第 43 卷 第 15 期/2023 年 8 月/光学学报 特邀综述

光学学报

基于点云数据的三维目标检测技术研究进展

李佳男[1,2]，王泽[1]，许廷发[1,2,3*]

[1]北京理工大学光电学院，北京 100081；

[2]北京理工大学光电成像技术与系统教育部重点实验室，北京 100081；

[3]北京理工大学重庆创新中心，重庆 401135

摘要 近年来，随着深度传感器和三维激光扫描设备的普及，点云数据引起了广泛关注。相对于二维图像，点云数据不仅包含场景的深度信息，还不受光照等环境因素的影响，能够更精确地实现目标识别和三维定位。因此，基于点云的三维目标检测技术已经成为智能空间感知和场景理解的关键技术。本文首先介绍了点云数据的特点，并探讨了不同类型的点云特征提取方法；其次，详细阐述了基于体素、点、图以及体素与点混合的点云目标检测方法的原理和发展历程；然后，介绍了常见的室内外点云目标检测数据集和评价指标，并对各类点云目标检测方法在KITTI和Waymo数据集上的性能进行了详细的比较和分析；最后，对点云目标检测技术的研究进展进行了总结和展望。

关键词 点云；三维目标检测；单模态；多模态

中图分类号 TP121 **文献标志码** A **DOI**：10.3788/AOS230745

1 引 言

目标检测是计算机视觉领域中一项关键任务，其主要目标是在输入的场景中准确地定位和分类目标对象，生成特定位置、大小和方向的边界框来框选目标，并对其进行分类。近年来，由于无人驾驶、机器视觉、增强现实等领域的快速发展，如何开展准确、实时的三维目标检测受到研究人员的广泛关注。

典型的三维目标检测技术使用图像和点云数据作为输入。由于图像数据缺乏深度信息，其在准确定位三维目标方面存在一定的限制。此外，基于图像的三维目标检测模型容易受到光照等外部环境因素的影响，从而降低了其可靠性。随着深度传感器和三维激光扫描设备的广泛应用，点云数据的采集成本降低，获取速度加快，数据精度提高。这些因素推动了基于点云的三维目标检测技术的快速发展。

点云是由点构成的数据集合，具有丰富的几何和深度信息。与二维图像不同，点云的结构不规则，数据排列无序，因此点云特征提取面临一些挑战。传统的点云特征提取方法通常基于局部点云的曲率、法向量、密度等信息，并结合高斯模型[1]、支持向量机[2]、随机森林[3]等方法，手动设计描述符来提取点云特征。然而，这些方法需要大量的先验知识，并忽略了点与点之间的关联，导致所构建模型的鲁棒性较差，容易受到点云噪声的干扰。虽然结合马尔可夫随机场[4]、条件随机场[5]等方法可以增强邻域间的关联性，但是由于过于依赖人工设计的规则，这些模型的泛化能力受限，无法适应更加复杂的场景。

2015年，LeCun等[6]对深度学习方法进行了系统阐述。深度学习方法因具有出色的特征表达能力和泛化能力而引起研究人员的广泛关注，它是一种数据驱动方法，其研究建立在大规模数据集的基础上。2012年，德国卡尔斯鲁厄理工学院和丰田美国技术研究院共同建立了KITTI数据集[7]，该数据集为深度学习方法的研究提供了重要的基础。随着对无人驾驶领域研究的深入，涌现出许多大规模的点云数据集，这进一步推动了基于深度学习的点云三维目标检测方法的发展。

由于基于深度学习的二维图像处理方法迅速发展，早期的研究提出了基于体素的方法[8-9]。这些方法将不规则的点云数据转换为规则的体素栅格，以便与二维方法结合进行目标检测。体素化的方法有效地解决了点云结构的不规则性问题，但同时也引入了量化误差，从而导致点云信息损失。为了缓解信息损失问题，可以提高量化精度，但这也会显著增加计算成本并提高网络训练的难度。

2017年，Qi等[10]提出了PointNet模型，该模型直接对点云数据进行处理，并充分利用点云所蕴含的丰

收稿日期：2023-03-29；**修回日期**：2023-05-31；**录用日期**：2023-06-05；**网络首发日期**：2023-06-28

基金项目：国家自然科学基金青年科学基金(62101032)

通信作者：*ciom_xtf1@bit.edu.cn

富几何信息，实现了高质量的点云特征提取。随后，基于PointNet的研究[11]验证了基于点的方法在目标检测中的可行性。近年来，将点云数据转换为图论中的图结构数据，并利用图神经网络间接处理点云数据，这为点云目标检测方法的发展注入了新的活力[12-15]。

为了帮助研究人员全面了解点云目标检测方法的发展脉络并快速聚焦领域热点，本文对基于点云的三维目标检测方法进行了整理和分析。首先，详细介绍了点云数据的特性，并全面阐述了主流的特征提取方法，包括基于体素、基于点和基于图的特征提取方法。其次，根据特征提取方法，将基于点云的单模态方法划分为四类：基于体素、基于点、基于图以及基于体素与点混合的检测方法。针对每类方法，重点介绍了其基本架构和发展历程，并对多模态方法进行了补充讨论。然后，详细介绍了典型的室内外点云数据集和评价指标，并通过在KITTI和Waymo数据集上的性能对比和分析，评估了不同方法的表现。最后，对点云目标检测方法未来发展的方向进行了展望。本文的梳理结果有助于研究人员系统地了解和掌握点云目标检测方法的发展动态，为相关研究提供参考和借鉴。

2 点云数据

点云，即一系列点的数据集合，可表征三维场景的空间结构信息，其表达式为

$$\boldsymbol{P}=\{\boldsymbol{M},\boldsymbol{A}\},\tag{1}$$

$$\boldsymbol{M}=\{\boldsymbol{m}_i=(x_i,y_i,z_i)|i=1,\cdots,N\},\tag{2}$$

$$\boldsymbol{A}=\{\boldsymbol{a}_i|i=1,\cdots,N\},\tag{3}$$

式中：$\boldsymbol{P}$为包含N个点的点云集合；$\boldsymbol{M}$为点云的坐标集合，$\boldsymbol{m}_i=(x_i,y_i,z_i)$表示第$i$个点的三维坐标；$\boldsymbol{A}$为点云的附加特征集合，$\boldsymbol{a}_i$为第$i$个点的附加特征信息，包括颜色信息、法向量信息、光谱强度信息等。

相较于二维图像，点云具有许多优势：1）点云不仅保留了场景的几何信息，还提供了丰富的深度信息；2）点云对光照条件的敏感性较小，因此在复杂环境中具有更好的适应性。点云所具有的独特性质，包括稀疏性、不规则性和无序性，使得高效合理地提取点云的信息成为点云目标检测的基础与关键。

2.1 点云数据特性

2.1.1 稀疏性

获取真实点云数据的主要设备是激光雷达，它通过发射和接收激光来感知物体的距离和方向，从而生成点云数据。受到激光雷达的发射频率和角分辨率等参数的限制，采样的点云密度通常非常稀疏。以经典数据集KITTI为例，将原始的激光雷达点云投影到对应的彩色图像上，只有约3%的像素点有点云信息。

此外，不同物体的表面属性不同，因此获取的点云在密度、强度等基本属性上存在明显差异。在不同相对位置或不同采样方向下获取的同一物体的点云也会有差异。点云在三维场景中的稀疏程度不一，呈现出近处密集、远处稀疏的特点，不同区域的点云存在过采样或欠采样的情况。

2.1.2 不规则性

在图像中，每个像素与其邻域像素之间的距离和方向是相同的，并且整个图像由规则排列的像素构成，因此图像的任何局部区域具有相同的结构和规则性，如图1(a)所示。在点云中，不同点之间的距离和方向是不同的，因此任何局部区域的结构都会存在较大差异，表现出不规则性，如图1(b)所示。

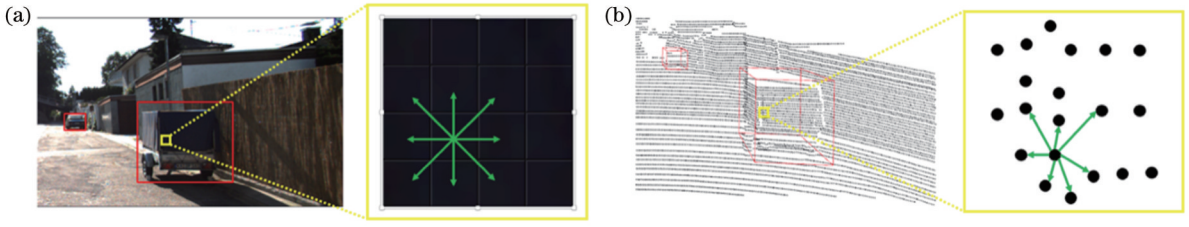

图1 图像与点云的结构差异。(a)规则结构；(b)不规则结构

Fig. 1 Structural differences between image and point cloud. (a) Regular structure; (b) irregular structure

2.1.3 无序性

图像的像素排列具有明确的顺序，对应一个唯一的矩阵。相比之下，点云具有无序性，其排列顺序受设备采集方式和数据读入方式等因素的影响。对点数为N的点云进行矩阵表示时，可得到$N!$个不同排列顺序的矩阵。如图2所示，对同一个点云$\{p_1,p_2,p_3,p_4,p_5\}$进行矩阵表示时，不同的读取顺序1→2→3→4→5(12345)、2→4→5→3→1(24531)将产生两个不同的矩阵$[p_1,p_2,p_3,p_4,p_5]$、$[p_2,p_4,p_5,p_3,p_1]$，这给后续的点云特征提取带来了挑战。

2.2 点云特征提取方法

点云的无规则性和无序性特点给点云数据处理带来了挑战，因此在点云特征提取方面需要克服这些困难。目前，点云特征提取方法主要分为三类：基于体素的方法、基于点的方法和基于图的方法。

2.2.1 基于体素的特征提取方法

基于体素的特征提取方法将点云区域划分为规则的三维体素栅格，并以体素为单位提取点云特征，生成结构规则的四维体素特征图。对于一个点云$\boldsymbol{P}$，它在直角坐标系各个坐标轴上的范围分别为$[D,H,W]$，

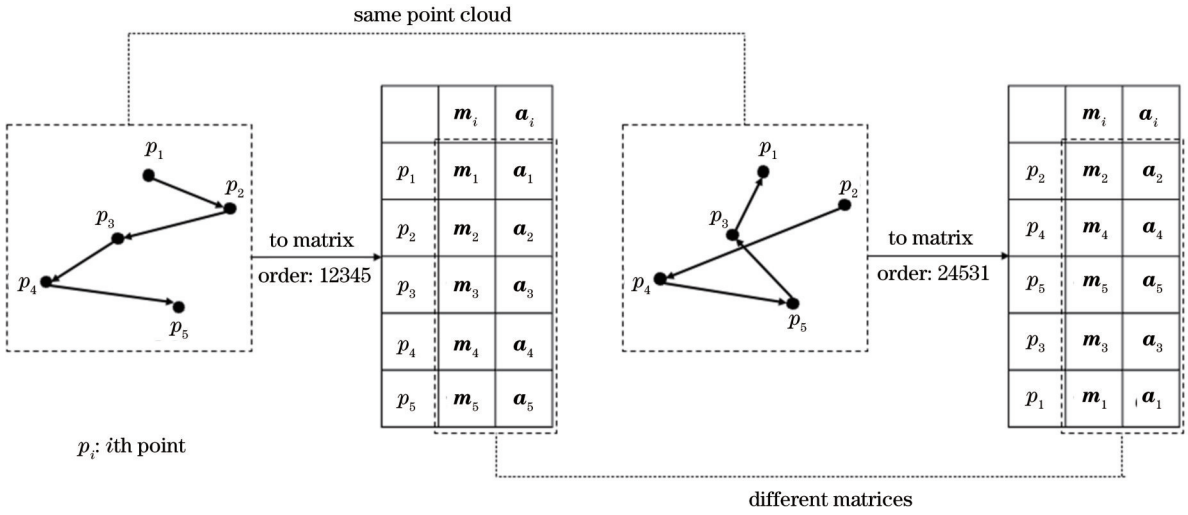

图 2 点云数据的无序性

Fig. 2 Disorder of point cloud data

如果设定体素大小为$[v_D, v_H, v_W]$，则体素化后的栅格分辨率为$[D', H', W']$，其中，$D'=D/v_D$，$H'=H/v_H$，$W'=W/v_W$。如图 3(a)所示，对于每一个体素，利用体素特征聚合模块聚合内部点云的特征$F\in R^{N\times C}$（R表示维度），可生成大小相同的体素特征向量$f\in R^{1\times C'}$，即

$$f=g_{\mathrm{VFA}}(F), \tag{4}$$

式中：g_{VFA}表示体素特征聚合函数，如平均池化函数。每个体素的输入特征维度取决于其包含的局部点数量，但经过体素特征聚合模块处理后，每个体素都对应一个具有相同维度的特征向量，这就将非结构化的点云P转换为规则的四维体素特征图$V\in R^{D'\times H'\times W'\times C'}$。

2.2.2 基于点的特征提取方法

基于点的方法直接处理点云，无需体素化等额外操作，最大限度地保留了点云的几何信息。对于一个点云P，采样一系列关键点并利用这些关键点对点云进行分组。如图 3(b)所示，对于包含N个点的组，局部点云的特征为$F\in R^{N\times C}$，先利用权重共享的特征变换函数进行特征变换，将特征的维度转换为$F'\in R^{N\times C'}$，再利用对称函数逐通道地聚合N个点的特征，生成分组特征向量$f\in R^{1\times C'}$，从而实现点云的特征提取，即

$$f=S\big(H(F)\big), \tag{5}$$

式中：H为特征变换函数，可以用多层感知器（MLP）学习实现；S为对称函数，如最大池化函数、平均池化函数。

2.2.3 基于图的特征提取方法

点云数据和图论中的图数据存在一定的相关性，可以将点云看作图中的节点，并通过连接不同节点来构建图，从而利用基于图的方法进行点云特征提取。基于图的方法包括图构建、图迭代和图聚合 3 个步骤，如图 3(c)所示。对于一个局部点云，首先需要构建一个点云图

$$G=(\boldsymbol{\Phi}, E), \tag{6}$$

式中：$\boldsymbol{\Phi}$为节点特征集合，$\boldsymbol{\varphi}_i\in\boldsymbol{\Phi}$表示第$i$个节点的特征；$E$为连接特征集合，$\boldsymbol{e}_{ij}\in E$表示第$i$个节点与第$j$个节点的连接特征。然后，通过迭代优化来逐步更新图节点的特征，在进行第t次图迭代操作时，节点特征更新为

$$\boldsymbol{\varphi}_i^{t+1}=g^t\left(\rho\left(\left\{\boldsymbol{e}_{ij}^t\middle|\left(\boldsymbol{e}_{ij}^t\in E\right)\right\}\right), \boldsymbol{\varphi}_i^t\right), \tag{7}$$

式中：ρ为连接特征聚合函数；g^t表示利用聚合的连接特征与节点特征实现特征更新。随着迭代次数的增加，节点的感受野不断扩大，T次迭代后，每个节点可以学习到丰富的特征。最后，利用图聚合模块完成节点特征的聚合，从而获得点云的特征表示。

3 点云目标检测方法

根据点云特征提取方式的差异，目前的点云目标检测方法可以分为四类：基于体素、基于点、基于图和基于体素与点混合。其中，基于体素、基于点和基于图的检测方法是三维目标检测的基础分支，基于体素与点混合的方法将两种不同的特征提取方式结合起来，实现了优势互补。在这些单模态方法的基础上，多模态融合方法通过引入其他传感器的信息来补充点云信息，突破了单一模态的信息限制，理论上可以达到更高的精度。

图 4 按照时间顺序列出了现有单模态方法的典型研究，并增加了一些多模态融合方法。基于体素的方法一直贯穿着三维目标检测的发展历程，并且一直占据着主导地位；基于点、基于图和基于体素与点混合的方法也不断推出新的研究成果，为三维目标检测领域的发展作出了重要贡献。

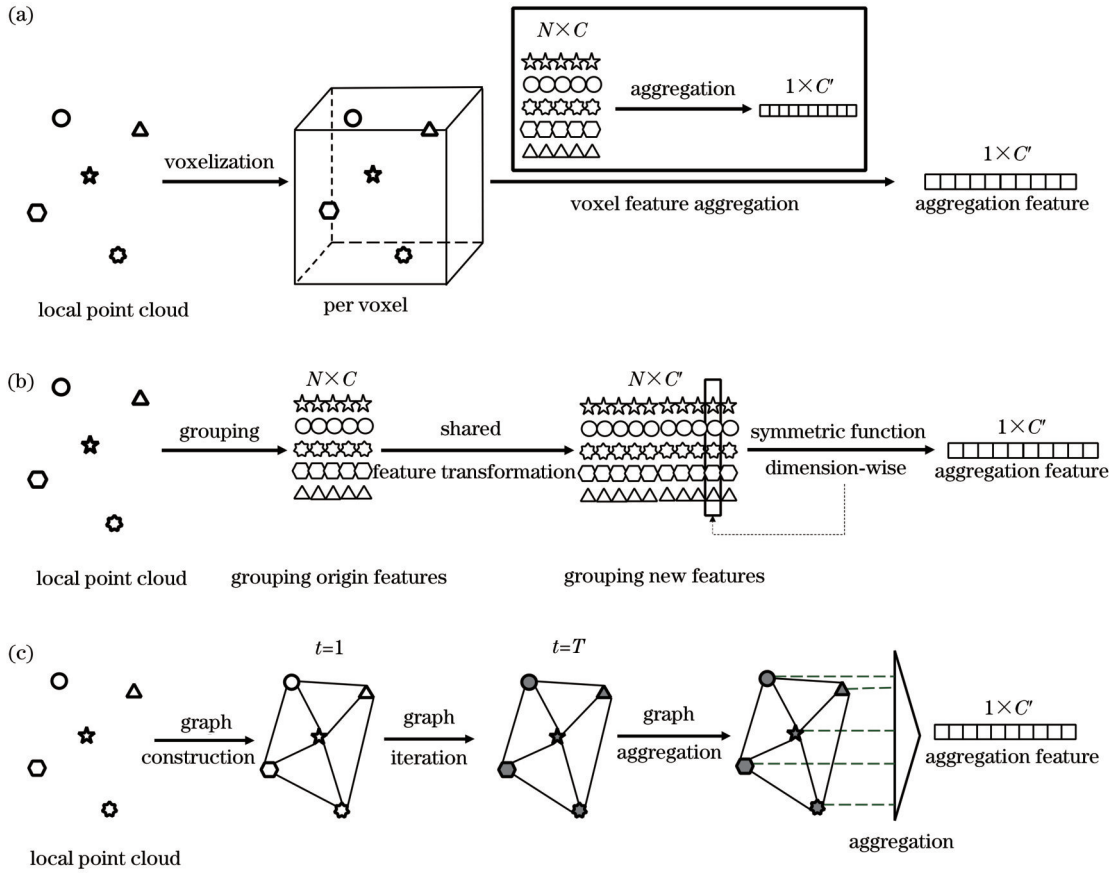

图 3 点云特征提取方法对比。(a)基于体素的点云特征提取方法;(b)基于点的点云特征提取方法;(c)基于图的点云特征提取方法

Fig. 3 Comparision of point cloud feature extraction methods. (a) Voxel-based point cloud feature extraction method; (b) point-based point cloud feature extraction method; (c) graph-based point cloud feature extraction method

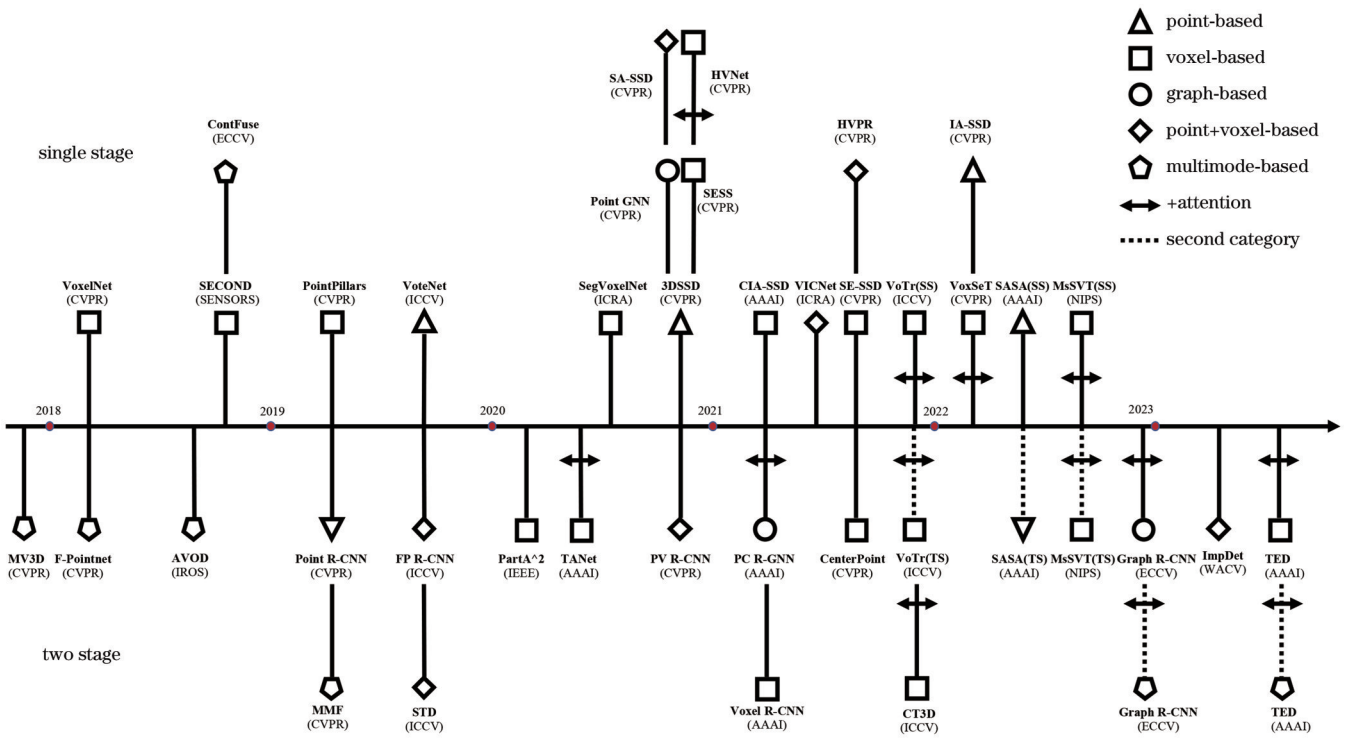

图 4 三维点云目标检测方法发展时间线

Fig. 4 Milestone timeline of 3D object detection in point clouds

3.1 基于体素的检测方法

基于体素的三维目标检测方法通过将点云转换为规则的四维体素特征图，并利用卷积神经网络实现目标检测。以单阶段检测方法为例，如图 5 所示：首先，特征图经过三维卷积模块的处理，实现了特征升维，并且在此过程中空间分辨率下降；其次，经过多层卷积操作，在高度方向实现了压缩，从而生成鸟瞰视角（BEV）的二维特征图；最后，二维特征图被送入二维特征提取网络，进行边界框预测和目标分类。基于体素的检测方法通过规则化点云数据，并利用卷积神经网络进行后续处理，具有高效的计算能力，因此成为三维目标检测的主流方法。

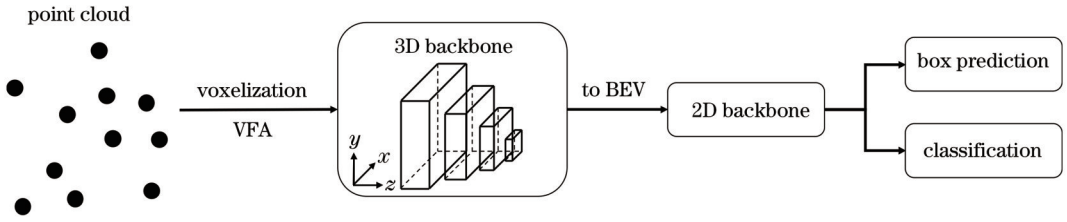

图 5 基于体素的目标检测流程

Fig. 5 Pipeline of voxel-based object detection method

早期的基于体素进行目标检测的尝试包括 Maturana 等[8]提出的 VoxNet 模型。该模型由两个部分组成：体素栅格模块和三维卷积模块。体素栅格模块用于将点云特征转换为体素特征，三维卷积模块则利用卷积操作对体素特征进行高维表示。然而，VoxNet 模型所使用的体素特征是手工设计的，因此在面对更为复杂的场景时无法适应。

Zhou 等[16]在 VoxNet 模型的基础上提出一种端到端的 VoxelNet 模型。该模型引入了体素特征编码模块，将逐点特征与局部聚合特征相结合，增强了体素内部点之间的交互作用。通过堆叠多个体素特征编码模块，VoxelNet 可以学习到更加丰富的特征，进而更好地表征三维目标的形状信息。然而，VoxelNet 在计算成本方面存在较高的开销，导致其训练过程缓慢且难以收敛。

Yan 等[17]提出的 SECOND 模型通过采用稀疏卷积的方法改进了 VoxelNet 模型的三维卷积模块，从而极大地提高了体素模型的训练和测试速度。稀疏卷积模块也成为后续体素模型中常用的组件之一。Lang 等[18]提出的 PointPillars 模型则移除了三维卷积模块，并在体素化过程中去除了 z 轴切分，仅在 x 和 y 方向上形成体柱排列。PointPillars 模型采用 PointNet 方法来聚合体柱内的局部点云特征，形成体柱特征，并将其映射为伪图像的特征，从而可以直接使用二维目标检测的特征网络进行处理。SECOND 和 PointPillars 被认为是体素方法中最经典的架构，为基于体素的三维目标检测的后续发展奠定了基础。

基于 SECOND 模型，Deng 等[19]提出了 Voxel R-CNN 模型。在该模型中，当生成候选框时，他们引入一系列格点，并以这些格点为中心，通过体素查询模块和改进的点云聚合模块提取了三维卷积模块中生成的体素特征。这样做可以感知体素栅格的空间结构信息，有效解决了体素化过程中信息损失的问题。

在三维场景中，目标的方向变化多样。为了提高模型对目标变换（如旋转、对称）的鲁棒性，传统的检测器通常采用数据增强或测试时间增强的方法。然而，这些方法的效果有限，且会增加大量的时间成本。为了解决这一问题，Wu 等[20]在 Voxel R-CNN 模型的基础上引入了多通道变换体素特征，该方法成功提高了模型对目标方向的预测精度，并在 KITTI 数据集上取得了出色的表现。

1）基于体素的检测方法结合二维经典检测架构

在三维目标检测领域，经典的架构相对较少。因此，研究人员开始尝试将二维目标检测模型扩展到三维领域，以实现对三维目标的检测。

Zheng 等[21]提出的 CIA-SSD 模型中包含置信度校准模块，用于感知交并比，从而提高单阶段目标检测的精度和速度。随后，Zheng 等[22]在 CIA-SSD 模型的基础上，结合知识蒸馏原理，提出一种名为 SE-SSD 的单阶段目标检测模型。SE-SSD 包括教师模型和学生模型，二者都以 CIA-SSD 模型为基础。教师模型用于对原始点云目标进行预测，生成信息丰富的软目标；学生模型综合学习软目标和对原始点云进行数据增强后生成的硬目标，这样可以显著提高检测精度。通过引入知识蒸馏的方法，SE-SSD 模型在保持高效率的同时，实现了更优的检测性能。

Yin 等[23]对二维目标检测领域的经典模型 CenterNet 进行了改进。他们将目标定位任务从生成目标候选框转变为生成目标中心点，然后利用中心点特征回归候选框的属性、类别等信息。通过结合中心点特征和周围点特征，在两个阶段完成置信度的预测和边界框的微调。这种方法简化了锚框生成的步骤，并在提高算法运行速度的同时增强了对目标方向预测的能力。

2）基于体素的检测方法结合注意力机制

2017 年，Google 提出了 Transformer 模型[24]，这个模型利用自注意力机制，有效地提取了全局信息。自那时以来，Transformer 模型在计算机视觉领域得到了

广泛的应用[25-36]。

Mao等[25]通过子流形体素模块和稀疏体素模块来解决非空体素稀疏的问题。此外，他们还结合了局部注意力和分散注意力两种注意力机制，并配合使用快速体素查询模块，有效地解决了非空体素数量庞大的问题。这种方法成功地利用Transformer替代传统的三维卷积模块。Sheng等[26]根据第一阶段检测生成的候选框，利用Transformer提取候选局部区域内点的特征，以生成更精确的边界框。He等[27]提出一种交叉注意力机制，用于对体素内的点云进行集合到集合的学习。通过引入交叉注意力机制，他们能够更好地感知体素内部的点云结构。

针对基于窗口的Transformer方法在获取点云长程依赖时可能导致细节信息模糊、目标定位和分类精度降低的问题，Dong等[28]提出了MsSVT模型。该模型通过将注意力头分组，每组负责感知特定范围的信息，并最终聚合各组的输出形成混合尺度特征。这种分组注意力机制可以在保持感受野范围的同时，更好地捕捉不同尺度的细节信息，提高目标定位和分类的精度。MsSVT模型不仅具有高效运行的优势，而且在单阶段模型中的性能也超越了许多双阶段模型，这在Waymo数据集上得到了验证。

3）小结

基于体素的检测方法在三维目标检测中具有一些显著的优点，如计算高效性和可迁移性，尤其是借鉴了二维目标检测方法的思想。此外，通过感知三维体素的空间结构等方法，可以在一定程度上缓解体素化过程中的信息损失问题。然而，基于体素的检测方法仍然存在一些限制和挑战。首先，体素化过程会引入量化损失，从而导致信息的部分丢失和模糊；其次，基于体素的方法在处理大规模点云数据时可能会面临较高的计算成本和较大的存储需求，导致训练和推断的效率较低。

3.2 基于点的检测方法

基于点的检测方法采用PointNet系列网络[10,37]作为骨干网络，直接对点云场景进行目标检测。如图6所示，该方法以单阶段检测方法为例，通过堆叠多层的采样、分组、聚合模块来扩大模型的感受野，并将特征聚合到关键点上。随后，利用关键点特征进行边界框预测和目标分类。基于点的检测方法具有简单的架构，并且能够保留更丰富的点云信息，因此可以达到较高的检测精度。

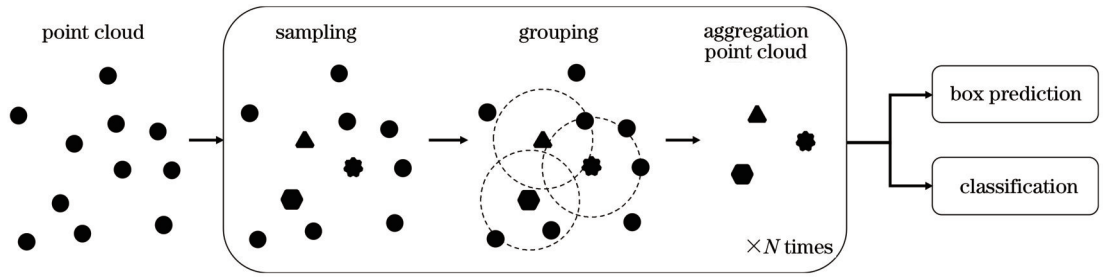

图6 基于点的目标检测流程

Fig. 6 Pipeline of point-based object detection method

Shi等[38]提出了Point R-CNN模型，该模型在点云目标检测中引入了语义分割的步骤。首先，对点云进行语义分割，将点云分为前景点和背景点；然后，生成高质量的候选框，并在第二阶段将候选框内的点云映射到规范坐标系中。在这个过程中，Point R-CNN模型利用点的语义标签来进一步提高目标检测的精度。

在PointNet++作为骨干网络的基础上，Qi等[11]提出了VoteNet模型。该模型通过以下步骤进行目标检测：首先，生成具有较大感受野的种子点特征；然后，利用霍夫投票的原理，回归种子点对应的目标中心点位置；最后，通过聚合中心点附近的种子点特征，实现目标边界框的预测和分类。

Yang等[39]提出了3DSSD模型，在最远点采样的基础上进行改进。该模型引入了语义距离来替代传统的欧氏距离，有效地保留了前景点。同时，3DSSD模型改进了边界框预测网络，并移除了特征传播模块和微调模块，以实现精度和运行速度的平衡。

Zhang等[40]在VoteNet模型的基础上进行改进，提出了IA-SSD模型。该模型设计了两种可学习的采样策略，即类别感知采样和中心感知采样，以更好地保存前景点。IA-SSD模型通过改进投票机制生成更高质量的种子点特征，并将其用于目标中心点预测、边界框预测和目标分类。IA-SSD模型在精度和计算效率方面表现出色，超越了PointPillars模型，它为基于点的检测方法注入了新的活力。

Chen等[41]提出了语义增强的集合抽象（SASA）模块，该模块在点的语义信息和坐标信息的基础上对传统的最远点采样方法进行改进。SASA模块为距离指标附加语义权重，使采样点在保留更多前景点的同时分布在更大的范围内。这个改进解决了保留简单目标的前景点而忽略难以检测目标的前景点的问题。研究人员在3DSSD模型和Point R-CNN模型上验证了SASA模块的有效性，实验结果表明，SASA模块不仅能够提高检测精度，而且能够与其他基于点的模型结

合使用，进一步增强检测性能。

基于点的检测方法具有结构简单、充分保留点云的几何信息以及容易实现较高精度的优点。此外，通过设计新的采样策略和聚合方式等方法，可以实现计算效率的优化。然而，由于重复执行点云采样和聚合的操作，计算开销大的问题尚未完全得到解决。此外，虽然基于点的检测方法充分保留了点云的几何信息，但如何实现高质量的点云特征提取仍需进一步研究。

3.3 基于图的检测方法

基于图的方法将点云数据表示为图结构，并借助图神经网络来间接处理点云数据，从而实现目标检测任务。以单阶段检测方法为例，其具体实现过程如图 7 所示：首先，构建图结构，将不同的节点连接起来；然后，通过利用连接边和邻域点的特征来迭代更新中心点的特征，以增强中心点与周围区域的联系。这样的迭代操作可以重复执行，从而生成一系列高质量的中心点特征，用于执行目标边界框的预测和分类任务。

Shi 等[42]提出了 Point GNN 模型，将图神经网络首次应用到点云目标检测任务中。该模型的核心思想是将点云表示转化为图表示，通过连接边将邻域内的节点特征传递给中心节点，并将这些特征与中心节点特征进行结合，实现中心节点特征的迭代更新。最终，利用最新的节点特征进行目标边界框和目标类别的预测，从而完成单阶段目标检测任务。

Point GNN 模型中所有节点对局部特征和全局特征的贡献被视为平等，并且该模型仅使用单一尺度的图形表示，未能充分利用上下文信息。为了解决这个问题，Zhang 等[43]提出了 PC R-GNN 模型。该模型是一个二阶段检测器：在第一阶段中，基于 Point R-GNN 模型，利用点的语义信息来筛选前景点并生成数量较少但质量较高的候选框；在第二阶段中，先使用点云补全模块对候选框内的点云进行补全，再利用基于注意力机制的多尺度图神经网络来充分提取几何信息。这样的设计使得 PC R-GNN 模型能够准确地检测稀疏或部分残缺的点云目标，提高检测的精度和鲁棒性。

Yang 等[44]提出了通用的二阶段修正模块 Graph R-CNN。该模块的第一步是执行动态点云聚合，动态地提取候选框内不同距离下的均匀采样点；第二步，通过应用图神经网络对提取的采样点进行上下文感知，生成高质量的聚合特征。Graph R-CNN 模块被设计为一个即插即用的二阶段修正模块，可以提高单阶段模型的精度。

基于图的检测方法结合了图神经网络的优势，成为一种新兴且表现优异的目标检测方法。尤其在二维领域，图神经网络方法不断发展，取得了显著的成果。然而，如何将这些图神经网络模型进一步迁移到三维点云数据上，仍然是需要进一步研究的方向。

3.4 基于体素与点混合的检测方法

基于体素的方法将点云转化为规则的体素，利用卷积神经网络高效提取特征。这种方法的优势在于可以利用卷积操作的并行性和高效性，但在体素化过程中会引入一定的量化误差。相比之下，基于点的方法保留了点云的原始结构，能够提取更丰富的点云信息。然而，重复执行点云采样与聚合的操作会增加计算成本。为了克服单模态方法的局限性，一些研究人员尝试结合基于体素的方法和基于点的方法，设计出兼具二者优势的混合模型。

Chen 等[45]提出了 Fast Point R-CNN 二阶段目标检测模型，该模型将点云表示和体素表示相结合。在第一阶段，该模型利用点云的体素表示作为输入，通过卷积神经网络生成候选框，并保留了卷积层的特征；在第二阶段，模型结合候选框内的点的坐标特征和对应的卷积特征，利用 PointNet 进行特征聚合，从而预测目标的类别和边界框。Fast Point R-CNN 模型成功地将点云表示和体素表示相结合，是早期基于体素和点混合的目标检测方法之一。

Yang 等[46]提出了 STD 模型，该模型利用 PointNet++生成每个点的语义特征作为补充特征。在第一阶段，该模型在每个点的位置生成球形锚框，并利用 PointNet 对锚框内的点云特征进行聚合，从而生成候选框；在第二阶段，该模型对候选框中的点云进行体素化操作，将稀疏的特征表示转化为稠密的特征表示，以提升检测效果。STD 模型在召回率方面表现出较优的性能，并在定位困难目标、小目标等方面具有优势。

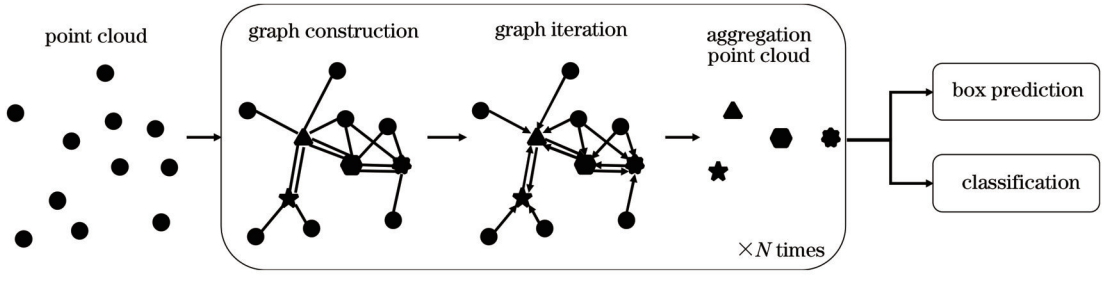

图 7 基于图的目标检测流程

Fig. 7 Pipeline of graph-based object detection method

Shi等[47]提出了PV R-CNN模型。该模型首先利用基于体素的方法，采用三维卷积神经网络作为骨干网络，高效地生成多尺度的体素特征和候选框；然后，利用基于点的方法，设计了体素聚合模块，将体素特征和原始点云特征聚合到关键点上进行采样；最后，通过格点池化模块，将关键点特征传递给候选框内的格点，实现目标边界框的预测和分类。PV R-CNN模型在性能上表现出色，在KITTI数据集中的Car、Pedestrian和Cyclist 3个类别及其不同难度级别的目标检测工作中都获得优异的检测结果。

He等[48]提出一种以基于体素的方法为主导的模型，其在训练过程中将卷积神经网络提取的特征映射回原始点云。该模型通过引入辅助模块进行前景点分割和目标中心点预测，以监督卷积神经网络感知原始点云的结构信息，从而解决了体素系列方法中信息损失的问题。在预测过程中，不再使用辅助模块，加快了模型的推理速度。这种网络设计方式为加速模型推理过程提供了新的思路。

综上所述，混合模型的设计可以充分发挥点云表示与体素表示的优势，从而提高点云目标检测的精度。随着基于体素和基于点的检测方法的不断发展，混合模型的潜力有待进一步挖掘和探索。

3.5 多模态融合的检测方法

不同类型的传感器所获取的数据具有各自的特点。例如：图像数据富含目标的边缘、颜色、亮度等信息，但缺乏深度信息，无法准确定位三维目标；点云数据富含目标的几何信息，但缺乏语义信息，难以精确区分邻近目标。因此，通过融合图像和点云数据，可以同时获取深度和语义信息，从而实现更精确的目标检测。基于图像和点云的多模态融合检测方法应运而生。根据融合方式的不同，可以将其分为两个级别[49]：候选框级别和点-像素级别。

1）候选框级别

候选框级别的多模态检测方法用于生成目标的候选框，并将候选框映射到不同的模态，从而提取对应的信息，并实现信息融合。在F-PointNet模型[50]和F-ConvNet模型[51]中，首先利用RGB图像生成二维候选框，并将这些候选框映射到三维点云区域，生成相应的锥形候选框；然后，使用PointNet提取候选框内的点云特征，用于后续的目标边界框预测与分类。虽然F-PointNet模型和F-ConvNet模型充分利用图像和点云的信息，但它们在不同的阶段发挥作用，没有真正实现信息的融合和互补。MV3D模型[52]在第一阶段将点云投影到BEV视图，并利用二维区域生成网络产生候选框；在第二阶段中，MV3D模型融合了BEV视图、前视图和RGB图像的相应特征，用于边界框回归。AVOD模型[53]在MV3D的基础上进一步拓展，在候选框生成过程中不再依赖单一的BEV视图，而是融合了RGB图像信息，以获得更高质量的候选框。FUTR3D模型[54]和TransFusion模型[55]采用Transformer架构，利用目标查询模块在点云空间生成高质量的候选框，并融合了候选框对应的图像特征，实现了更优质的信息融合。

2）点-像素级别

点-像素级别的多模态检测方法旨在为点云附加图像的语义特征，以结合几何信息和语义信息进行目标检测。PointPaint模型[56]将点云投影到预训练的图像语义分割网络的预测图上，并为每个点赋予语义类别分数。MVP模型[57]基于二维检测结果生成稠密的三维虚拟点云，以增强原始稀疏点云数据。PointPaint模型和MVP模型都在输入层增加了语义信息，但在特征层面实现融合也是一个值得注意的方向。ContFuse模型[58]使用连续卷积在多个尺度上融合来自不同传感器的卷积特征。MMF模型[59]引入地面估计和深度补全辅助任务，以提高网络的特征提取能力并加强特征融合。

综上所述，多模态检测方法通过结合不同模态的数据，能够获得更加丰富和全面的信息。随着融合技术的不断发展，多模态方法的检测精度有望超越单模态方法，展现出良好的发展前景。

4 数据集与评价指标

4.1 点云目标检测数据集

公开的大规模三维目标检测数据集对于推动点云目标检测技术的发展起到了巨大的推动作用。这些数据集根据场景类型的不同可以分为室内场景数据集和室外场景数据集[60]，具体信息如表1所示。

4.1.1 室外数据集

常用的室外数据集包括KITTI、Waymo、nuScenes和STCrowd。图8展示了不同室外数据集的可视化样例，包括图像、点云以及待检目标（方框标示）。下面对各室外数据集的场景特点和规模等特性进行介绍。

KITTI数据集[7]由德国卡尔斯鲁厄理工学院和丰田美国技术研究院联合制作而成。该数据集收集了真实世界中城市和乡村等不同场景的图像数据。每幅图像最多包含15辆车和30个行人，并且存在各种程度的遮挡和截断情况。整个数据集包括389对立体图像和超过20万个带有三维边框标注的目标，涵盖了轿车、货车、卡车、行人、自行车和有轨电车6个类别。此外，数据集还提供了完全可见、半遮挡、全遮挡和截断4种标注信息，以及目标数量、朝向、大小等属性信息。

Waymo数据集[61]是由Waymo公司和Google公司联合制作的一个多模态、场景多样化的大规模数据集。该数据集包含激光雷达点云和图像数据，并采集于市区和郊区等不同场景，覆盖了白天、黑夜的不同时段。该数据集中共包含1150个场景，每个场景的时长为20 s；标注了约1200万个目标的三维边界框，涵盖了机

表 1 常用的三维目标检测数据集对比

Table 1 Comparison of commonly used 3D object detection datasets

Dataset	Scene	Year	Data type	3D bounding box	Category
KITTI[7]	Outdoor	2012	Point cloud +image	2×10^5	8
Waymo[61]	Outdoor	2019	Point cloud +image	1.2×10^6	4
nuScenes[62]	Outdoor	2019	Point cloud +image	4×10^5	23
STCrowd[63]	Outdoor	2022	Point cloud +image	2.19×10^5	1
NYU-Depth[64]	Indoor	2012	Image+depth map	3.5×10^4	40
SUN3D[65]	Indoor	2013	Image+depth map	—	—
SUN RGB-D[66]	Indoor	2015	Image+depth map	5.8×10^4	800
ScanNet[67]	Indoor	2017	Image+depth map	—	—

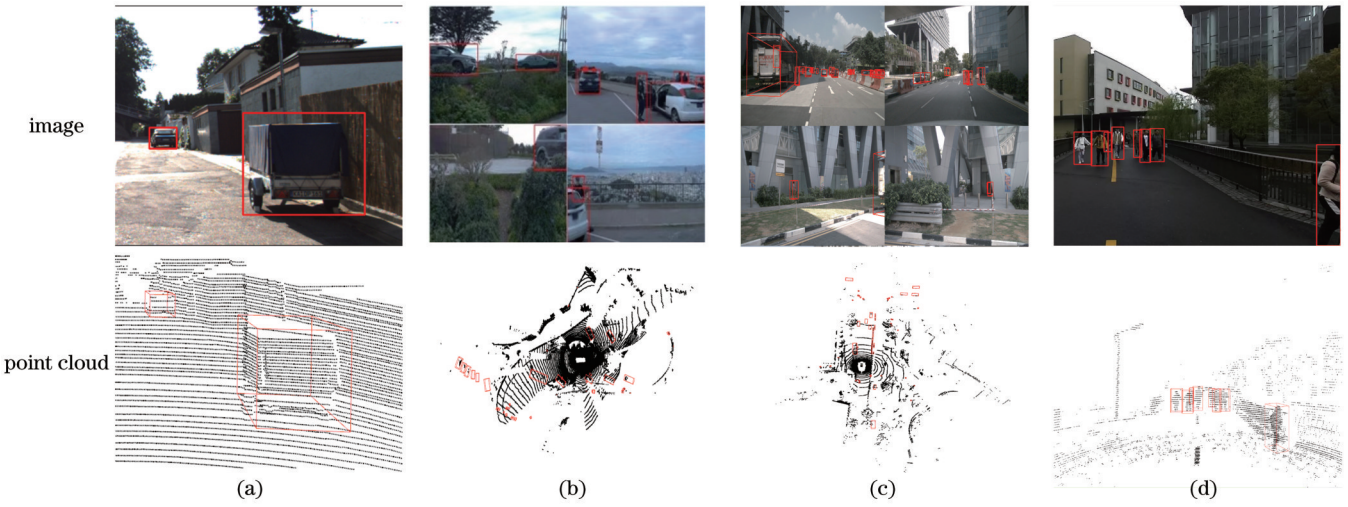

图 8 室外三维目标检测数据集样例。(a) KITTI;(b) Waymo;(c) nuScenes;(d) STCrowd

Fig. 8 Samples of outdoor 3D object detection datasets. (a) KITTI; (b) Waymo; (c) nuScenes; (d) STCrowd

动车辆、行人、自行车和信号灯 4 个类别。

nuScenes 数据集[62]是由 nuTonomy 公司创建的一个大规模、多模态的数据集。该数据集利用各种类型的传感器采集数据,提供 360°的全方位视野,以城市为采集地点,并涵盖白天、黑夜,以及晴天、雨天和多云等不同时间和天气条件下的场景。该数据集共包含 1000 个场景,每个场景的时长为 20 s;标注了约 40 万个三维边界框,涵盖了 23 个目标类别,并提供了可见度、活动度、姿态等附加信息。

STCrowd 数据集[63]是由上海科技大学和香港中文大学等机构联合制作的高密度行人数据集。该数据集具有大规模、多模态和多样化的特点,包含丰富多样的场景、天气条件、行人密度和姿态等内容。整个数据集包含 84 个序列和 10891 帧场景,平均每帧场景包含约 20 个行人目标,并且伴随着不同程度的遮挡情况。该数据集总共包含 21.9 万个三维边界框,每个边界框都附带有密度和遮挡等信息。

4.1.2 室内数据集

常用的室内数据集主要包括 NYU-Depth、SUN3D、SUN RGB-D 和 ScanNet。下面对各室内数据集的场景特点和规模等特性进行介绍。

NYU-Depth 数据集[64]采集于美国 3 个不同城市的商业区与住宅区。该数据集包含了 1449 幅室内 RGB-D 图像,涵盖了 464 个场景。数据集中标注了 35064 个目标,并将它们分为 40 个不同的类别。

SUN3D 数据集[65]是一个大规模的 RGB-D 序列数据集,其中包含相机姿态和目标标签信息。与传统数据集基于特定视角的限制不同,SUN3D 数据集以场景为中心,全面地呈现了场景的三维信息。该数据集包含来自 254 个场景的 415 个序列,这些序列经过机器与人工混合标注处理。

SUN RGB-D 数据集[66]采用 4 个经过同步与校准的深度传感器进行数据采集。该数据集包含 10335 幅 RGB-D 图像,共有 146617 个二维多边形标注框和 64595 个三维边界框,这些边界框还附带了目标的方向信息。此外,每幅图像还提供对应的三维空间布局和场景类别等详细注释信息。

ScanNet 数据集[67]收集了来自 1513 个场景的 250 万幅 RGB-D 图像,并提供了三维相机姿态和表面重构等详细信息。该数据集涵盖了各种场景,包括小范围的浴室、壁橱等,以及大范围的公寓、教室等,这给三维目标检测带来了更大的挑战。

4.2 评价指标

在目标检测任务中，样本通常被分为正样本和负样本。根据检测算法对样本的预测结果，可以将其归类为以下4种情况：真正（TP）例、假负（FN）例、假正（FP）例、真负（TN）例。基于这些情况，可以计算以下3个指标：精确率（p）、召回率（r）、准确率（R_{acc}）。

$$p=\frac{n_{\mathrm{TP}}}{n_{\mathrm{TP}}+n_{\mathrm{FP}}}, \tag{8}$$

$$r=\frac{n_{\mathrm{TP}}}{n_{\mathrm{TP}}+n_{\mathrm{FN}}}, \tag{9}$$

$$R_{\mathrm{acc}}=\frac{n_{\mathrm{TP}}+n_{\mathrm{TN}}}{n_{\mathrm{TP}}+n_{\mathrm{TN}}+n_{\mathrm{FP}}+n_{\mathrm{FN}}}, \tag{10}$$

式中：n_{TP}、n_{FP}、n_{FN}、n_{TN}为不同情况对应的样本数量。

目标检测任务关注预测结果的精确率p与召回率r。为了综合评价算法的性能，研究人员通过计算p-r曲线下方的面积，提出了新的指标——平均精度（p_{AP}）。由于实际p-r曲线下方的面积难以计算，因此研究人员使用离散化的插值指标$p_{\mathrm{AP}\text{-}R_n}$来代替$p_{\mathrm{AP}}$[68]，即

$$p_{\mathrm{AP}\text{-}R_n}=\frac{1}{n}\sum_{r\in R_n}P^*(r), \tag{11}$$

$$P^*(r)=\max_{r^*:r^*\geqslant r}P(r^*), \tag{12}$$

$$R_n=\left\{r_0,r_0+\frac{r_1-r_0}{n-1},r_0+\frac{2(r_1-r_0)}{n-1},\cdots,r_1\right\}, \tag{13}$$

式中：R_n为r的取值集合；n为召回率采样数量；r_0和r_1分别为召回率的最小值和最大值，通常取0和1；$P^*(r)$为召回率r对应的准确率；r^*为大于或等于r的所有真实召回率。

KITTI数据集的三维目标检测任务使用$p_{\mathrm{AP}\text{-}R_{11}}$作为评价指标，对机动车、行人、自行车3类目标分别设置了不同的交并比（IoU）阈值0.7、0.5、0.5。在2019年以后，KITTI数据集使用$p_{\mathrm{AP}\text{-}R_{40}}$代替$p_{\mathrm{AP}\text{-}R_{11}}$来评价三维目标检测算法的性能[69]。

Waymo数据集的三维目标检测任务使用$p_{\mathrm{AP}\text{-}R_{21}}$作为评价指标，并对车辆和行人分别设置了不同的IoU阈值0.7和0.5。为了更好地评估检测算法在目标方向估计方面的性能，Waymo对传统的p_{AP}指标进行改进，引入方向评估信息，并为每个预测正确的正样本（TP）附加了加权因子：$\min\left(|\tilde{\theta}-\theta|,2\pi-|\tilde{\theta}-\theta|\right)\Big/\pi$，其中$\theta$表示预测框方位角，$\tilde{\theta}$表示真实框方位角，范围均在$[-\pi,\pi]$。

5 性能对比与分析

本节以p_{AP}为评价指标，对比了不同类别检测算法在KITTI和Waymo数据集上的性能。具体的性能对比结果见表2和表3。在分析中，重点参考了在KITTI数据集上对车辆的中等难度检测的p_{AP}指标。

早期的基于体素的检测方法如VoxelNet[16]、SECOND[17]、PointPillars[18]等，在体素化过程中会引入量化误差，导致信息损失，因此它们的平均精度指标都低于80%。随后的研究工作加强了对体素空间结构的感知，从而实现了更高的检测精度。例如：Voxel R-CNN[19]通过设置格点感知的体素结构，将精度提升到81.62%；VoTr-TSD[25]和VoxSeT[27]通过引入注意力机制加强了体素之间的联系，分别达到了82.09%和82.06%的精度。此外，基于体素的方法将点云转化为规则的体素特征图，因此可以借鉴二维目标检测领域的成熟技术来提升检测性能。例如，SE-SSD[22]基于蒸馏模型，CenterPoint[23]基于CenterNet，它们的精度都与较新的工作VoxSeT[27]相当。

基于点的检测方法和基于图的检测方法具有较高的精度，它们直接在点云上进行处理，保留了丰富的几何信息，从而能够更准确地完成目标检测任务。例如：基于点的检测方法3DSSD[39]和IA-SSD[40]的精度均超过80%；基于图的检测方法Graph R-CNN[44]的精度接近83%。基于点和基于图的检测方法相比于基于体素的检测方法保留了更多的信息，理论上可以实现更高的精度。然而，目前这三类方法的最新模型的精度相当，都在82%左右，这表明基于点的检测方法和基于图的检测方法的检测精度仍有进一步提升的空间。

基于体素和点的混合检测方法结合了基于体素和基于点的检测方法的优点，以实现精度和速度之间的平衡。然而，目前这种混合方法的精度并没有明显超越单一检测方法，精度仍然在80%左右。此外，多模态方法利用不同模态的数据作为输入，能够提供更加丰富和全面的信息，从理论上讲，可以达到更高的精度。随着融合技术的不断发展，多模态方法的精度有望逐渐接近并超越单模态方法。

6 展 望

基于点云的三维目标检测技术已经取得一系列的成果，但是仍然存在一些可以优化的方面，包括提高模型的鲁棒性、加快检测速度与提高准确性、降低对大量标注数据的依赖、提高在真实场景下的适用性等。

1）多分支融合

基于体素、基于点和基于图的单分支检测方法各有优缺点，因此多分支融合可以结合不同分支方法的优势，理论上有望实现更好的性能。然而，目前基于体素和点云的混合检测方法在精度和速度方面并未明显超越单一分支的方法。为了进一步提高精度和速度，未来的研究可以探索不同的分支组合和优化算法结构。

2）多模态融合

多模态数据融合可以结合更多传感器的信息，从而提高系统的鲁棒性。由于不同传感器的数据形式存在巨大差异，信息的配准和融合成为一个具有挑战性的任务。在早期的研究中，一些方法尝试将点云数据转化

表 2 点云目标检测算法在KITTI数据集上的平均精度(p_{AP})对比

Table 2 Average precision (p_{AP}) comparison of point cloud object detection methods on KITTI dataset unit: %

Type	Method	Car			Pedestrian			Cyclist		
		Easy	Moderate	Hard	Easy	Moderate	Hard	Easy	Moderate	Hard
Voxel-based	VoxelNet[16]	77.47	65.11	57.73	39.48	33.69	31.50	61.22	48.36	44.37
	SECOND[17]	84.65	75.96	68.71	45.31	35.52	33.14	75.83	60.82	53.67
	PointPillars[18]	82.58	74.31	68.99	51.45	41.92	38.89	77.10	58.65	51.92
	PartA^2[70]	87.81	78.49	73.51	53.10	43.35	40.06	**79.17**	**63.52**	**56.93**
	TANet[71]	84.39	75.94	68.82	**53.72**	**44.34**	**40.49**	75.70	59.44	52.53
	SegVoxelNet[72]	86.04	76.13	70.76	—	—	—	—	—	—
	CIA-SSD[21]	89.59	80.28	72.87	—	—	—	—	—	—
	Voxel R-CNN[19]	90.90	81.62	77.06	—	—	—	—	—	—
	SE-SSD[22]	**91.49**	**82.54**	77.15	—	—	—	—	—	—
	VoTr-TSD[25]	89.90	82.09	**79.14**	—	—	—	—	—	—
	CT3D[26]	87.83	81.77	77.16	—	—	—	—	—	—
	VoxSeT[27]	88.53	82.06	77.46	—	—	—	—	—	—
Point-based	Point R-CNN[38]	86.96	75.64	70.70	47.98	39.37	36.01	74.96	58.82	52.53
	3DSSD[39]	88.36	79.57	74.55	**54.64**	**44.27**	**40.23**	**82.48**	64.10	56.90
	IA-SSD (single)[40]	**88.87**	80.32	75.10	47.90	41.03	37.98	82.36	**66.25**	**59.70**
	IA-SSD (multi)[40]	88.34	80.13	75.04	46.51	39.03	35.61	78.35	61.94	55.70
	SASA[41]	88.76	**82.16**	**77.16**	—	—	—	—	—	—
Graph-based	Point-GNN[42]	88.33	79.47	72.29	**51.92**	**43.77**	**40.14**	**78.6**	**63.48**	**57.08**
	PC R-GNN[43]	89.13	79.90	75.54	—	—	—	—	—	—
	GraR-Vol[44]	**91.89**	**83.27**	77.78	—	—	—	—	—	—
	GraR-Po[44]	91.79	83.18	**77.98**	—	—	—	—	—	—
	GraR-Vo[44]	91.29	82.77	77.20	—	—	—	—	—	—
	GraR-Pi[44]	90.94	82.42	77.00	—	—	—	—	—	—
Voxel+point-based	FP R-CNN[45]	85.29	77.40	70.24	—	—	—	—	—	—
	STD[46]	87.95	79.71	75.09	**53.29**	42.47	38.35	**78.69**	61.59	55.30
	PV R-CNN[47]	**90.25**	81.43	76.82	52.17	**43.29**	**40.29**	78.60	**63.71**	**57.65**
	SA-SSD[48]	88.75	79.79	74.16	—	—	—	—	—	—
	ImpDet[73]	88.39	**82.14**	**76.98**	—	—	—	—	—	—
Multimode-based	MV3D[52]	74.97	63.63	54.00	—	—	—	—	—	—
	F-PointNet[50]	82.19	69.79	60.59	**50.53**	**42.15**	**38.08**	**72.27**	**56.12**	**49.01**
	AVOD[53]	76.39	66.47	60.23	36.10	27.86	25.76	57.19	42.08	38.29
	ContFuse[58]	83.68	68.78	61.67	—	—	—	—	—	—
	MMF[59]	**88.40**	**77.43**	**70.22**	—	—	—	—	—	—

为伪图像数据，并在图像层面进行特征融合。然而，这种方法牺牲了点云数据的空间几何信息，导致算法的检测效果不佳。Wang等[79]提出一种将图像数据转化为伪点云数据的方法，在点云层面进行特征融合，以兼顾点云数据和图像数据的优势，并实现特征的深度融合。未来，多模态融合技术的精度有望进一步提升。

3) 迁移二维目标检测方法

二维目标检测技术已经相对成熟，为三维点云目标检测提供了强大的指导作用。因此，可以从二维图像领域的模型设计思路和检测策略中获得灵感，并将其应用到三维点云目标检测任务中，例如结合Transformer的VoTr[25]和CT3D[26]。另外，SE-SSD[22]和CenterPoint[23]等方法借鉴了蒸馏模型和CenterNet的思想。这些方法的出现表明二维目标检测技术对于三维点云目标检测具有迁移潜力。

4) 弱监督学习与自监督学习

标注三维目标检测数据集需要大量的人力和时间投入。为了降低对标注数据的依赖性，研究人员已经开始探索弱监督学习和无监督学习这两种方法在三维目标检测领域的应用[80-82]。弱监督学习只需要一组弱标注的场景和一些精确标注的目标，就能实现良好的目标检测效果；无监督学习则是利用无标注数据进行

表 3 点云目标检测算法在 Waymo 数据集上的平均精度(p_{AP})对比

Table 3 Average precision (p_{AP}) comparison of point cloud object detection methods on Waymo dataset unit: %

Level	Method	3D				BEV			
		Overall	0-30 m	30-50 m	50 m-Inf	Overall	0-30 m	30-50 m	50 m-Inf
LEVEL_1 (IoU is 0.7)	PointPillars[18]	56.62	81.01	51.75	27.94	75.57	92.10	74.06	55.47
	MVF[74]	62.93	86.30	60.02	36.02	80.40	93.59	79.21	63.09
	PV R-CNN[47]	70.30	91.92	69.21	42.17	82.96	97.35	82.99	64.97
	Pillar-OD[75]	69.80	88.53	66.50	42.93	87.11	95.78	84.87	72.12
	Voxel R-CNN[19]	75.59	92.49	74.09	53.15	88.19	97.62	87.34	77.70
	LiDAR R-CNN[76]	76.00	92.10	74.60	54.50	90.10	97.00	89.50	78.90
	CenterPoint[23]	76.86	92.27	75.31	54.10	91.61	97.19	91.05	82.06
	PVGNet[77]	74.00	—	—	—	—	—	—	—
	VoTR-TSD[25]	74.95	92.28	73.36	51.09	—	—	—	—
	CT3D[26]	76.30	92.51	75.07	55.36	90.50	**97.64**	88.06	78.89
	Pyramid-PV[78]	76.30	92.67	74.91	54.54	—	—	—	—
	VoxSeT[27]	76.02	91.13	75.75	54.23	89.12	95.12	87.36	77.78
	GraR-Ce[44]	**80.77**	**93.59**	**79.68**	**60.41**	**92.69**	97.56	**92.15**	**84.13**
	ImpDet[73]	74.38	91.98	72.86	49.13	—	—	—	—
LEVEL_2 (IoU is 0.7)	PV R-CNN[47]	65.36	91.58	65.13	36.46	77.45	94.64	80.39	55.39
	Voxel R-CNN[19]	66.59	91.74	67.89	40.80	81.07	96.99	81.37	63.26
	LiDAR R-CNN[76]	68.30	91.30	68.50	42.40	81.70	94.30	82.30	65.80
	CenterPoint[23]	69.09	91.41	69.43	42.40	85.43	96.35	86.44	70.06
	VoTR-TSD[25]	65.91	—	—	—	—	—	—	—
	CT3D[26]	69.04	91.76	68.93	42.60	81.74	**97.05**	82.22	64.34
	Pyramid-PV[78]	67.23	—	—	—	—	—	—	—
	VoxSeT[27]	68.16	91.03	67.13	42.23	76.13	94.13	81.78	58.13
	GraR-Ce[44]	**72.55**	**92.75**	**73.74**	**47.84**	**86.56**	96.79	**87.59**	**72.06**

训练，通过学习数据集内部的特征和分布来推断目标信息。进一步深入研究这两种学习方法的原理，可以帮助降低算法对大规模标注数据的依赖性，从而提高三维目标检测的效率和可扩展性。

5) 复杂数据集的构建与使用

深度学习方法是数据驱动型算法，数据集的质量对网络学习的效果具有重要影响。为了更好地模拟真实情况并提高模型的适应能力，构建复杂和多样的数据集是至关重要的。在三维目标检测领域，数据集的构建需要考虑多种因素，包括不同天气条件下的场景，如晴天、雨天、雪天和雾天等。此外，数据集还应覆盖不同的时间段，包括白天和黑夜，以及不同地区，如城市、农村等。这样能够使模型在面对复杂和多样化的情况时具有更好的适应能力。

7 结束语

点云数据具有丰富的几何信息，这为实现精确的三维目标检测提供了有力支持。近年来，深度学习方法的广泛应用使得基于深度学习的点云目标检测技术得到了快速发展，在计算机视觉领域成为研究的热点。本文首先对点云的独特性和特征提取方法进行介绍，并根据不同分类阐述点云目标检测方法的相关原理和发展历程；然后，介绍常用的数据集和评价指标，并对不同方法的性能进行量化比较。整体而言，基于点云数据的三维目标检测技术具有巨大的研究价值和良好的应用前景。未来的研究方向包括多分支和多模态融合、迁移二维目标检测方法、弱监督学习和无监督学习，以及复杂数据集的构建和使用等。这些方向的探索将进一步推动点云目标检测技术的发展，提高其在实际应用中的可靠性。

参 考 文 献

[1] Lalonde J F, Unnikrishnan R, Vandapel N, et al. Scale selection for classification of point-sampled 3D surfaces[C]// Fifth International Conference on 3-D Digital Imaging and Modeling (3DIM'05), June 13-16, 2005, Ottawa, ON, Canada. New York: IEEE Press, 2005: 285-292.

[2] Gao Z H, Liu X W. Support vector machine and object-oriented classification for urban impervious surface extraction from satellite imagery[C]//2014 The Third International Conference on Agro-Geoinformatics, August 11-14, 2014, Beijing, China. New York: IEEE Press, 2014.

[3] Zheng G, Zhong L, Li Y F, et al. A random forest based

method for urban object classification using lidar data and aerial imagery[C]//2015 23rd International Conference on Geoinformatics, June 19-21, 2015, Wuhan, China. New York: IEEE Press, 2016.

[4] Munoz D, Bagnell J A, Vandapel N, et al. Contextual classification with functional Max-Margin Markov Networks [C]//2009 IEEE Conference on Computer Vision and Pattern Recognition, June 20-25, 2009, Miami, FL, USA. New York: IEEE Press, 2009: 975-982.

[5] Niemeyer J, Rottensteiner F, Soergel U. Contextual classification of lidar data and building object detection in urban areas[J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2014, 87: 152-165.

[6] LeCun Y, Bengio Y, Hinton G. Deep learning[J]. Nature, 2015, 521(7553): 436-444.

[7] Geiger A, Lenz P, Urtasun R. Are we ready for autonomous driving? The KITTI vision benchmark suite[C]//2012 IEEE Conference on Computer Vision and Pattern Recognition, June 16-21, 2012, Providence, RI, USA. New York: IEEE Press, 2012: 3354-3361.

[8] Maturana D, Scherer S. VoxNet: a 3D Convolutional Neural Network for real-time object recognition[C]//2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), September 28 - October 2, 2015, Hamburg, Germany. New York: IEEE Press, 2015: 922-928.

[9] Xu Y, Hoegner L, Tuttas S, et al. Voxel- and graph-based point cloud segmentation of 3D scenes using perceptual grouping laws[J]. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 2017, IV-1/W1: 43-50.

[10] Qi C R, Hao S, Mo K C, et al. PointNet: deep learning on point sets for 3D classification and segmentation[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 77-85.

[11] Qi C R, Litany O, He K M, et al. Deep Hough voting for 3D object detection in point clouds[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV), October 27-November 2, 2019, Seoul, Korea (South). New York: IEEE Press, 2020: 9276-9285.

[12] Qi X J, Liao R J, Jia J Y, et al. 3D graph neural networks for RGBD semantic segmentation[C]//2017 IEEE International Conference on Computer Vision (ICCV), October 22-29, 2017, Venice, Italy. New York: IEEE Press, 2017: 5209-5218.

[13] Landrieu L, Simonovsky M. Large-scale point cloud semantic segmentation with superpoint graphs[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 4558-4567.

[14] Bi Y, Chadha A, Abbas A, et al. Graph-based object classification for neuromorphic vision sensing[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV), October 27 - November 2, 2019, Seoul, Korea (South). New York: IEEE Press, 2020: 491-501.

[15] Wang Y, Sun Y B, Liu Z W, et al. Dynamic graph CNN for learning on point clouds[J]. ACM Transactions on Graphics, 38 (5): 1-12.

[16] Zhou Y, Tuzel O. VoxelNet: end-to-end learning for point cloud based 3D object detection[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 4490-4499.

[17] Yan Y, Mao Y X, Li B. SECOND: sparsely embedded convolutional detection[J]. Sensors, 2018, 18(10): 3337.

[18] Lang A H, Vora S, Caesar H, et al. PointPillars: fast encoders for object detection from point clouds[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 15-20, 2019, Long Beach, CA, USA. New York: IEEE Press, 2020: 12689-12697.

[19] Deng J J, Shi S S, Li P W, et al. Voxel R-CNN: towards high performance voxel-based 3D object detection[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 35(2): 1201-1209.

[20] Wu H, Wen C L, Li W, et al. Transformation-equivariant 3D object detection for autonomous driving[EB/OL]. (2022-11-22) [2023-02-04]. https://arxiv.org/abs/2211.11962.

[21] Zheng W, Tang W L, Chen S J, et al. CIA-SSD: confident IoU-aware single-stage object detector from point cloud[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 35(4): 3555-3562.

[22] Zheng W, Tang W L, Jiang L, et al. SE-SSD: self-ensembling single-stage object detector from point cloud[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 20-25, 2021, Nashville, TN, USA. New York: IEEE Press, 2021: 14489-14498.

[23] Yin T W, Zhou X Y, Krähenbühl P. Center-based 3D object detection and tracking[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 20-25, 2021, Nashville, TN, USA. New York: IEEE Press, 2021: 11779-11788.

[24] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems, December 4-9, 2017, Long Beach, California, USA. New York: ACM, 2017: 6000-6010.

[25] Mao J G, Xue Y J, Niu M Z, et al. Voxel transformer for 3D object detection[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV), October 10-17, 2021, Montreal, QC, Canada. New York: IEEE Press, 2022: 3144-3153.

[26] Sheng H L, Cai S J, Liu Y, et al. Improving 3D object detection with channel-wise transformer[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV), October 10-17, 2021, Montreal, QC, Canada. New York: IEEE Press, 2022: 2723-2732.

[27] He C H, Li R H, Li S, et al. Voxel set transformer: a set-to-set approach to 3D object detection from point clouds[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 18-24, 2022, New Orleans, LA, USA. New York: IEEE Press, 2022: 8407-8417.

[28] Dong S, Ding L, Wang H, et al. MsSVT: mixed-scale sparse voxel transformer for 3D object detection on point clouds[C]// Advances in Neural Information Processing Systems, May 16-19, 2022, New Orleans, LA, USA. Canada: NIPS, 2022.

[29] Ding L H, Dong S C, Xu T F, et al. FH-net: a fast hierarchical network for scene flow estimation on real-world point clouds [M]//Avidan S, Brostow G, Cissé M, et al. Computer vision-ECCV 2022. Lecture notes in computer science. Cham: Springer, 2022, 13699: 213-229.

[30] Zhang Y N, Chen J X, Huang D. CAT-det: contrastively augmented transformer for multimodal 3D object detection[C]// 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 18-24, 2022, New Orleans, LA, USA. New York: IEEE Press, 2022: 898-907.

[31] Xie Q, Lai Y K, Wu J, et al. MLCVNet: multi-level context VoteNet for 3D object detection[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 13-19, 2020, Seattle, WA, USA. New York: IEEE Press, 2020: 10444-10453.

[32] Xie Q, Lai Y K, Wu J, et al. Vote-based 3D object detection with context modeling and SOB-3DNMS[J]. International Journal of Computer Vision, 2021, 129(6): 1857-1874.

[33] Chen X X, Zhao H, Zhou G Y, et al. PQ-transformer: jointly parsing 3D objects and layouts from point clouds[J]. IEEE Robotics and Automation Letters, 2022, 7(2): 2519-2526.

[34] Misra I, Girdhar R, Joulin A. An end-to-end transformer model for 3D object detection[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV), October 10-17, 2021, Montreal, QC, Canada. New York: IEEE Press, 2022: 2886-2897.

[35] Xu X, Dong S, Xu T, et al. FusionRCNN: LiDAR-camera fusion for two-stage 3D object detection[J]. Remote Sensing, 2023, 15(7): 1839.

[36] Hu J S K, Kuai T S, Waslander S L. Point density-aware voxels for LiDAR 3D object detection[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 18-24, 2022, New Orleans, LA, USA. New York: IEEE Press, 2022: 8459-8468.

[37] Qi C R, Yi L, Su H, et al. PointNet++: deep hierarchical feature learning on point sets in a metric space[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems, December 4-9, 2017, Long Beach, California, USA. New York: ACM Press, 2017: 5105-5114.

[38] Shi S S, Wang X G, Li H S. PointRCNN: 3D object proposal generation and detection from point cloud[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 15-20, 2019, Long Beach, CA, USA. New York: IEEE Press, 2020: 770-779.

[39] Yang Z T, Sun Y N, Liu S, et al. 3DSSD: point-based 3D single stage object detector[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 13-19, 2020, Seattle, WA, USA. New York: IEEE Press, 2020: 11037-11045.

[40] Zhang Y F, Hu Q Y, Xu G Q, et al. Not all points are equal: learning highly efficient point-based detectors for 3D LiDAR point clouds[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 18-24, 2022, New Orleans, LA, USA. New York: IEEE Press, 2022: 18931-18940.

[41] Chen C, Chen Z, Zhang J, et al. SASA: semantics-augmented set abstraction for point-based 3D object detection[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2022, 36(1): 221-229.

[42] Shi W J, Rajkumar R. Point-GNN: graph neural network for 3D object detection in a point cloud[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 13-19, 2020, Seattle, WA, USA. New York: IEEE Press, 2020: 1708-1716.

[43] Zhang Y N, Huang D, Wang Y H. PC-RGNN: point cloud completion and graph neural network for 3D object detection[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 35(4): 3430-3437.

[44] Yang H H, Liu Z L, Wu X P, et al. Graph R-CNN: towards accurate 3D object detection with semantic-decorated local graph [M]//Avidan S, Brostow G, Cissé M, et al. Computer vision - ECCV 2022. Lecture notes in computer science. Cham: Springer, 2022, 13668: 662-679.

[45] Chen Y L, Liu S, Shen X Y, et al. Fast point R-CNN[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV), October 27-November 2, 2019, Seoul, Republic of Korea. New York: IEEE Press, 2020: 9774-9783.

[46] Yang Z T, Sun Y N, Liu S, et al. STD: sparse-to-dense 3D object detector for point cloud[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV), October 27-November 2, 2019, Seoul, Korea (South). New York: IEEE Press, 2020: 1951-1960.

[47] Shi S S, Guo C X, Jiang L, et al. PV-RCNN: point-voxel feature set abstraction for 3D object detection[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 13-19, 2020, Seattle, WA, USA. New York: IEEE Press, 2020: 10526-10535.

[48] He C H, Zeng H, Huang J Q, et al. Structure aware single-stage 3D object detection from point cloud[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 13-19, 2020, Seattle, WA, USA. New York: IEEE Press, 2020: 11870-11879.

[49] Liu Z J, Tang H T, Amini A, et al. BEVFusion: multi-task multi-sensor fusion with unified bird's-eye view representation [EB/OL]. (2022-05-26) [2023-02-01]. https://arxiv. org/abs/2205.13542.

[50] Qi C R, Liu W, Wu C X, et al. Frustum PointNets for 3D object detection from RGB-D data[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 918-927.

[51] Wang Z X, Jia K. Frustum ConvNet: sliding Frustums to aggregate local point-wise features for amodal 3D object detection[C]//2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), November 3-8, 2019, Macau, China. New York: IEEE Press, 2020: 1742-1749.

[52] Chen X Z, Ma H M, Wan J, et al. Multi-view 3D object detection network for autonomous driving[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 6526-6534.

[53] Ku J, Mozifian M, Lee J, et al. Joint 3D proposal generation and object detection from view aggregation[C]//2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), October 1-5, 2018, Madrid, Spain. New York: IEEE Press, 2019.

[54] Chen X Y, Zhang T Y, Wang Y, et al. FUTR3D: a unified sensor fusion framework for 3D detection[EB/OL]. (2022-03-20) [2023-02-03]. https://arxiv.org/abs/2203.10642.

[55] Bai X Y, Hu Z Y, Zhu X G, et al. TransFusion: robust LiDAR-camera fusion for 3D object detection with transformers[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 18-24, 2022, New Orleans, LA, USA. New York: IEEE Press, 2022: 1080-1089.

[56] Vora S, Lang A H, Helou B, et al. PointPainting: sequential fusion for 3D object detection[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 13-19, 2020, Seattle, WA, USA. New York: IEEE Press, 2020: 4603-4611.

[57] Yin T W, Zhou X Y, Krähenbühl P. Multimodal virtual point 3D detection[EB/OL]. (2021-11-12)[2023-02-06]. https://arxiv.org/abs/2111.06881.

[58] Liang M, Yang B, Wang S L, et al. Deep continuous fusion for multi-sensor 3D object detection[M]//Ferrari V, Hebert M, Sminchisescu C, et al. Computer vision - ECCV 2018. Lecture notes in computer science. Cham: Springer, 2018, 11220: 663-678.

[59] Liang M, Yang B, Chen Y, et al. Multi-task multi-sensor fusion for 3D object detection[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 15-20, 2019, Long Beach, CA, USA. New York: IEEE Press, 2020: 7337-7345.

[60] 王亚东, 田永林, 李国强, 等. 基于卷积神经网络的三维目标检测研究综述[J]. 模式识别与人工智能, 2021, 34(12): 1103-1119.
Wang Y D, Tian Y L, Li G Q, et al. 3D object detection based on convolutional neural networks: a survey[J]. Pattern Recognition and Artificial Intelligence, 2021, 34(12): 1103-1119.

[61] Sun P, Kretzschmar H, Dotiwalla X, et al. Scalability in perception for autonomous driving: waymo open dataset[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 13-19, 2020, Seattle, WA, USA. New York: IEEE Press, 2020: 2443-2451.

[62] Caesar H, Bankiti V, Lang A H, et al. nuScenes: a multimodal dataset for autonomous driving[C]//2020 IEEE/CVF

Conference on Computer Vision and Pattern Recognition (CVPR), June 13-19, 2020, Seattle, WA, USA. New York: IEEE Press, 2020: 11618-11628.

[63] Cong P S, Zhu X G, Qiao F, et al. STCrowd: a multimodal dataset for pedestrian perception in crowded scenes[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 18-24, 2022, New Orleans, LA, USA. New York: IEEE Press, 2022: 19608-19617.

[64] Silberman N, Hoiem D, Kohli P, et al. Indoor segmentation and support inference from RGBD images[M]//Fitzgibbon A, Lazebnik S, Perona P, et al. Computer vision - ECCV 2012. Lecture notes in computer science. Heidelberg: Springer, 2012, 7576: 746-760.

[65] Xiao J X, Owens A, Torralba A. SUN3D: a database of big spaces reconstructed using SfM and object labels[C]//2013 IEEE International Conference on Computer Vision, December 1-8, 2013. Sydney, Australia. New York: IEEE Press, 2013: 1625-1632.

[66] Song S R, Lichtenberg S P, Xiao J X. SUN RGB-D: a RGB-D scene understanding benchmark suite[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 7-12, 2015, Boston, MA, USA. New York: IEEE Press, 2015: 567-576.

[67] Dai A, Chang A X, Savva M, et al. ScanNet: richly-annotated 3D reconstructions of indoor scenes[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 2432-2443.

[68] Qian R, Lai X, Li X R. 3D object detection for autonomous driving: a survey[J]. Pattern Recognition, 2022, 130: 108796.

[69] Simonelli A, Bulò S R, Porzi L, et al. Disentangling monocular 3D object detection[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV), October 27-November 2, 2019, Seoul, Republic of Korea. New York: IEEE Press, 2020: 1991-1999.

[70] Shi S S, Wang Z, Shi J P, 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, 2021, 43(8): 2647-2664.

[71] Liu Z, Zhao X, Huang T T, et al. TANet: robust 3D object detection from point clouds with triple attention[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(7): 11677-11684.

[72] Yi H W, Shi S S, Ding M Y, et al. SegVoxelNet: exploring semantic context and depth-aware features for 3D vehicle detection from point cloud[C]//2020 IEEE International Conference on Robotics and Automation (ICRA), May 31-August 31, 2020, Paris, France. New York: IEEE Press, 2020: 2274-2280.

[73] Qian X L, Wang L, Zhu Y, et al. ImpDet: exploring implicit fields for 3D object detection[C]//2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), January 2-7, 2023, Waikoloa, HI, USA. New York: IEEE Press, 2023: 4249-4259.

[74] Zhou Y, Sun P, Zhang Y, et al. End-to-end multi-view fusion for 3D object detection in LiDAR point clouds[EB/OL]. (2019-10-23)[2023-02-01]. https://arxiv.org/abs/1910.06528.

[75] Wang Y E, Fathi A, Kundu A, et al. Pillar-based object detection for autonomous driving[M]//Vedaldi A, Bischof H, Brox T, et al. Computer vision - ECCV 2020. Lecture notes in computer science. Cham: Springer, 2020, 12367: 18-34.

[76] Li Z C, Wang F, Wang N Y. LiDAR R-CNN: an efficient and universal 3D object detector[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 20-25, 2021, Nashville, TN, USA. New York: IEEE Press, 2021: 7542-7551.

[77] Miao Z W, Chen J K, Pan H Y, et al. PVGNet: a bottom-up one-stage 3D object detector with integrated multi-level features [C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 20-25, 2021, Nashville, TN, USA. New York: IEEE Press, 2021: 3278-3287.

[78] Mao J G, Niu M Z, Bai H Y, et al. Pyramid R-CNN: towards better performance and adaptability for 3D object detection[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV), October 10-17, 2021, Montreal, QC, Canada. New York: IEEE Press, 2022: 2703-2712.

[79] Wang Y, Chao W L, Garg D, et al. Pseudo-LiDAR from visual depth estimation: bridging the gap in 3D object detection for autonomous driving[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 15-20, 2019, Long Beach, CA, USA. New York: IEEE Press, 2020: 8437-8445.

[80] Meng Q H, Wang W G, Zhou T F, et al. Weakly supervised 3D object detection from lidar point cloud[M]//Vedaldi A, Bischof H, Brox T, et al. Computer Vision - ECCV 2020. Lecture notes in computer science. Cham: Springer, 2020, 12358: 515-531.

[81] Zhang Z W, Girdhar R, Joulin A, et al. Self-supervised pretraining of 3D features on any point-cloud[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV), October 10-17, 2021, Montreal, QC, Canada. New York: IEEE Press, 2022: 10232-10243.

[82] Luo Z P, Cai Z A, Zhou C Q, et al. Unsupervised domain adaptive 3D detection with multi-level consistency[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV), October 10-17, 2021, Montreal, QC, Canada. New York: IEEE Press, 2022: 8846-8855.

Three-Dimensional Object Detection Technology Based on Point Cloud Data

Li Jianan[1,2], Wang Ze[1], Xu Tingfa[1,2,3*]

[1]*School of Optoelectronics, Beijing Institute of Technology, Beijing* 100081, *China*;

[2]*Key Laboratory of Photoelectronic Imaging Technology and System, Ministry of Education, Beijing Institute of Technology, Beijing* 100081, *China*;

[3]*Chongqing Innovation Center, Beijing Institute of Technology, Chongqing* 401135, *China*

Abstract

Significance In recent years, self-driving technology has garnered considerable attention from both academia and industry. Autonomous perception, which encompasses the perception of the vehicle′s state and the surrounding environment, is a critical component of self-driving technology, guiding decision-making and planning modules. In order to perceive the environment accurately, it is necessary to detect objects in three-dimensional (3D) scenes. However, traditional 3D object detection techniques are typically based on image data, which lack depth information. This makes it challenging to use image-based object detection in 3D scene tasks. Therefore, 3D object detection predominantly relies on point cloud data obtained from devices such as lidar and 3D scanners.

Point cloud data consist of a collection of points, with each containing coordinate information and additional attributes such as color, normal vector, and intensity. Point cloud data are rich in depth information. However, in contrast to two-dimensional images, point cloud data are sparse and unordered, and they exhibit a complex and irregular structure, posing challenges for feature extraction processes. Traditional methods rely on local point cloud information such as curvature, normal vector, and density, combined with methods such as the Gaussian model to manually design descriptors for processing point cloud data. However, these methods rely heavily on *a priori* knowledge and fail to account for the relationships between neighboring points, resulting in low robustness and susceptibility to noise.

In recent years, deep learning methods have gained significant attention from researchers due to their robust feature representation and generalization capabilities. The effectiveness of deep learning methods relies heavily on high-quality datasets. To advance the field of point cloud object detection, numerous companies such as Waymo and Baidu, as well as research institutes have produced large-scale point cloud datasets. With the help of such datasets, point cloud object detection combined with deep learning has rapidly developed and demonstrated powerful performance. Despite the progress made in this field, challenges related to accuracy and real-time performance still exist. Therefore, this paper provides a review of the research conducted in point cloud object detection and looks forward to future developments to promote the advancement of this field.

Progress The development of point cloud object detection has been significantly promoted by the recent emergence of large-scale open-source datasets. Several standard datasets for outdoor scenes, including KITTI, Waymo, and nuScenes, as well as indoor scenes, including NYU-Depth, SUN RGB-D, and ScanNet, have been released, which have greatly facilitated research in this field. The relevant properties of these datasets are summarized in Table 1.

Point cloud data are characterized by sparsity, non-uniformity, and disorder, which distinguish them from image data. To address these unique properties of point clouds, researchers have developed a range of object detection algorithms specifically designed for this type of data. Based on the methods of feature extraction, point cloud-based single-modal methods can be categorized into four groups: voxel-based, point-based, graph-based, and point+voxel-based methods. Voxel-based methods divide the point cloud into regular voxel grids and aggregate point cloud features within each voxel to generate regular four-dimensional feature maps. VoxelNet, SECOND, and PointPillars are classic architectures of this kind of method. Point-based methods process the point cloud directly and utilize symmetric functions to aggregate point cloud features while retaining the geometric information of the point cloud to the greatest extent. PointNet, PointNet++, and Point R-CNN are their classic architectures. Graph-based methods convert the point cloud into a graph representation and process it through the graph neural network. Point GNN and Graph R-CNN are classic architectures of this approach. Point+voxel-based methods combine the methods based on point and those based on voxel, with STD and PV R-CNN as classic architectures. In addition, to enhance the semantic information of point cloud data, researchers have used image data to supplement secondary information to design multi-modal methods. MV3D, AVOD, and MMF are classic

architectures of multi-modal methods. A chronological summary of classical methods for object detection from point clouds is presented in Fig. 4.

Conclusions and Prospects The field of 3D object detection from point clouds is a significant research area in computer vision that is gaining increasing attention from scholars. The foundational branch of 3D object detection from point clouds has flourished, and future research may focus on several areas. These include multi-branch and multi-mode fusion, the integration of two-dimensional detection methods, weakly supervised and self-supervised learning, and the creation and utilization of complex datasets.

Key words point cloud; 3D object detection; single modality; multi-modality