集包含1 150个场景，每个场景时长为20秒，共有23万帧数据，以10 Hz的频率获取，划分为1 000个训练样本和159个测试样本，以距离图像的形式提供相机和激光雷达读数的同步。数据集对车辆、行人、骑自行车的人和标志四个类别进行标注，并将其划分为两个难度级别，若标注的3D检测框中的激光雷达点数大于5则为LEVEL_1(L1)，其余为LEVEL_2(L2)。

4.2 评价指标

3D目标检测任务中常用的评价指标有交并比(IoU，Intersection over Union)、查准率(Precision)、查全率(Recall)、平均精度(AP，Average Precision)和平均精度均值(mAP，mean Average Precision)。

① IoU

IoU表示检测框与真值框的交集与并集之比。IoU越高，检测越准确。计算如下式所示：

$$IoU=\frac{Area\ of\ Overlap(BBox_{\text{pred}},BBox_{\text{gt}})}{Area\ of\ Union(BBox_{\text{pred}},BBox_{\text{gt}})} \tag{4}$$

其中，$BBox_{\text{pred}}$为检测框，$BBox_{\text{gt}}$为真值框。

② 查准率和查全率

查准率P是指被预测为正样本中实际为正样本所占的比例，查全率R是指测试集中所有正样本被正确预测为正样本的比例，计算如下式所示：

$$\begin{aligned}P&=\frac{TP}{TP+FP}\\R&=\frac{TP}{TP+FN}\end{aligned} \tag{5}$$

其中，TP为判断检测结果中被正确识别的正样本的数量，FP为被错误识别为正样本的负样本数量，FN为被错误识别为负样本的正样本的数量。

③ AP和mAP

AP是Precision-Recall(PR)曲线下的面积。AP越高，检测精度越高。mAP是指对所有c个类别的平均精度Average Precision(AP)进行平均，计算如下式所示：

$$mAP=\frac{1}{|C|}\sum_{c\in C}AP_c \tag{6}$$

由于在3D检测任务中，IoU通常有两种计算方式，故AP也有相应的两种计算方式。具体如下：AP_{3D}对应于直接在3D空间中计算检测框和真值框的IoU，AP_{BEV}对应于将检测框和真值框投影到BEV上再计算IoU。

4.2.1　KITTI 数据集评价指标

KITTI[61]数据集提出平均方向相似性(Average Orientation Similarity，AOS)用于评价物体方向的检测结果，计算如下式所示：

$$AOS=\frac{1}{11}\sum_{r\in\{0,0.1,\cdots,1\}}\max_{\tilde{r}:\tilde{r}\geq r}s(\tilde{r}) \tag{7}$$

其中，r为PASCAL(结构化编程语言)目标检测的查全率，方向相似性$s(r)$计算如下：

$$s(r)=\frac{1}{|D(r)|}\sum_{i\in D(r)}\frac{1+\cos\Delta_\theta^{(i)}}{2}\delta_i \tag{8}$$

其中，$D(r)$表示查全率r下的所有目标检测结果的集合，是检测物体i的估计方向与真值方向的角度差。为了防止多个检出结果匹配到同一真值，如果检出目标i已经匹配到真值(IoU至少为50%)，$\delta_i=1$，否则$\delta_i=0$。

4.2.2　nuScenes 数据集评价指标

nuScenes[62]数据集通过在地平面上设定2D中心距离d的阈值匹配来度量AP，而不使用IoU，以将检测性能与对象的大小和方向解耦。计算如下式所示：

$$mAP=\frac{1}{|C||D|}\sum_{c\in C}\sum_{d\in D}AP_{c,d} \tag{9}$$

其中，C为类别集合，D为匹配距离阈值{0.5,1,2,4}。

除AP外，nuScenes[62]数据集为每个与真值框匹配的检测框测量了一组True Positive metrics(TP metrics，真正的正度量)，包括以下指标：平均平移误差(ATE)为二维欧几里得中心距离；平均尺度误差(ASE)为调整方向和平移后的3D IoU(1-IoU)；平均方向误差(AOE)为预测值和真实值之间的最小偏航角差；平均速度误差(AVE)为二维速度差的L2范数；平均属性误差(AAE)定义为1减去属性分类精度(1-acc)。对于每个TP指标，计算所有类的平均TP指标(mTP)：

$$mTP=\frac{1}{|C|}\sum_{c\in C}TP_c \tag{10}$$

其中，所有TP指标在匹配过程中使用d=2 m中心距离计算，每个指标为召回率达到10%的累计均值。

nuScenes[62]数据集提出nuScenes检测分数(NDS)，同时考虑检测性能与检测框的位置、大小、方向、属性和速度，计算如下式所示：

$$NDS=\frac{1}{10}\left[5mAP+\sum_{mTP\in TP}(1-\min(1,mTP))\right] \tag{11}$$

4.2.3　Waymo[63]数据集评价指标

由于通用的目标检测的平均精度AP没有考虑方向，Waymo[63]数据集提出APH，加入了方向信息。AP和APH计算如下：

$$\begin{aligned}AP&=100\int_0^1\max\{p(r')|r'\geq r\}\mathrm{d}r\\APH&=100\int_0^1\max\{h(r')|r'\geq r\}\mathrm{d}r\end{aligned} \tag{12}$$

其中，$p(r)$是PR曲线，$h(r)$的计算与$p(r)$类似，其使用的TP由方向精度$\min(|\tilde{\theta}-\theta|,2\pi-|\tilde{\theta}-\theta|)/\pi$加权，$\tilde{\theta}$和$\theta$分别是预测方向和真值方向。

5　性能对比与分析

本节对比了不同类型检测方法在KITTI[61]、nuScenes[62]和Waymo[63]数据集上的性能，分别如表1、表2和表3所示。

由表格可以看出，基于原始点云方法的精度总体高于基于体素的方法，但由于处理原始点云带来的巨大计算消耗，更多方法选择采用基于体素的处理方式。早期基于体素的方法精度较低，这是因为体素化过程不可避免地会引入量化误差。而一些方法通过两阶段方法有效地聚合细粒度信息作为细化，或者结合点云和体素两者的优势，或者引入Transformer机制，以增加推理时间为代价来提高体素方法的准确性[64-66]。基于鸟瞰图的方法虽然速度较快但精度普遍较低，即使加入一些策略在一定程度上牺牲速度以提升精度也无法得到令人满意的效果。这是由于投影引入的不可逆的信息损失总是存在。基于距离视图的方法相对于基于鸟瞰图的方法效果更好，因为距离视图相对地保留了更丰富的原始信息，但是距离视图中的尺度和遮挡问题难以解决。基于图结构的方法的检测性能总体令人满意，展示了该方向的研究潜力，但此类方法计算效率很低，难以满足实时性的要求。多模态方法性能总体高于单模态方法，但在KITTI[61]数据集上提升效果并不明显，如何更好地对齐不同模态的异构特征仍然需要进一步研究。

表 1 3D 目标检测方法在 KITTI 测试集的结果对比

Table 1 Comparison of 3D object detection methods on KITTI test set

%

类型	方法	Car			Pedestrian			Cyclist		
		Easy	Mod	Hard	Easy	Mod	Hard	Easy	Mod	Hard
基于原始点云	PointRCNN	86.96	75.64	70.70	47.98	39.37	36.01	74.96	58.82	52.53
	3DSSD	88.36	79.57	74.55	54.64	44.27	40.23	82.48	64.10	56.90
	Pointformer	87.13	77.06	69.25	50.67	42.43	39.60	75.01	59.80	53.99
基于体素	VoxelNet	77.47	65.11	57.73	39.48	33.69	31.51	61.22	48.36	44.37
	SECOND	83.13	73.66	66.20	51.07	42.56	37.29	70.51	53.85	46.90
	PointPillars	82.58	74.31	68.99	51.45	41.92	38.89	77.10	58.65	51.92
	Voxel R-CNN	90.90	81.62	77.06		—	—	—	—	—
	VoTr	89.90	82.09	79.14	—	—	—	—	—	—
基于体素与点云	SA-SSD	88.75	79.79	74.16	—	—	—	—	—	—
	Part-A^2 Net	87.81	78.49	73.51	53.10	43.35	40.06	79.17	63.52	56.93
	PV-RCNN	90.25	81.43	76.82	52.17	43.29	40.29	78.60	63.71	57.65
基于距离视图	RangeRCNN	88.47	81.33	77.09	—	—	—	—	—	—
	RangeDet	85.41	77.36	72.60	—	—	—	—	—	—
基于BEV	BirdNet	40.99	27.26	25.32	22.04	17.08	15.82	43.98	30.25	27.21
	BirdNet+	76.15	64.04	59.79	41.55	35.06	32.93	65.67	53.84	49.06
基于图	PointRGCN	85.97	75.73	70.60	—	—	—	—	—	—
	Point-GNN	88.33	79.47	72.29	51.92	43.77	40.14	78.60	63.48	57.08
	PC-RGNN	89.13	79.90	75.54	—	—	—	—	—	—
	Graph R-CNN	91.89	83.27	77.78	—	—	—	—	—	—
多模态融合	F-PointNet	82.19	69.79	60.59	50.53	42.15	38.08	72.27	56.12	49.01
	MVX-Net	83.20	72.70	65.20	—	—	—	—	—	—
	PointPainting	82.11	71.70	67.08	50.32	40.97	37.87	77.63	63.78	55.89
	MV3D	74.97	63.63	54.00	—	—	—	—	—	—
	AVOD	81.94	71.88	66.38	50.80	42.81	40.88	64.00	52.18	46.61
	3D-CVF	89.20	80.05	73.11	—	—	—	—	—	—
	CLOCs	89.16	82.28	77.23	—	—	—	—	—	—

表 2 3D 目标检测方法在 nuScenes 测试集的结果对比

Table 2 Comparison of 3D object detection methods on nuScenes test set

%

方法	mAP	NDS	Car	Ped	Bus	Barrier	TC	Truck	Trailer	Motor	CV	Bicycle
3DSSD	42.6	56.4	81.2	70.1	61.4	47.9	31.0	47.1	30.4	35.9	12.6	8.6
Pointformer	53.6		82.3	81.8	55.6	66.0	72.2	48.1	43.4	55.0	8.3	22.7
PointPillars	30.5	45.3	68.4	59.7	28.2	38.9	30.8	23.0	23.4	27.4	4.1	1.1
CenterPoint	58.0	65.5	84.6	83.4	60.2	70.9	76.7	51.0	53.2	53.7	17.5	28.7
BirdNet+	39.2		67.7	48.7	39.7	60.5	28.0	43.6	47.2	28.9	16.3	11.0
PointPainting	46.4	58.1	77.9	73.3	36.1	60.2	62.4	35.8	37.3	41.5	15.8	24.1
MVP	66.4	70.5	86.8	89.1	67.4	74.8	85.0	58.5	57.3	70.0	26.1	49.3
TransFusion	68.9	71.7	87.1	88.4	68.3	78.1	86.7	60.0	60.8	73.6	33.1	52.9

注：Ped—Pedestrian；TC—Traffic Cone；Motor—Motorcycle；CV—Construction Vehicle

表 3　3D 目标检测方法在 Waymo 测试集的结果对比

Table 3　Comparison of 3D object detection methods on Waymo test set

%

难度级别	方法	Vehicle		Pedestrian		Cyclist	
		AP	APH	AP	APH	AP	APH
LEVEL_1	SECOND	72.27	71.69	68.70	58.18	60.62	59.28
	PointPillars	56.62	—	59.25	—	—	—
	Voxel R-CNN	75.59	—	—	—	—	—
	CenterPoint	76.70	76.20	79.00	72.90	—	—
	VoTr	74.95	74.25	—	—	—	—
	Part-A^2 Net	77.05	76.51	75.24	66.87	68.60	67.36
	PV-RCNN	77.51	76.89	75.01	65.65	67.81	66.35
	PV-RCNN++	79.25	78.78	81.83	76.28	73.72	72.66
	LaserNet	52.11	50.05	63.40	—	—	—
	RangeRCNN	75.43	74.97	—	—	—	—
	RangeDet	72.85	—	75.94	—	—	—
	RSN	78.40	78.10	79.40	76.20	—	—
	Graph R-CNN	80.77	80.28	82.35	76.64	75.28	74.21
	DeepFusion	83.60	83.20	87.10	84.70	—	—
LEVEL_2	SECOND	63.85	63.33	60.72	51.31	58.34	57.05
	Voxel R-CNN	66.59	—	—	—	—	—
	CenterPoint	68.80	68.30	71.00	65.30	—	—
	VoTr	65.91	65.29	—	—	—	—
	Part-A^2 Net	68.47	67.97	66.18	58.62	66.13	64.93
	PV-RCNN	68.98	68.41	66.04	57.61	65.39	63.98
	PV-RCNN++	70.61	70.18	73.17	68.00	71.21	70.19
	RSN	69.50	69.10	69.90	67.00	—	—
	Graph R-CNN	72.55	72.10	74.44	69.02	72.52	71.49
	TransFusion	—	65.10	—	64.00	—	67.40
	DeepFusion	76.00	75.60	80.40	78.10	—	—

6　研究趋势和挑战

尽管 3D 点云目标检测已经得到了广泛的研究与应用，但其性能要得到进一步提升与落地应用仍然存在许多挑战。

6.1　数据集的局限性与开放世界中的 3D 目标检测

目前的研究主要集中在主流数据集上寻求性能的突破，但这些数据集在场景多样性与目标的类别数量上具有很强的局限性。首先，数据大部分来源于交通流量压力不大的城市或农村道路场景，且这些道路大都平坦无斜坡，而现实中很多道路有一定的倾角，检测的物体与车辆自身可能并不总是在同一地平面上，且存在交通流量压力突出的情况。其次，数据大部分采集于良好的天气情况下，但现实中雨雪和雾天不可避免，激光雷达在恶劣天气中衰减加大，容易产生噪声点，影响算法的鲁棒性。此外，数据集中感兴趣的目标往往只有车、人等且分布不均，在现实交通场景中存在很多未知障碍，未知类别也会影响行驶的安全性。尽管 nuScenes[62] 和 Waymo[63] 数据集已经涵盖了部分特殊场景，且目前更多的算法考虑在这两种更大规模的数据集上进行测试，但仍然不足以满足复杂开放世界中的应用，这使得现有的 3D 目标检测器处理未知场景与识别未知目标的能力有限。

目前，在 2D 目标检测中出现了一些针对开放世界目标检测的研究工作，其任务是在检测已知的目标类别的同时还能识别场景中未知的目标。如 OWOD[67](面向开放世界的目标检测)提出一种基于对比聚类和能量的开放世界目标检测器，实现

了在没有明确监督的情况下，将未被引入的对象识别为“未知”，且在不忘记之前学习过的类的情况下，在接收相应标签的同时逐步学习这些未知类别。而在3D目标检测领域，也出现一些针对开放世界的检测[68,69]。如CoDA[70]设计了一种协同式3D新目标发掘与跨模态对齐方法，在训练中同时学习对新类别目标的定位和分类，对开放词汇3D目标检测中的新类别目标的定位和分类问题进行研究。然而，基于开放世界的3D目标检测方法仍处于起步阶段，设计满足复杂开放世界中的3D目标检测器仍是一个值得研究的方向。

近年来，随着硬件计算能力的不断提升、大数据的涌现以及深度学习的迅猛发展，大模型成为机器学习各领域的重要研究方向和热点话题[71,72]。与其他模型相比，大模型具有更好的性能和泛化能力，可以实现自监督学习和多任务学习。因此，将3D目标检测与大模型相结合，以解决现有数据集的局限性，实现复杂真实世界中的高性能检测，是未来的发展方向。

此外，由于难以对大量数据进行准确详细的标注，3D目标检测器的自监督和半监督学习[73−75]也是未来的研究趋势。

6.2 点云的稀疏性

激光雷达点云本质上是稀疏的，且存在遮挡与密度不均的问题，导致对远处或点数很少的目标特征提取困难，难以精确检测3D框甚至丢失目标。此外，整个场景点云庞大，各类方法为了提升特征提取的效率都采用了不同程度的下采样，使得点云的稀疏性更加突出。因此，设计更有效的特征提取方式，尽可能保留有用的稀疏点的信息还需要进一步研究。此外，利用多帧点云的时序信息和互补信息可以得到更完整的目标结构和更精确的定位信息，从而缓解点云稀疏性带来的影响，值得进一步探讨。一些方法提出对稀疏点云进行补全[76−78]或者从局部点云推断完整形状[79,80]，以补全被遮挡的不完整目标或小目标点云，也是未来的研究方向之一。

6.3 多模态融合算法

多模态数据信息的互补性有助于提升模型的检测性能与鲁棒性。相机图像具有丰富的颜色和纹理信息，可以补充激光雷达点云数据的语义信息。但在光照条件不佳时，相机往往难以有效记录信息，而激光雷达虽然不易受到光照等环境因素的影响，但得到的点云比较稀疏。因此，将不同模态的数据融合可以充分利用不同传感器对同一环境的感知信息，为缓解点云的稀疏性与恶劣天气下单一传感器的局限性提供了可能方案[81−83]。

然而目前多模态信息融合还面临诸多挑战，包括不同测量空间数据的对齐与信息的丢失等问题。由于来自图像和点云的语义特征本质上是异构的，而且激光雷达点与图像像素的转换是多对一映射，在实际应用中特征的对齐和信息的配准都面临挑战。此外，在实际场景中，融合系统还需考虑时间戳、多传感器的校准匹配与合作的优先级等。

6.4 实时性

目前，大多数方法更多地关注检测精度的提升，忽略了在嵌入式系统中的边缘设备上的运行情况。但在实际落地应用中，3D目标检测的实时性是至关重要的。为了确保安全和高效，系统必须能够在极短的时间内处理传感器数据并做出决策。因此，考虑在嵌入式系统中进行部署的效率必不可少。尤其是在自动驾驶场景中，实时地检测目标是3D目标检测最重要的挑战之一。

因此，为了在实际场景中实现高效的模型部署，未来工作还需要探索特征的高效提取与融合以及模型的修剪和量化等技术[84−86]，以平衡对精度和实时性的要求。

7 结束语

随着全球汽车产业规模的日益扩大和产业分工格局的不断变革，自动驾驶技术作为现代工业转型升级、助推产业和国家工业实力的重要体现，得到了各国各界研究机构与制造公司的高度重视和广泛研究。基于激光雷达点云的3D目标检测是自动驾驶汽车感知系统的重要组成部分，负责对车辆周边物体的类别、位置、大小和方向进行检测，为后续路径规划与控制决策提供可靠的环境信息。本文对基于激光雷达点云的3D目标检测进行综述，包括数据的特点与处理、基于不同处理方法的检测算法、常用数据集和评价指标以及各类算法的性能对比分析。在未来，激光雷达3D目标检测技术的发展同时面临机遇与挑战，需要对其进一步研究与探索。

参考文献

[1] QIAN R, LAI X, LI X. 3D object detection for autonomous driving: A survey[J]. Pattern Recognition, 2022, 130: 108796.

[2] 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 Hanqing, et al. A survey on deep learning based visual object detection[J]. Journal of Image and Graphics, 2022, 27(6): 1697-1722.

[3] MAO J, SHI S, WANG X, et al. 3D object detection for autonomous driving: A comprehensive survey[J]. International Journal of Computer Vision, 2023, 131(8): 1909-1963.

[4] FERNANDES D, SILVA A, NEVOA R, et al. Point-cloud based 3D object detection and classification methods for self-driving applications: A survey and taxonomy[J]. Information Fusion, 2021, 68: 161-191.

[5] 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 Meiying, YUAN Zhengqian, et al. 3D object detection algorithms in autonomous driving: A review[J]. Control and Decision, 2023, 38(4): 865-889.

[6] 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.
HUANG Zhe, WANG Yongcai, LI Deying. A survey of 3D object detection algorithms[J]. Chinese Journal of Intelligent Science and Technology, 2023, 5(1): 7-31.

[7] 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): 286-302.
LI Jianan, WANG Ze, XU Tingfa. Three-dimensional object detection technology based on point cloud data[J]. Acta Optica Sinica, 2023, 43(15): 286-302.

[8] ZAMANAKOS G, TSOCHATZIDIS L, AMANATIADIS A, et al. A comprehensive survey of LIDAR-based 3D object detection methods with deep learning for autonomous driving[J]. Computers & Graphics, 2021, 99: 153-181.

[9] WU Y, WANG Y, ZHANG S, et al. Deep 3D object detection networks using LiDAR data: A review[J]. IEEE Sensors Journal, 2020, 21(2): 1152-1171.

[10] RAJ T, HASHIM F H, HUDDIN A B, et al. A survey on LiDAR scanning mechanisms[J]. Electronics, 2020, 9(5): 741.

[11] RORIZ R, CABRAL J, GOMES T, et al. Automotive LiDAR technology: A survey[J]. IEEE Transactions on Intelligent Transportation Systems, 2022, 23(7): 6282-6297.

[12] LI Y, IBANEZ-GUZMAN J. Lidar for autonomous driving: The principles, challenges, and trends for automotive lidar and perception systems[J]. IEEE Signal Processing Magazine, 2020, 37(4): 50-61.

[13] ROYO S, BALLESTA-GARCIA M. An overview of lidar imaging systems for autonomous vehicles[J]. Applied sciences, 2019, 9(19): 4093.

[14] CAMUFFO E, MARI D, MILANI S. Recent advancements in learning algorithms for point clouds: An updated overview[J]. Sensors, 2022, 22(4): 1357.

[15] XU Q, ZHONG Y, NEUMANN U. Behind the curtain: Learning occluded shapes for 3D object detection[C]// AAAI Conference on Artificial Intelligence, 2021.

[16] ZHANG H, WANG C, TIAN S, et al. Deep learning-based 3D point cloud classification: A systematic survey and outlook[J]. Displays, 2023, 79: 102456.

[17] 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 of 3D point cloud analysis methods based on deep learning[J]. Chinese Journal of Scientific Instrument, 2023, 44(11): 130-158.

[18] QI C R, SU H, MO K, et al. PointNet: Deep learning on point sets for 3D classification and segmentation[C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2017: 77-85.

[19] Qi C R, YI L, SU H, et al. PointNet++: Deep hierarchical feature learning on point sets in a metric space[C]// Conference on Neural Information Processing Systems. 2017: 5099-5108.

[20] ZHOU Y, TUZEL O. VoxelNet: End-to-end learning for point cloud based 3D object detection[C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2018: 4490-4499.

[21] YAN Y, MAO Y, LI B. Second: Sparsely embedded convolutional detection[J]. Sensors, 2018, 18(10): 3337.

[22] LANG A H, VORA S, CAESAR H, et al. PointPillars: Fast encoders for object detection from point clouds[C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2019: 12689-12697.

[23] LI B, ZHANG T, XIA T. Vehicle detection from 3D lidar using fully convolutional network[C]// Conference on Robotics: Science and Systems, 2016.

[24] MILIOTO A, VIZZO I, BEHLEY J, et al. RangeNet++: Fast and accurate LiDAR semantic segmentation[C]// IEEE/RSJ International Conference on Intelligent Robots and Systems. New York: IEEE, 2019: 4213-4220.

[25] LIANG Z, ZHANG M, ZHANG Z, et al. RangeRCNN: Towards fast and accurate 3D object detection with range image representation[EB/OL]. (2021-03-23)[2024-06-04].https://urxiv.org/abs/2009.00206.

[26] YANG B, LUO W, URTASUM R. PIXOR: Real-time 3D object detection from point clouds[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2018: 7652-7660.

[27] SHI W, RAJKUMAR R. Point-GNN: Graph neural network for 3D object detection in a point cloud[C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2020: 1711-1719.

[28] SHI S, WANG X, LI H. PointRCNN: 3D object proposal generation and detection from point cloud[C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2019: 770-779.

[29] YANG Z, SUN Y, LIU S, et al. 3DSSD: Point-based 3D single stage object detector[C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2020: 11037-11045.

[30] PAN X, XIA Z, SONG S, et al. 3D object detection with pointformer[C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2021: 7463-7472.

[31] DENG J, SHI S, LI P, et al. Voxel R-CNN: Towards high performance voxel-based 3D object detection[C] // AAAI Conference on Artificial Intelligence. 2021: 1201-1209.

[32] ZHOU X, WANG D, KRAHENBUHL P. Objects as points[EB/OL]. (2019-04-25)[2024-06-04]. https://arxiv.org/abs/1904.07850.

[33] YIN T, ZHOU X, KRAHENBUHL P. Center-based 3D object detection and tracking[C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021: 11784-11793.

[34] MAO J, XUE Y, NIU M, et al. Voxel transformer for 3D object detection[C]// IEEE/CVF International Conference on Computer Vision. 2021: 3164-3173.

[35] HE C, ZENG H, HUANG J, et al. Structure aware single-stage 3D object detection from point cloud[C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2020: 11870-11879.

[36] SHI S, WANG Z, SHI J, 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.

[37] SHI S, GUO C, JIANG L, et al. PV-RCNN: Point-voxel feature set abstraction for 3D object detection[C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2020: 10526-10535.

[38] SHI S, JIANG L, DENG J, et al. PV-RCNN++: Point-voxel feature set abstraction with local vector representation for 3D object detection[J]. International Journal of Computer Vision, 2023, 131(2): 531-551.

[39] MEYER G P, LADDHA A, KEE E, et al. Lasernet: An efficient probabilistic 3D object detector for autonomous driving[C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2019: 12677-12686.

[40] FAN L, XIONG X, WANG F, et al. RangeDet: In defense of range view for LiDAR-based 3D object detection[C]// IEEE/CVF International Conference on Computer Vision. New York: IEEE, 2021: 2898-2907.

[41] SUN P, WANG W, CHAI Y, et al. RSN: Range sparse net for efficient, accurate LiDAR 3D object detection [C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2021: 5721-5730.

[42] BELTRAN J, GUINDEL C, MORENO F M, et al. Birdnet: A 3D object detection framework from lidar information[C]// International Conference on Intelligent Transportation Systems. 2018: 3517-3523.

[43] BARRERA A, BELTRAN J, GUINDEL C, et al. Birdnet+: Two-stage 3D object detection in lidar through a sparsity-invariant bird's eye view[J]. IEEE Access, 2021, 9: 160299-160316.

[44] SIMON M, MILZ S, AMENDE K, et al. Complex-YOLO: An euler-region-proposal for real-time 3D object detection on point clouds[C]// European Conference on Computer Vision. Cham: Springer, 2019: 197-209.

[45] ZARZAR J, GIANCOLA S, GHANEM B. PointRGCN: Graph convolution networks for 3D vehicles detection refinement[EB/OL]. (2019-11-27) [2024-06-04]. https://arxiv.org/abs/1911.12236..

[46] ZHANG Y, HUANG D, WANG Y. PC-RGNN: Point cloud completion and graph neural network for 3D object detection[C]// AAAI Conference on Artificial Intelligence. 2021, 35(4): 3430-3437.

[47] YANG H, LIU Z, WU X, et al. Graph R-CNN: Towards accurate 3D object detection with semantic-decorated lo-

cal graph[C]// European Conference on Computer Vision. 2022: 662-679.

[48] WANG Y, MAO Q, ZHU H, et al. Multi-modal 3D object detection in autonomous driving: A survey[J]. International Journal of Computer Vision, 2023, 131(8): 2122-2152.

[49] WANG L, ZHANG X, SONG Z, et al. Multi-modal 3D object detection in autonomous driving: A survey and taxonomy[J]. IEEE Transactions on Intelligent Vehicles, 2023, 8(7): 3781-3798.

[50] HUANG K, SHI B, LI X, et al. Multi-modal sensor fusion for auto driving perception: A survey[EB/OL]. (2022-02-27) [2024-06-04]. https://arxiv. org/abs/2202. 02703.

[51] QI C R, LIU W, WU C, et al. Frustum PointNets for 3D object detection from RGB-D data[C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2018: 918-927.

[52] SINDAGI V A, ZHOU Y, TUZEL O. MVX-Net: Multimodal voxelnet for 3D object detection[C]// International Conference on Robotics and Automation. 2019: 7276-7282.

[53] VORA S, LANG A H, HELOU B, et al. PointPainting: Sequential fusion for 3D object detection[C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2020: 4603-4611.

[54] YIN T, ZHOU X, KRAHENBUHL P. Multimodal virtual point 3D detection[C]// Conference on Neural Information Processing Systems. 2021.

[55] CHEN X, MA H, WAN J, et al. Multi-view 3D object detection network for autonomous driving[C]// IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2017: 6526-6534.

[56] KU J, MOZIFIAN M, LEE J, et al. Joint 3D proposal generation and object detection from view aggregation [C]// IEEE/RSJ International Conference on Intelligent Robots and Systems. New York: IEEE, 2018: 1-8.

[57] YOO J H, KIM Y, KIM J, et al. 3D-CVF: Generating joint camera and LiDAR features using cross-view spatial feature fusion for 3D object detection[C]// European Conference on Computer Vision. 2020: 720-736.

[58] BAI X, HU Z, ZHU X, et al. Transfusion: Robust lidar-camera fusion for 3D object detection with transformers [C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2022: 1090-1099.

[59] LI Y, YU A W, MENG T, et al. DeepFusion: Lidar-camera deep fusion for multi-modal 3D object detection [C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2022: 17161-17170.

[60] PANG S, MORRIS D, RADHA H. CLOCs: Camera-LiDAR object candidates fusion for 3D object detection [C]//IEEE/RSJ International Conference on Intelligent Robots and Systems. New York: IEEE, 2020: 10386-10393.

[61] GEIGER A, LENZ P, URTASUM R. Are we ready for autonomous driving? The KITTI vision benchmark suite [C]// IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2012: 3354-3361.

[62] CAESAR H, BANKITI V, LANG A H, et al. NuScenes: A multimodal dataset for autonomous driving[C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2020: 11621-11631.

[63] SUN P, KRETZSCHMAR H, DOTIWALLA X, et al. Scalability in perception for autonomous driving: Waymo open dataset[C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2020: 2446-2454.

[64] GUAN T, WANG J, LAN S, et al. M3DETR: Multi-representation, multi-scale, mutual-relation 3D object detection with transformers[C]// IEEE/CVF Winter Conference on Applications of Computer Vision. New York: IEEE, 2022: 772-782.

[65] ERABATI G K, ARAUJO H. LI3DETR: A lidar based 3D detection transformer[C]// IEEE/CVF Winter Conference on Applications of Computer Vision. New York: IEEE, 2023: 4250-4259.

[66] LIU C, QIAN X, HUANG B, et al. Multimodal transformer for automatic 3D annotation and object detection [C]//European Conference on Computer Vision. 2022: 657-673.

[67] JOSEPH K J, KHAN S, Khan F S, et al. Towards open world object detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2021: 5830-5840.

[68] CEN J, YUN P, CAI J, et al. Open-set 3D object detection[C]// International Conference on 3D Vision. 2021: 869-878.

[69] ALLIEGRO A, CAPPIO BORLINO F, TOMMASI T. 3DOS: Towards 3D open set learning-benchmarking and understanding semantic novelty detection on point clouds[J]. Advances in Neural Information Processing Systems. 2022, 35: 21228-21240.

[70] CAO Y, ZENG Y, XU H, et al. Coda: Collaborative

novel box discovery and cross-modal alignment for open-vocabulary 3D object detection[EB/OL]. (2023-10-04)[2024-06-04]. https://arxiv.org/abs/2310.02960.

[71] ACHIAM J, ADLER S, AGARWAL S, et al. GPT-4 technical report[EB/OL]. (2024-03-04) [2024-06-04]. https://arxiv.org/abs/2303.08774.

[72] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[J]. arXiv preprint arXiv: 2010.11929, 2020.

[73] YANG J, SHI S, WANG Z, et al. ST3D: Self-training for unsupervised domain adaptation on 3D object detection [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2021: 10368-10378.

[74] LI H, YANG J, XU Y, et al. A level set annotation framework with single-point supervision for infrared small target detection[J]. IEEE Signal Processing Letters, 2024, 31: 451-455.

[75] LI H, YANG J, XU Y, et al. Click on mask: A labor-efficient annotation framework with level set for infrared small target detection[EB/OL]. (2023-10-19) [2024-06-04]. https://arxiv.org/abs/2310.12562.

[76] YUAN W, KHOT T, HELD D, et al. PCN: Point completion network[C]// International Conference on 3D Vision. 2018: 728-737.

[77] YU X, RAO Y, WANG Z, et al. Pointr: Diverse point cloud completion with geometry-aware transformers[C]// IEEE/CVF International Conference on Computer Vision. New York: IEEE, 2021: 12498-12507.

[78] WU X, PENG L, YANG H, et al. Sparse fuse dense: Towards high quality 3D detection with depth completion [C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2022: 5418-5427.

[79] WANG T, HU X, LIU Z, et al. Sparse2Dense: Learning to densify 3D features for 3D object detection[J]. Advances in Neural Information Processing Systems, 2022, 35: 38533-38545.

[80] QIN Y, WANG C, KANG Z, et al. Supfusion: Supervised lidar-camera fusion for 3D object detection[C]// IEEE/CVF International Conference on Computer Vision. New York: IEEE, 2023: 22014-22024.

[81] ZHU M, MA C, JI P, et al. Cross-modality 3D object detection[C]// IEEE/CVF Winter Conference on Applications of Computer Vision. New York: IEEE, 2021: 3772-3781.

[82] CAI Q, PAN Y, YAO T, et al. Objectfusion: Multi-modal 3D object detection with object-centric fusion[C]// IEEE/CVF International Conference on Computer Vision. New York: IEEE, 2023: 18067-18076.

[83] XIE Y, XU C, RAKOTOSAONA M J, et al. Sparsefusion: Fusing multi-modal sparse representations for multi-sensor 3D object detection[C]// IEEE/CVF International Conference on Computer Vision. New York: IEEE, 2023: 17591-17602.

[84] ZHANG L, DONG R, TAI H S, et al. Pointdistiller: Structured knowledge distillation towards efficient and compact 3D detection[C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2023: 21791-21801.

[85] ZHENG W, TANG W, JIANG L, et al. SE-SSD: Self-ensembling single-stage object detector from point cloud [C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2021: 14494-14503.

[86] YANG J, SHI S, DING R, et al. Towards efficient 3D object detection with knowledge distillation[J]. Advances in Neural Information Processing Systems, 2022, 35: 21300-21313.

[作者简介]

武淑文　1999年生，硕士研究生。

李燕烯　2003年生，本科。

张少琛　1999年生，硕士研究生。

杨金福　1977年生，教授，博士生导师。

(本文编辑：傅　杰)

(英文编辑：赵尹默)