计算机科学
COMPUTER SCIENCE

面向自动驾驶的三维目标检测综述

霍威乐，荆涛，任爽

引用本文

霍威乐，荆涛，任爽.面向自动驾驶的三维目标检测综述[J]．计算机科学, 2023, 50(7): 107-118.

HUO Weile, JING Tao, REN Shuang. Review of 3D Object Detection for Autonomous Driving[J]. Computer Science, 2023, 50(7): 107-118.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.220700090

面向自动驾驶的三维目标检测综述

霍威乐　荆　涛　任　爽
北京交通大学计算机与信息技术学院　北京 100044
(wlhuo0365@bjtu.edu.cn)

摘　要　近年来,随着自动驾驶行业的蓬勃发展,作为感知系统核心的三维目标检测技术受到越来越多的关注,已成为当前热门的研究方向。同时,深度学习的广泛应用,使得最近的三维目标检测技术有了很大的突破,大批优秀的算法涌现。文中系统地总结了面向自动驾驶领域的三维目标检测方法,并按传感器类型将现有的算法分为3类,即基于图像的三维目标检测、基于LiDAR的三维目标检测和基于多传感器的三维目标检测;其次,详细分析了3种方法的优缺点,并对基于LiDAR的三维目标检测算法进行了深入调研和细分;然后,介绍了自动驾驶领域常用的三维目标检测数据集,包括KITTI,nuScenes和Waymo Open Dataset,并对比了最新的三维目标检测算法在不同数据集上的性能表现;最后探讨了三维目标检测技术未来的发展方向。

关键词:自动驾驶;三维目标检测;深度学习;点云;激光雷达
中图法分类号　TP391.41

Review of 3D Object Detection for Autonomous Driving

HUO Weile,JING Tao and REN Shuang
School of Computer and Information Technology,Beijing Jiaotong University,Beijing 100044,China

Abstract　In recent years,with the rapid development of autonomous driving,3D object detection technology as the core of perception systems has received more and more attention and become a hot research direction. At the same time,the wide application of deep learning has made a great breakthrough in 3D object detection technology recently. A large number of excellent algorithms have emerged. This paper systematically summarizes 3D object detection methods for the autonomous driving field and divides the existing algorithms into three categories according to sensor types:image-based 3D object detection,LiDAR-based 3D object detection,and multi-sensor-based 3D object detection. After that,it analyzes the advantages and disadvantages of the three methods in detail. The LiDAR-based 3D object detection algorithms are thoroughly investigated and subdivided. Then it introduces the commonly used 3D object detection datasets in autonomous driving,including KITTI,nuScenes,and Waymo Open Dataset,and compares the performance of the latest 3D object detection algorithms on different datasets. Finally,the future research direction of 3D object detection technology is discussed.

Keywords　Autonomous driving,3D object detection,Deep learning,Point cloud,LiDAR

1　引言

近年来,随着深度学习的崛起,二维目标检测取得了重大突破[1-4],如人脸检测、行人检测等已经在现实生活中得到了广泛应用。同时,自动驾驶技术的兴起使得三维目标检测也逐步受到重视。三维目标检测可细分为分类和定位两个子任务,利用紧密的三维边界框定位目标,来确定目标的类别、位置、大小、方向等信息[5]。

在自动驾驶场景中,通常使用相机或激光雷达(Light Detection And Ranging,LiDAR)来捕捉RGB图像或三维点云。早期的三维目标检测通常仅使用图像作为输入[6-10],首先利用二维目标检测模型检测出二维的边界框,再通过深度估计将二维边界框转换到三维空间。但由于图像中缺乏精确的深度信息,这种检测方法性能较差。

三维点云是由激光雷达对物体表面进行扫描得到的一组无序点的集合,相比图像,它能够提供物体在三维空间中精确的位置信息。然而,由于点云具有不规则性和高度稀疏性,直接在三维点云数据上移植二维检测方法比较困难。受图像中"像素"的影响,大多数研究人员将点云数据转换为规则的网格形式,即"体素"[11-16]。一些方法[11-12]将体素编码为手工设计的特征,继而使用支持向量机(Support Vector Machines,SVM)或卷积神经网络(Convolutional Neural Network,

到稿日期:2022-07-11　返修日期:2022-11-20
基金项目:国家自然科学基金(62072025)
This work was supported by the National Natural Science Foundation of China(62072025).
通信作者:任爽(sren@bjtu.edu.cn)

CNN)完成检测。然而这些手工设计的特征会带来信息瓶颈,难以应对不同场景的检测任务。此后,随着对人工智能技术的深入研究,机器学习的特征[13-14]替代了手工设计特征,三维目标检测技术有了重大突破。之后,随着三维稀疏卷积的引入[14],基于体素的三维目标检测方法表现出了更好的性能,能够达到实时检测的效果。然而,点云的体素化过程不可避免地引入了量化误差,从而导致了细粒度信息的丢失。

2017 年,Qi 等提出了 PointNet[17],开创了直接使用原始点云进行特征提取的先河。PointNet 使用最大池化和 T-Net 解决了点云数据的无序性和旋转不变性问题。之后,针对 PointNet 中局部信息缺失的问题,他们又提出了 PointNet++[18],通过集抽象(Set Abstraction,SA)操作学习不同尺度的局部特征。PointNet 系列可以作为一个独立的结构,广泛用于不同的点云处理任务中[19-24]。在三维目标检测领域,一些方法使用 PointNet++对点云的原始点特征进行提取[19,25-29],其灵活的接受域使得网络能够学习到多尺度的点特征。

除了激光雷达,雷达(Radio Detection and Ranging,Radar)也可以用来获取点云。雷达通过发射无线电波来探测物体的距离、速度和方向,工作在毫米波段的雷达被称为毫米波雷达。由于雷达无法探测出物体的高度信息,因此通常使用雷达数据作为其他模态数据的补充信息[30-31]。4D 毫米波雷达的出现弥补了这一不足。所谓 4D,就是在传统雷达的基础上增加了一维高度信息,这使得雷达点云更加适用于三维目标检测任务[32-33]。Xu 等[32]提出了一种 RPFA-Net 检测网络,将 4D 雷达点云处理成柱形体素,并基于自注意力机制提取全局特征。Li[33]基于 PointRCNN[19]提出了一种改进的三分支密度感知网络,对 4D 雷达点云的密度不均匀问题进行了研究。然而,目前包含 4D 雷达数据的数据集较少,常用的一个数据集是 Astyx HiRes2019[34],但其数据量较小。因此,总的来说,基于 4D 毫米波雷达的三维目标检测发展较为缓慢。

本文系统地总结了近年来自动驾驶场景下三维目标检测领域的研究进展,并对不同方法进行了细致的分类。总的来说,按照传感器的类型可将现有方法分为 3 类:基于图像、基于 LiDAR,以及基于多传感器的三维目标检测方法,并分别对其进行了综述。其中,我们对当前的主流方法,即基于 LiDAR 的检测方法进行了深入调研。现有文献大都根据点云的表示形式将其分为点和体素两条支线,并据此概括为基于体素的(Voxel-based)、基于点的(Point-based)、基于点和体素的(Point-Voxel-based)3 个子类[35]。本文扩展了基于 LiDAR 的三维目标检测方法的分类,依据是否将三维点云二维化提出了一种新的基于视图的(View-based)子类。遵从这种思想,本文还将基于多传感器的三维目标检测方法细分为多视图融合方法和多模态融合方法。此外,本文还介绍了自动驾驶领域常用的数据集,并在不同数据集上对现有算法的性能进行了对比。最后,根据目前的三维目标检测算法存在的问题,对未来的发展方向进行了探讨。

2 基于图像的三维目标检测

早期的三维目标检测仅使用图像作为输入,基于经典的二维目标检测网络完成检测。Chen 等[7-8,10]将候选框的生成问题转化为最小化能量函数的问题,然后利用几何先验构造能量函数,使用 SVM 学习能量函数的权重,最后基于 Faster r-cnn[36]构建检测网络。Song 等[37]在单目视频序列中联合 SFM(Structure from Motion)线索和二维目标检测线索,同时定位 3D 场景中的近距离和远距离对象。还有一些方法利用滑动窗口进行检测,Song 等[6]首先将 CAD 模型渲染为深度图,并提取深度图的特征,最后使用滑动窗口检测物体。Deep Sliding Shapes[9]基于滑动窗口的思想设计了 RGB-D 图像上的多尺度区域生成网络,并根据 3D 识别和 2D 识别的结果进行联合目标识别。

基于图像的三维目标检测方法,其本质还是二维的。相比三维点云数据,图像上物体所占的面积大小受距离的影响,当物体距离相机较远时,该物体在图像上所占据的像素点的个数就少,且图像中缺乏精确的深度信息,这些特点使得大部分基于图像的三维目标检测方法的精度落后于基于 LiDAR 的检测方法。

尽管基于图像的三维目标检测方法性能还有待提升,但由于相比激光雷达,相机的成本更低,因此在某些场景下相机仍然是昂贵的激光雷达的替代品。近期有一些方法致力于将 RGB 图像生成伪激光雷达(Pseudo-LiDAR)[38-39]后再进行处理。这些方法通常由深度估计和目标检测两个网络组成,首先对 RGB 图像进行深度估计,然后利用深度将其投影为一个伪激光雷达,最后像处理一般的点云那样使用三维方法进行检测。AM3D[40]在目标检测网络中还进一步融合了 RGB 图像的特征和 2D 检测结果。针对伪激光雷达中两个网络须独立训练的问题,Qian 等[41]引入了一个可微的表示变换(Change of Representation,CoR),将深度估计网络的输出作为三维目标检测网络的输入,实现了端到端的训练。

总的来说,基于图像的三维目标检测由于受到本身数据形式的限制,在检测精度上很难有较大的突破,但在某些场景下,仍然具有广阔的应用前景。另一方面,伪激光雷达方法的提出,也为其提供了新的解决思路,有利于缩小基于图像和基于 LiDAR 检测之间的精度差异。

3 基于 LiDAR 的三维目标检测

基于 LiDAR 的三维目标检测方法仅以点云作为输入,在精度和效率上都能达到较高水平,是目前的主流算法。与图像不同,LiDAR 扫描的点云数据具有一些特殊的性质。一方面,点云提供了三维空间信息,包括精确的深度及位置,为三维目标检测提供了数据支撑;另一方面,点云具有稀疏性、无序性以及分布不均匀性,给数据的处理带来了极大的挑战。

按照对 LiDAR 点云数据的处理方式,可以将现有基于 LiDAR 的三维目标检测方法分为以下 4 类:基于视图的 View-based 方法[42-46]、基于体素的 Voxel-based 方法[11-16,47-54]、基于原始点的 Point-based 方法[19,26-28,55-57]以及基于点和体素的 Point-Voxel-based 方法[29,58-62]。View-based 方法不直接对三维点云进行处理,而是将点云投影为鸟瞰图等二维视图,然后运用成熟的 2D 目标检测算法进行检测。Voxel-based 方法将三维空间划分为均匀的网格,每个网格

称为一个体素，将体素内部的点特征编码为体素特征，从而获得规则的网络输入。而 Point-based 方法直接对原始点云进行处理，利用多层感知机（Multilayer Perceptrons，MLPs）[17-18] 或图神经网络（Graph Neural Networks，GNNs）[56-57] 提取点的特征。最后，Point-Voxel-based 方法在特征提取中同时利用了点云的点表示与体素表示，充分结合了二者的优势。根据上述分类，我们在图 1 中列出了近年来具有代表性的三维目标检测算法。

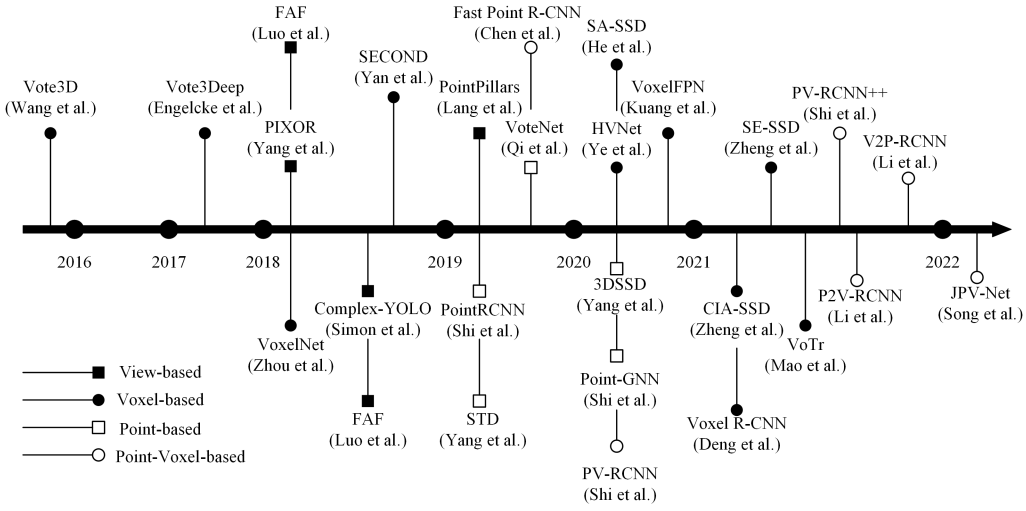


图 1　近年来三维目标检测算法时间轴

Fig. 1　Chronological overview of 3D object detection algorithms in recent years

3.1　View-based 方法

一些方法首先将三维点云数据投影到二维平面上，如鸟瞰图（Bird Eye View，BEV）、前视图（Front View，FV），再使用二维卷积网络处理映射后的点云，大大降低了计算的复杂度，能够达到实时检测的效果。相比稀疏的点云，在更为密集的二维平面，尤其是 BEV 上进行目标检测有 4 个方面的优势：1）BEV 中保留了物体在三维空间上的原始物理尺寸；2）由于对象在三维空间中是自然分离的，因此 BEV 上的所有对象之间均无遮挡；3）实际的交通场景中，所有物体均是位于地面的，垂直方向上的方差较小；4）相比三维卷积，二维卷积大大减少了计算量和复杂度。现有的文献中少有将这类方法独立为子类的，大多将其粗糙地归为 Voxel-based，然而这样的分类并不准确。相比直接处理原始点云的方法，这类方法的本质是将三维数据二维化，通过将其映射成其他视图对点云进行降维处理，究其根本，仍是在二维数据上进行的检测。因此本文提出了一种 View-based 的子类方法，以弥补现有分类方法的不足。

Yang 等提出的 PIXOR[42] 利用高度和反射率将点云转化为 BEV 表示，并添加小通道卷积层来提取更加精细的特征，使网络更有利于对小物体的检测，其网络结构如图 2 所示。Simon 等[43] 提出了 YOLOv2 网络[63] 在三维场景下的应用——Complex-YOLO，首先将点云转换为鸟瞰图，使用高度、强度、密度填充为图像的 RGB 通道，接着采用欧拉区域建议网络来回归边界框。得益于 YOLOv2 的架构，Complex-YOLO 具备了实时的检测速度，但在精度方面相较于其他方法还有很多不足，尤其是对小目标检测的性能不佳，如行人和骑行者。YOLO3D[44] 同样是在 YOLOv2 的架构上进行扩展，它将偏航角以及边界框的回归作为直接回归问题，并以此扩展了 YOLOv2 的损失函数。PointPillars[45] 将常用的三维卷积网络简化为二维卷积网络，通过将点云离散成 X-Y 平面上的一组柱形体素（Pillars），来将其映射为二维伪图像（Pseudo Image）。PointPillars 在 KITTI 数据集上的检测速度达到了 62FPS，远超大部分基于 LiDAR 的方法。Luo 等[46] 也将点云表示为 Pillars，并把高度作为特征通道，联合三维目标检测、跟踪与运动预测多个任务，减小了各任务的累计误差，并利用运动轨迹有效减少了当前帧中假阳性或假阴性的情况。

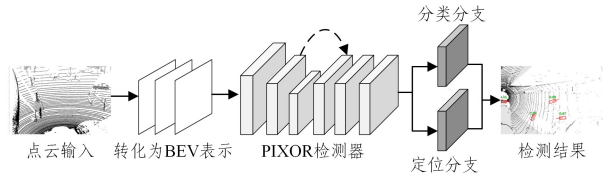


图 2　PIXOR 网络结构图

Fig. 2　Structure of PIXOR

利用投影等方式将点云处理为鸟瞰图或其他视图，简化了数据的处理流程，同时，使用二维卷积代替三维卷积，也极大地降低了计算时间成本。但另一方面，高度的压缩使得 BEV 丢失了很多信息，如目标内部各点之间的结构信息，从而导致网络对小目标的检测效果较差。

3.2　Voxel-based 方法

由于点云是一种不规则数据，受图像中像素的启发，研究人员倾向于将点云表示为体素网格的形式，进而将点云的特征量化为体素的特征，然后再使用神经网络进行训练[11-16,47-54]。

早期基于体素的方法大都使用一些手工设计的特征来表示体素。Vote3D[11] 使用一个固定维度的特征向量来表示体素网格，这些特征包括占用单元内点的分散度、反射率以及二元占用特征，然后采用滑动窗口的方式使用 SVM 分类器判断窗口内是否包含 RoI，最后通过投票得出每个窗口的得分。Engelcke 等[12] 基于 Vote3D 中的投票算法，构造卷积网络进行检测。3DFCN[15] 简化了体素的特征表示，以体素的占用

状态对体素进行二进制编码，并使用全卷积神经网络进行 3D 目标检测。这些手工设计的特征虽然在一些特定的数据集上表现良好，但难以适应复杂多变的真实自动驾驶场景。

随后，VoxelNet[13]的出现使得体素特征的获得从手工编码转向了利用机器学习获得，带来了性能上的重大突破。VoxelNet 引入了体素特征编码层（Voxel Feature Encoding，VFE），通过全连接网络（Fully Connected Network，FCN）对体素内部的所有点进行编码，再通过最大池化获得逐体素特征，其网络结构如图 3 所示。Shen 等[64]还在 VoxelNet 的基础上实现了一个两阶段的网络，进一步修正了区域生成网络（Region Proposal Network，RPN）的检测结果。VoxelNet 使用三维卷积作为骨干网络，来学习更多的体素上下文信息，但同时也引入了新的问题：随着体素分辨率的提高，计算复杂度呈指数级增长。因此，SECOND[14]使用稀疏三维卷积[65-67]替代传统的三维卷积，大大降低了体素表示的计算量，实现了点云数据的高效处理，从而使得基于体素的检测成为当前的主流方法。

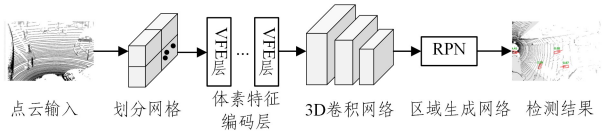


图 3 VoxelNet 网络结构图

Fig. 3 Structure of VoxelNet

Voxel-based 方法中体素的大小是一个很重要的影响因素，体素尺寸越小，特征体的分辨率就越高，但随之而来的计算成本也会急剧增加。因此，Ye 等[50]设计了混合体素特征提取（Hybrid Voxel Feature Extractor，HVFE）模块，融合了不同尺度的 VFE 层，并通过注意力机制将体素特征投影为不同比例的伪图像特征，最后使用特征融合金字塔融合特征生成最终的检测结果，实现了推断速度和精度上的平衡。

特征提取中采用的三维卷积操作会使特征体的分辨率逐渐降低，导致空间信息丢失。为此，Wang 等[16]在 VoxelNet[13]的结构上引入了特征金字塔，将具有不同分辨率的多尺度体素特征相融合用于最后的检测。而 He 等[49]在训练中添加了一个辅助网络，来学习点云的空间结构信息，在原始点云的分辨率上聚合体素特征，用作前景分割和中心点预测任务的特征输入。受人类的联想识别过程的启发，Du 等[51]提出将弱感知特征和鲁棒的概念特征联系起来，通过对齐感知域与概念域，使得网络能够自适应地生成具有更多信息的概念特征，解决了真实的雷达数据中某些感知对象由于距离和遮挡而出现结构不完整的问题。

一般来讲，体素化的过程会导致点云细节信息的丢失，然而 Deng 等[48]提出了不同的看法：精确的点位置信息可能不是必须的。他们认为 Voxel-based 方法的瓶颈在于缺失体素上下文信息的聚合，并由此提出了 Voxel R-CNN，与其他大部分基于体素的方法不同的是，该方法为两阶段检测网络。在网络的细化阶段中，将 RoI 划分为网格，利用 PointNet++的结构将邻居体素的特征聚合到中心体素上，最终获得了较好的检测结果。

同样是聚焦于体素空间特征，Zheng 等[52]设计了一个 SSFA（Spatial-Semantic Feature Aggregation）模块，该模块对传统的 BEV 特征提取网络进行了改进，将浅层网络得到的低级空间特征与深层网络得到的高级语义特征利用注意力机制进行自适应地融合。Mao 等[54]对体素特征提取网络进行了改进，基于 Transformer[68]设计了子流形体素模块和稀疏体素模块，用来获取体素之间的远程关系，以解决 Voxel-based 方法接收域有限的问题。

基于体素的方法由于具有高效性，经常被用在其他模型中作为骨干网络来提取特征，有助于添加其他创新性结构。Zheng 等[53]提出的 SE-SSD 利用了知识蒸馏[69]的思想，通过 Teacher-SSD 生成软标签（Soft Target），Student-SSD 则同时利用软标签和硬标签（Hard Target）进行训练，大大提升了模型检测的性能。由于 Teacher-SSD 仅在训练过程中使用，因此没有在检测中引入额外的计算量，这使得 SE-SSD 兼具了性能与速度，其网络结构如图 4 所示。CenterPoint[70]使用 VoxelNet 作为体素特征提取器，把 2D 检测器 CenterNet[71]的思想迁移到了 3D 检测中，通过关键点检测器找到目标的中心点，并根据中心点特征回归出三维边界框，减小了目标检测器的搜索空间，且与点云对象的旋转不变特性相契合。

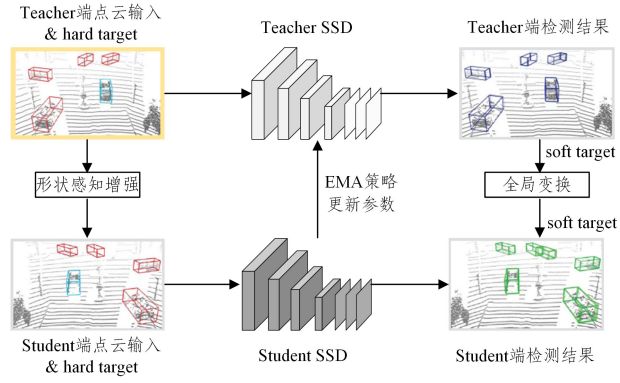


图 4 SE-SSD 网络结构图

Fig. 4 Structure of SE-SSD

总的来说，将点云表示为体素的形式，并使其规则化，便于运用卷积网络进行高效处理。因此，为了实现实时检测，目前大多数算法均采用基于体素的主体框架，主要的改进方向包括以下几个：1）扩大体素的接受域，如 Voxel R-CNN[48]和 VoTr[54]；2）对体素空间进行结构感知[16,47,49,52]；3）与其他经典结构相结合，如 SE-SSD[53]引入了知识蒸馏的思想，CenterPoint[70]将 center-based 结构[71]迁移至三维中。Voxel-based 方法由于具有高效性，受到很多学者的青睐，但这种方法仍存在一些问题，其中最重要的就是体素化引入的量化误差。因此，如何最大限度地消除量化误差的影响，仍将是未来的重点研究方向。

3.3 Point-based 方法

点云是一组点的坐标的集合，集合内部的所有点之间具有排列不变性，即无论数组中点的顺序如何排列，都不会改变点云本身所表示的物体；此外，点云还具有旋转不变性，当数组发生旋转时，点云的形状不会发生变化。这些特性使得研究者难以像处理图像、文字等其他规则序列一样直接处理点云。直到 2017 年，Qi 等[17]提出了 PointNet，使用最大池化作为对称

函数来解决点云的排列不变性，并利用一个小型的 T-Net 网络预测仿射变换矩阵，避免了点云几何变换带来的影响。之后，针对局部特征提取，Qi 等[18]又提出了 PointNet++，利用采样分组、集抽象的操作对点云的局部特征进行聚合，大大提升了网络的鲁棒性。PointNet 系列的提出，开创了 Point-based 方法的先河，促使更多基于点表示的三维目标检测算法出现，如 PointRCNN[19]，STD[27]，3DSSD[55]。

PointRCNN[19]是首个基于点的两阶段检测网络，其网络结构如图 5 所示。

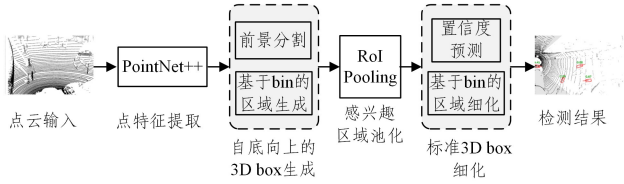


图 5 PointRCNN 网络结构图

Fig. 5 Structure of PointRCNN

它在第一阶段中使用多尺度 PointNet++网络[18]对点云进行前景分割，并基于 bin 为每个前景点生成初始预测；在第二阶段，通过坐标的标准变换进行提案的细化。Yang 等[27]通过一种球形锚框来生成提案，实现了高召回率。在第二阶段，提出了一种新的区域池化方式——PointsPool，类似于 VoxelNet[13]，PointsPool 将感兴趣区域划分为体素后，再基于 VFE 提取特征。由于点云中的点都是来自物体表面的点，不包含物体中心点，直接使用点云中的点生成提案会存在一定的偏差。因此，Qi 等[28]在深度网络中引入了霍夫投票机制，利用投票产生一组靠近物体中心的新点，然后再进行提案的预测生成。

基于点的方法大多采用 PointNet++来提取特征，而 PointNet++中的特征传播操作比较耗时，且此前的大多方法都属于两阶段架构，因此基于点的方法难以实现实时检测。针对此弊端，3DSSD[55]采用了单阶段架构，并删除了特征传播操作，最终的检测速度达到了 26FPS。3DSSD 中还提出了一种基于特征空间的最远点采样方法（F-FPS），用以增加采样点中前景点的比例，其网络结构如图 6 所示。Chen 等[72]还将语义信息引入 SA 阶段，提出了语义引导的最远点采样算法 S-FPS，使得网络能够更加关注信息丰富的前景点。Zhang 等[73]认为，在传统的随机采样及最远点采样中，前景点会逐渐丢失，因此他们设计了 IA-SSD，通过类感知采样和中心感知采样策略来保证采样到的点中前景点的概率最大，从而实现高召回率。IA-SSD 同样采用单阶段架构，在推理阶段其速度超越了 PointPillars，可达 83FPS，是至今最快的基于 LiDAR 的检测方法。

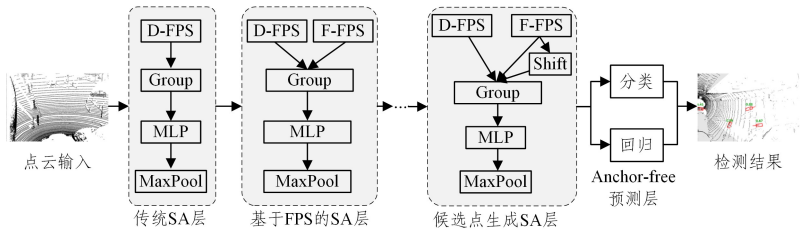


图 6 3DSSD 网络结构图

Fig. 6 Structure of 3DSSD

PointNet 系列结构对点特征的提取是通过 MLP 实现的，还有一些方法[56-57]将点云转换为图表示，使用 GNN[74]或 GCN（Graph Convolutional Network）[75]进行特征的提取。PointRGCN[57]是首个在三维目标检测领域引入 GCN 的方法，它的第一阶段沿用了 PointRCNN[19]，在第二阶段中使用图卷积进行提案内部和提案之间的特征提取与聚合。Point-GNN[56]是一种单阶段的检测网络，它直接在原始点云上构建图结构，使用图卷积提取顶点特征，并对每个顶点预测其边界框和类别。

总的来说，目前 Point-based 方法主要是基于 PointNet 系列进行特征提取的，能够达到比较高的检测精度，但在规模较大的自动驾驶场景中，由于点的数量十分庞大，其计算成本相对较高，难以达到实时检测的效果。

3.4 Point-Voxel-based 方法

随着 Voxel-based 和 Point-based 方法的发展，一些方法提出将点和体素结合起来[29,58-62,76]，以期同时利用点表示和体素表示的优势。

Chen 等[58]提出的 Fast Point R-CNN 是首个将点和体素结合起来进行检测的方法。它首先使用体素表示生成少量高质量的初始预测，接着在第二阶段使用注意力机制融合原始点的坐标信息和第一阶段的卷积特征，对初始预测框进行细化。PV-RCNN[29]同样是在第二阶段引入了点特征进行区域细化，但不同于 Fast Point R-CNN[58]中简单的特征拼接，它首先使用 FPS 在点云中采样得到少量关键点，接着使用 VSA（Voxel Set Abstraction）模块将多尺度体素特征聚合到少量关键点上，并与原始点的坐标特征进行融合，最后通过 RoI grid pooling 得到每个 RoI 的特征，如图 7 所示。PV-RCNN 取得了较优的检测性能，但由于特征聚合操作较为复杂，检测速度较慢。之后，Shi 等[59]又提出了 PV-RCNN 的改进版本 PV-RCNN++，对关键点采样和特征聚合操作进行了改进。对于关键点采样，首先将整个点云替换为 RoI 周围的点，以减少点的数量，接着将整个激光雷达场景划分为多个扇区，然后在这些扇区并行地进行 FPS 采样。对于特征聚合，PV-RCNN[29]采用了 PointNet++中的 SA 操作进行聚合，但 SA 中存在大量距离查询操作，时间复杂度较高，为此，Shi 等提出了向量聚合（VectorPool），通过在关键点附近构建体素邻域，快速聚合体素特征。改进后的 PV-RCNN++相较于 PV-RCNN 在速度上有了较大的提升。Song 等[62]提出的 JPV-Net 中使用了一种三线性 SA 模块，用来实现点到体素的投影，相比 PV-RCNN 中的 SA 操作，避免了多尺度分组操作，

同样节省了计算成本。Zhao 等[77]在 PV-RCNN 的基础上设计了一种 SegFPS 算法来进行关键点的采样，通过前景分割网络使关键点中保留更多的前景点。

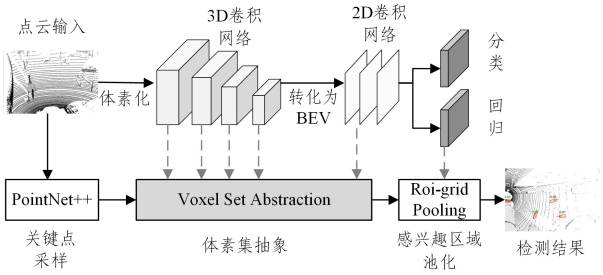


图 7 PV-RCNN 网络结构图

Fig. 7 Structure of PV-RCNN

与上述方法[29,58-59]不同，Li 等[60]在第一阶段就引入了原始点特征，使用 PointNet++网络提取点特征，并通过前背景分割网络得到每个点的评分，之后利用前景评分分别对点特征和局部体素特征加权聚合得到最终的体素特征。其中点和体素特征融合的策略是分别提取点和体素的特征，然后加以融合，不涉及点和体素的对应关系。而后，他们又提出了另一种融合策略[61]，即通过一组体素到点的解码器，将体素特征利用逆距离加权插值的方式得到与初始点云分辨率相同的点特征，然后与上一层解码器得到的点特征相加，得到该层解码器解码出的点特征，并利用最后一个解码器的点特征进行初始检测框的细化。

从上述方法来看，基于点和体素特征融合的检测结合了两者的优势，提升了检测性能。目前所提出的 Point-Voxel-based 方法均采用两阶段的架构，使用体素特征提取网络作为主干部分，在第一阶段生成初始检测框，差别仅在于融合点的方式和阶段不同。有的方法仅在特征层面进行融合，如 Fast Point R-CNN[58]和 P2V-RCNN[60]，其他方法[29,59,62,76]利用点和体素的空间信息进行对应聚合。然而，受限于两阶段网络的固有特性，这些方法大都比较耗时，因此，Point-Voxel-based 方法的研究难点在于如何以一种更加轻量化的网络结构实现点和体素特征的融合。

4 基于多传感器的三维目标检测

自动驾驶的安全离不开冗余传感系统。在实际场景中，自动驾驶车辆上一般会同时搭载激光雷达、雷达、摄像头等多个传感器。不同的传感器具有不同的特点，具体如表 1 所列。

表 1 不同传感器的优缺点对比

Table 1 Advantages and disadvantages of different sensors

传感器	相机	雷达	激光雷达
优点	纹理信息丰富；成本低	探测距离长；能够探测目标的径向速度；成本低；不受恶劣天气影响	能提供精确的三维空间信息；近距离探测分辨率高；噪声较少
缺点	缺乏深度信息；易受恶劣天气影响	探测精度和分辨率较激光雷达低；数据更为稀疏；噪声多	成本高；易受恶劣天气影响；探测距离相比雷达短

为了避免过于依赖单一传感器，一些方法致力于融合不同传感器的信息，实现优势互补。基于多传感器的三维目标检测方法最常使用的是相机和激光雷达这两种传感器，也有一些利用雷达作为图像或对激光雷达的补充信息加以融合。依据对点云数据处理方式的不同，可以将其分为多视图融合与多模态融合两类。

多视图融合方法通常将点云数据投影为 2D 视图，如鸟瞰图(BEV)、前视图(FV)，然后与 RGB 图像特征进行融合。MV3D[78]将点云分别投影到鸟瞰图和前视图上，然后在鸟瞰图上生成 3D 候选框，并将候选框分别投影到前视图和图像中，获得对应视图的 RoI 特征，最后对 3 种视图的 RoI 特征进行深度融合得到最终的 3D 边界框。不同于 MV3D 在 RoI 细化阶段中融合进其他视图特征，AVOD[79]在提案生成阶段就对不同视图(鸟瞰图和 RGB 图像)的特征进行了融合。此外，AVOD 还基于 FPN 改进了 RPN，以保证最后的特征图是全分辨率的，改善了 MV3D 上小目标检测效果不佳的问题。Liang 等[80]通过连续融合层将图像的多尺度特征融合到不同尺度的 BEV 特征中，创建出一个密集的 BEV 特征图。但当雷达点十分稀疏时，这种融合的作用十分有限。由此，MMF[81]引入了深度补全的辅助任务，来寻找 LiDAR 的鸟瞰图与图像之间的密集对应关系。此外，MMF 还引入了一个地面估计的辅助任务，这种多任务间的相互增益，使得网络能够学习到更优的特征表示。CenterFusion[31]使用视锥关联方法将雷达检测到的目标与 RGB 图像上检测到的目标关联起来，并将这些关联雷达检测的深度和速度特征与图像特征连接后用于二次回归的输入。Cui 等[82]将 4D 毫米波雷达转化为 BEV 和 FV，利用 RGB 图像和 BEV 生成提案，并分别投影到 3 个视图的特征图中进行交叉融合。基于多视图的融合方法本质上仍是基于二维数据的，这种降维再投影的操作消除了空间上的相关性，同时简单的高度压缩也使得点之间的拓扑关系丢失，导致检测精度降低。

基于多模态融合的方法同样以图像和点云为输入，不同于多视图方法，其对点云的处理是直接在原始点云上进行的。Frustum-PointNets[25]是第一个使用原始点云的多传感器检测方法，它使用级联的方式融合多个传感器，首先对图像进行检测得到 2D 边界框，然后将二维检测结果投影到三维空间中生成截锥体，最后使用 PointNet[17]回归边界框参数。F-ConvNet[83]使用平行的平面生成一组截锥，以获取每一个截锥内部的局部特征，最后通过全卷积网络进行上采样和下采样融合特征。这类方法可统称为基于视锥的方法，由于其将 2D 检测作为网络的第一阶段，使得最终的检测结果过于依赖所选择的 2D 检测器。

基于多视图融合的方法[78-79]在点云的 BEV 层面与图像进行融合，但投影到 BEV 的过程不可避免地会造成信息损失；而基于视锥的方法[25,83-84]过于依赖 2D 检测器的性能，且生成的视锥中会存在遮挡问题。因此，一些方法尝试在更细粒度的点或体素上进行多模态融合[85-88]，被称为早期融合。Xie 等[85]提出了一种基于点的注意连续卷积融合模块，该模块连接图像语义分割子网络和点特征提取子网络，直接将 3D 点和 2D 语义特征进行融合。PointPainting[86]同样将点云投影到图像上，从而将图像的语义信息融合到点上。PointAugmenting[88]进一步提出了一种跨模态的数据增强方法，其

不只对点云进行真值粘贴，还根据遮挡关系在图像上也粘贴真值，以保证 2D 和 3D 数据在数据增强中的一致性。然而这种将点云投影到图像上检索特征的方法[85-86]，图像和点云之间分辨率的差异可能会导致图像的采样率较低。因此，Zhu 等[87]提出使用一组虚拟点作为多模态特征的聚合点，其密度介于图像密度和点云密度之间，这种方法很好地弥合了不同模态数据之间分辨率的差距。由于早期融合对数据对齐很敏感，因此 Pang 等[89]提出了晚期融合的 CLOCs，对 2D 候选结果和 3D 候选结果进行融合，以最大限度地提取所有潜在的正确结果。

多模态融合的方法结合图像和点云进行检测，以实现优势互补，但总的来说，由于不同模态数据之间视角、维度的差异，目前的研究还不够成熟。因此，如何设计一种融合方法，充分地利用二者的优势，消除或进一步减弱跨模态数据之间的融合鸿沟，仍值得研究人员进一步思考。

5 三维目标检测常用数据集

三维目标检测的快速发展离不开数据集的支撑，在自动驾驶领域使用较为广泛的公开数据集有 3 个：KITTI 数据集[90-91]、nuScenes 数据集[92]和 Waymo Open Dataset[93]。接下来，我们将对这 3 个数据集展开介绍，并对不同算法在这些数据集上的性能进行对比。

5.1 KITTI 数据集

KITTI 数据集[90-91]是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办的，可用于多种自动驾驶场景下的视觉算法的评测，如立体图像评测、光流、深度估计、二维/三维目标检测、目标追踪等，是目前使用最广泛的一个公开数据集。KITTI 数据集使用 Velodyne 64 线 3D 激光雷达完成点云数据的采集，还使用摄像机同步进行对应图像的采集。KITTI 数据集包含 7481 个训练样本和 7518 个测试样本，其主要的检测目标有 3 类：Car，Pedestrian 和 Cyclist。并依据遮挡、截断等情况，按难度将目标分为 Easy，Moderate 和 Hard 3 个等级。由于 KITTI 的测试集无法获取，大多数研究将训练集划分为训练和验证两个集合，分别包含 3712 和 3769 个样本。

与二维目标检测相同，平均准确率（mean Average Precision，mAP）是三维目标检测任务的主要评估指标。准确率是将预测框与真值框的交并比（Intersection over Union，IoU）和某阈值进行比较得来的，若 IoU 大于阈值，则视为真阳性（True Positive，TP），否则为假阳性（False Positive，FP）。据此，得到召回率 r 和精度 p，如式（1）所示：

$$r=\frac{TP}{TP+FN}$$
$$p=\frac{TP}{TP+FP} \tag{1}$$

其中，FN 表示假阴性（False Negative）。通常使用插值的方法计算 AP，如式（2）所示。其中 $\mathbb{R}$ 表示预定义的召回位置，在 KITTI 测试集中通常使用 $\mathbb{R}=40$。

$$AP=\frac{1}{|\mathbb{R}|}\sum_{r\in\mathbb{R}}\max_{r':r'\geqslant r}p(r') \tag{2}$$

KITTI 数据集为不同的类别设置了不同的 IoU 阈值，其中 Car 类别的阈值为 0.7，Cyclist 和 Pedestrian 类别的阈值为 0.5。KITTI 官网设立了三维目标检测任务的排行榜，并按 Car 类别中等难度的准确率进行排序。表 2 列出了一些代表性算法在 KITTI 测试集上的三维目标检测性能。

表 2 不同算法在 KITTI 测试集上的检测性能

Table 2 Detection performance of different algorithms on KITTI test set

方法	来源	类型	Car-3D			Car-BEV		
			Easy	Moderate	Hard	Easy	Moderate	Hard
MV3D[78]	CVPR(2017)	多传感器	74.97	63.63	54.00	86.62	78.93	69.80
AVOD[79]	IROS(2018)	多传感器	83.07	71.76	65.73	89.75	84.95	78.32
F-PointNets[25]	CVPR(2018)	多传感器	82.19	69.79	60.59	91.17	84.67	74.77
VoxelNet[13]	CVPR(2018)	Voxel-based	77.47	65.11	57.73	87.95	78.39	71.29
SECOND[14]	Sensors(2018)	Voxel-based	83.13	73.66	66.20	88.07	79.37	77.95
PointRCNN[19]	CVPR(2019)	Point-based	86.96	75.64	70.70	92.13	87.39	82.72
PointPillars[45]	CVPR(2019)	View-based	82.58	74.31	68.99	90.07	86.56	82.81
Fast Point R-CNN[58]	ICCV(2019)	Point-Voxel-based	85.29	77.40	70.24	90.87	87.84	80.52
STD[27]	ICCV(2019)	Point-based	87.95	79.71	75.09	94.74	89.19	86.42
PI-RCNN[85]	AAAI(2020)	多传感器	84.37	74.82	70.03	—	—	—
3DSSD[55]	CVPR(2020)	Point-based	88.36	79.57	74.55	92.66	89.02	85.86
PV-RCNN[29]	CVPR(2020)	Point-Voxel-based	90.25	81.43	76.82	**94.98**	**90.65**	**86.14**
Part-A^2[47]	TPAMI(2021)	Voxel-based	85.94	77.86	72.00	89.52	84.76	81.47
Voxel R-CNN[48]	AAAI(2021)	Voxel-based	**90.90**	81.62	**77.06**	—	—	—
CIA-SSD[52]	AAAI(2021)	Voxel-based	89.59	80.28	72.87	—	—	—
SE-SSD[53]	CVPR(2021)	Voxel-based	**91.49**	**82.54**	**77.15**	**95.68**	**91.84**	**86.72**
JPV-Net[62]	AAAI(2022)	Point-Voxel-based	88.66	**81.73**	76.94	—	—	—
IA-SSD[73]	CVPR(2022)	Point-based	88.87	80.32	75.10	—	—	—

5.2 nuScenes 数据集

nuScenes 数据集[92]由在波士顿和新加坡拍摄的 1000 个驾驶场景组成，涵盖了不同地点、不同时间和不同天气的情况。每个场景的时长为 20 s，其中标注的三维边界框分为 23 个类，带有 8 个属性的标记值。nuScenes 数据集比 KITTI 数据集大得多，其标注量是 KITTI 的 7 倍，且 nuScenes 数据集是这 3 个数据集中唯一一个包含了三维 Radar 数据的数据集。此外，KITTI 数据集多是在白天和良好天气条件下拍摄

的，而 nuScenes 数据集还包含了夜晚及雨天等场景。

在三维目标检测任务中，nuScenes 数据集采用的评估指标包括 mAP，NDS（nuScenes Detection Score）和 PKL（Planning KL-Divergence）。mAP 指标中使用鸟瞰图的中心距离 d 代替三维包围框的交并比进行阈值匹配，d 的取值范围为 $\{0.5,1,2,4\}$m，mAP 的计算如式(3)所示：

$$mAP=\frac{1}{|\mathbb{C}||\mathbb{D}|}\sum_{c\in\mathbb{C}}\sum_{d\in\boldsymbol{D}}AP_{c,d} \tag{3}$$

其中，$\mathbb{C}$ 表示类别集合，$\mathbb{D}=\{0.5,1,2,4\}$。

由于 mAP 中仅考虑了包围框的位置信息，不包括尺寸和方向，因此，nuScenes 还设计了一系列的 TP 指标，分别对预测的三维包围框的平移、尺度、方向、速度和属性进行评估，具体包括 ATE（Average Translation Error），ASE（Average Scale Error），AOE（Average Orientation Error），AVE（Average Velocity Error）和 AAE（Average Attribute Error）。每个 TP 指标的所有类别的平均 TP 值 mTP 计算式如式(4)所示：

$$mTP=\frac{1}{|\mathbb{C}|}\sum_{c\in\mathbb{C}}TP_c \tag{4}$$

NDS 将 5 个 TP 指标集成到一个标量中，如式(5)所示：

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

其中，$\mathbb{TP}$ 是 5 个 TP 指标的集合。

最近，nuScenes 还增加了一个神经规划指标 PKL[94-95]，用于对检测器的感知性能进行评估。它测量了规划器在接收到检测结果时与看到现实对象时的规划差异，并以此检验 3D 目标检测对下游自动驾驶任务的影响。PKL 是非负值，其值越大，代表检测性能越差。

nuScenes 数据集中需检测的类别数为 10 类，表 3 列出了一些代表性算法在 nuScenes 测试集上的三维目标检测性能。

表 3 不同算法在 nuScenes 测试集上的检测性能

Table 3 Detection performance of different algorithms on nuScenes test set

方法	来源	类型	$mAP\uparrow$	$NDS\uparrow$	$PKL\downarrow$
3DSSD[55]	CVPR(2020)	Point-based	42.6	56.4	—
PointPainting[86]	CVPR(2020)	多传感器	46.4	58.1	0.89
PointAugmenting[88]	CVPR(2021)	多传感器	**66.8**	**71.0**	**0.59**
CenterPoint[70]	CVPR(2021)	Voxel-based	58.0	65.5	0.69

5.3 Waymo Open Dataset

Waymo Open Dataset[93]是目前最大的自动驾驶数据集，使用 5 个激光雷达传感器和 5 个高分辨率针孔相机收集数据，包含 798 个训练场景、202 个验证场景、150 个测试场景，每个场景的时长为 20 s。Waymo Open Dataset 的注释频率比 nuScenes 高 5 倍，共有 2500 万个 3D 标签和 2200 万个 2D 标签。

在三维目标检测任务中，Waymo Open Dataset 使用的评价指标除了 AP 外，增加了一个包含方向角信息的 APH，其计算式如式(6)所示：

$$\begin{aligned}AP&=100\int_0^1\max\{p(r')\mid r'\geqslant r\}\mathrm{d}r\\APH&=100\int_0^1\max\{h(r')\mid r'\geqslant r\}\mathrm{d}r\end{aligned} \tag{6}$$

其中，$p(r)$为 P-R 曲线，$h(r)$与 $p(r)$类似，但其每个 TP 值都由方向角信息 $\frac{\min(|\hat{\theta}-\theta|,2\pi-|\hat{\theta}-\theta|)}{\pi}$ 进行加权，$\hat{\theta}$ 和 θ 分别表示预测方向角和方向角真值。

与 KITTI 类似，Waymo 依据 3D 边界框中包含的激光雷达点数划分了两个难度级别：LEVEL_1 中标注物体的雷达点数不少于 5 个，LEVEL_2 中标注物体的雷达点数少于 5 个。表 4 列出了一些代表性算法在 Waymo Open Dataset 的 202 个验证场景上的三维目标检测性能。

表 4 不同算法在 Waymo 验证集上的车辆检测性能

Table 4 Detection performance of different algorithms on Waymo validation set for vehicle detection

方法	来源	类型	LEVEL_1 mAP/mAPH	LEVEL_2 mAP/mAPH
PV-RCNN[29]	CVPR(2020)	Point-Voxel-based	70.30/69.69	65.36/64.79
PV-RCNN++[59]	arXiv(2021)	Point-Voxel-based	**78.79/78.21**	**70.26/69.71**
Voxel R-CNN[48]	AAAI(2021)	Voxel-based	75.59/—	66.59/—
PointAugmenting[88]	CVPR(2021)	多传感器	67.41/—	62.70/—
VoTr-TSD[54]	ICCV(2021)	Voxel-based	74.95/74.25	65.91/65.29
V2P-RCNN[61]	ACM MM(2021)	Point-Voxel-based	77.24/—	69.77/—
IA-SSD[73]	CVPR(2022)	Point-based	70.53/69.67	61.55/60.80

6 未来发展方向

随着无人驾驶逐渐进入实际应用阶段，准确而快速的目标检测是其感知系统中不可或缺的一部分。相比成熟的二维目标检测，三维目标检测技术仍存在一定的局限性，但也具有广阔的发展空间。结合上文所述内容，对三维目标检测未来的发展方向展望如下。

(1)二维到三维的深度估计。从整体上来说，基于图像的三维目标检测方法性能不如基于 LiDAR 的方法，其关键在于二维向三维转换时的深度估计误差较大。因此，提升深度估计网络的准确性对基于图像的检测方法性能的提升具有重要意义；另一方面，基于图像的检测方法的性能提升也能够为后续的多模态融合奠定基础。

(2)基于 LiDAR 的小目标检测。目前的算法大多能够识别出距离较近、具有稠密点云的目标，但对于一些远处的、包含雷达点极少的小目标仍然无法正确检测。点云特征提取过程通常会涉及降采样操作，导致部分点丢失，对小目标来说尤其如此，因此基于点云的小目标检测问题将是一个重要的研究方向。

(3)检测速度与精度的平衡。现有的三维目标检测网络可以分为单阶段和两阶段两种架构，单阶段架构能够实现较快的速度，但其精度大多不及两阶段架构；两阶段架构由于增加了候选框细化的步骤，在检测速度上稍显逊色。因此如何提高单阶段架构的检测精度或是两阶段架构的检测速度，

实现速度与精度的平衡将是未来三维目标检测的重点所在。

(4)4D毫米波雷达研究前景广阔。传统的毫米波雷达仅能探测物体的距离信息，无法感知物体的高度，而4D毫米波雷达则弥补了这一缺陷。4D毫米波雷达能够适应各种恶劣环境，且成本较激光雷达低很多，这使得其有望替代激光雷达。目前国内外针对4D毫米波雷达的研究还较少，检测算法也多是直接对激光雷达算法的改进[32-33]，或作为补充信息与其他模态数据融合[30-31,82]，较少针对4D毫米波雷达特性开展算法研究。2022年6月KAIST最新发布了一个包含4D雷达数据的大规模数据集K-Radar[96]，弥补了4D毫米波雷达领域数据集稀少这一短板。大规模数据集的发布也必将催生出大量相关研究，基于4D毫米波雷达的三维目标检测研究前景广阔。

(5)多模态数据融合。在自动驾驶中的冗余传感系统的支持下，多模态融合将是一种不可避免的趋势。然而事实上，由于不同传感器数据之间分辨率、语义的差异，目前的多模态融合方法还远不及基于点云的单模态方法。因此，如何有效地消除不同模态数据之间的鸿沟，达到更好的融合效果将是三维目标检测未来发展的一大挑战。

结束语 本文总结了自动驾驶场景下三维目标检测领域的最新进展，将现有的方法按传感器类型分为3类，即基于图像的三维目标检测、基于LiDAR的三维目标检测和基于多传感器的三维目标检测，并着重对基于LiDAR的三维目标检测方法进行了综述。本文还介绍了3种常用的自动驾驶数据集及其评价指标，比较了一些最新的具有代表性的算法在不同数据集上的检测效果。最后，本文对三维目标检测技术存在的问题和挑战进行了分析，并对其未来发展方向进行了展望。

参 考 文 献

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

[2] REDMON J, DIVVALA S, GIRSHICK R, et al. You Only Look Once: Unified, Real-time Object Detection[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016:779-788.

[3] LIU W, ANGUELOV D, ERHAN D, et al. SSD: Single Shot Multibox Detector[C]// Proceedings of the European Conference on Computer Vision. Amsterdam: Springer, 2016:21-37.

[4] GIRSHICK R, DONAHUE J, DARRELL T, et al. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Columbus: IEEE, 2014: 580-587.

[5] ZHANG P, SONG Y F, ZONG L B, et al. Advances in 3D Object Detection: a Brief survey[J]. Computer Science, 2020, 47(4): 94-102.

[6] SONG S, XIAO J. Sliding Shapes for 3 D Object Detection in Depth Images[C]// Proceedings of the European Conference on Computer Vision. Zurich: Springer, 2014:634-651.

[7] CHEN X, KUNDU K, ZHU Y, et al. 3D Object Proposals for Accurate Object Class Detection[J]. Advances in Neural Information Processing Systems, 2015, 28: 424-432.

[8] CHEN X, KUNDU K, ZHANG Z, et al. Monocular 3D Object Detection for Autonomous Driving[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016: 2147-2156.

[9] SONG S, XIAO J. Deep Sliding Shapes for Amodal 3D Object Detection in RGB-D Images[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016: 808-816.

[10] CHEN X, KUNDU K, ZHU Y, et al. 3D Object Proposals Using Stereo Imagery for Accurate Object Class Detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 40(5): 1259-1272.

[11] WANG D Z, POSNER I. Voting for Voting in Online Point Cloud Object Detection[C]// Robotics: Science and Systems. 2015: 10-15.

[12] ENGELCKE M, RAO D, WANG D Z, et al. Vote3deep: Fast Object Detection in 3D Point Clouds Using Efficient Convolutional Neural Networks[C]// Proceedings of the IEEE International Conference on Robotics and Automation. Singapore: IEEE, 2017: 1355-1361.

[13] ZHOU Y, TUZEL O. VoxelNet: End-to-End Learning for Point Cloud Based 3D Object Detection[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 4490-4499.

[14] YAN Y, MAO Y, LI B. SECOND: Sparsely Embedded Convolutional Detection[J]. Sensors, 2018, 18(10): 3337-3353.

[15] LI B. 3D Fully Convolutional Network for Vehicle Detection in Point Cloud[C]// Proceedings of the IEEE International Conference on Intelligent Robots and Systems. Vancouver: IEEE, 2017: 1513-1518.

[16] KUANG H, WANG B, AN J, et al. Voxel-FPN: Multi-scale Voxel Feature Aggregation for 3D Object Detection from LiDAR Point Clouds[J]. Sensors, 2020, 20(3): 704.

[17] QI C R, SU H, MO K, et al. PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017: 652-660.

[18] QI C R, YI L, SU H, et al. PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space[J]. Advances in Neural Information Processing Systems, 2017, 30: 5099-5108.

[19] SHI S, WANG X, LI H. PointRCNN: 3D Object Proposal Generation and Detection from Point Cloud[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 770-779.

[20] ZHAO J H, DOU X T, CAO Y E, et al. A Method for 3D Point Cloud Classification Based on Segmentation Results[J]. Science of Surveying and Mapping, 2022, 47(3): 85-95.

[21] DU Z J, CAO F L, YE H L, et al. 3D Point Cloud Classification Algorithm Based on Residual Edge Convolution[J]. Pattern Recognition and Artificial Intelligence, 2021, 34(9): 836-843.

[22] YANG X W, WANG A B, HAN X, et al. Point Cloud Semantic

Segmentation Based on KNN-PointNet[J]. Laser & Optoelectronics Progress, 2021, 58(24): 272-279.

[23] AOKI Y, GOFORTH H, SRIVATSAN R A, et al. PointNetlk: Robust & Efficient Point Cloud Registration Using PointNet [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 7163-7172.

[24] GE L, REN Z, YUAN J. Point-to-Point Regression PointNet for 3D Hand Pose Estimation[C]// Proceedings of the European Conference on Computer Vision. Munich: Springer, 2018: 475-491.

[25] QI C R, LIU W, WU C, et al. Frustum PointNets for 3D Object Detection from RGB-D Data[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 918-927.

[26] YANG Z, SUN Y, LIU S, et al. IPOD: Intensive Point-based Object Detector for Point Cloud[J]. arXiv: 1812. 05276, 2018.

[27] YANG Z, SUN Y, LIU S, et al. STD: Sparse-to-Dense 3D Object Detector for Point Cloud[C]// Proceedings of the IEEE International Conference on Computer Vision. Seoul: IEEE, 2019: 1951-1960.

[28] QI C R, LITANY O, HE K, et al. Deep Hough Voting for 3D Object Detection in Point Clouds[C]// Proceedings of the IEEE International Conference on Computer Vision. Seoul: IEEE, 2019: 9277-9286.

[29] SHI S, GUO C, JIANG L, et al. PV-RCNN: Point-Voxel Feature Set Abstraction for 3D Object Detection[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020: 10529-10538.

[30] YANG B, GUO R, LIANG M, et al. RadarNet: Exploiting Radar for Robust Perception of Dynamic Objects[C]// Proceedings of the European Conference on Computer Vision. Springer, 2020: 496-512.

[31] NABATI R, QI H. CenterFusion: Center-based Radar and Camera Fusion for 3D Object Detection[C]// Proceedings of the IEEE Winter Conference on Applications of Computer Vision. IEEE, 2021, 1526-1535.

[32] XU B, ZHANG X, WANG L, et al. RPFA-Net: a 4D RaDAR Pillar Feature Attention Network for 3D Object Detection[C]// Proceedings of the IEEE International Intelligent Transportation Systems Conference. Indianapolis: IEEE, 2021: 3061-3066.

[33] LI X L. Research on Key Technologies of Object Detection for Vehicular 4D Millimeter Wave Radar[D]. Nanjing: Nanjing University Of Science And Technology, 2021.

[34] MEYER M, KUSCHK G. Automotive Radar Dataset for Deep Learning Based 3D Object Detection[C]// Proceedings of the European Radar Conference. Paris: IEEE, 2019: 129-132.

[35] QIAN R, LAI X, LI X. 3D Object Detection for Autonomous Driving: a Survey[J]. Pattern Recognition, 2022, 130: 108796.

[36] REN S, HE K, GIRSHICK R, et al. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks [J]. Advances in Neural Information Processing Systems, 2015, 28: 91-99.

[37] SONG S, CHANDRAKER M. Joint SFM and Detection Cues for Monocular 3D Localization in Road Scenes[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston: IEEE, 2015: 3734-3742.

[38] 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]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 8445-8453.

[39] YOU Y, WANG Y, CHAO W L, et al. Pseudo-LiDAR++: Accurate Depth for 3D Object Detection in Autonomous Driving [J]. arXiv: 1906. 06310, 2019.

[40] MA X, WANG Z, LI H, et al. Accurate Monocular 3D Object Detection via Color-Embedded 3D Reconstruction for Autonomous Driving[C]// Proceedings of the IEEE International Conference on Computer Vision. Seoul: IEEE, 2019: 6851-6860.

[41] QIAN R, GARG D, WANG Y, et al. End-to-End Pseudo-LiDAR for Image-based 3D Object Detection[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020: 5881-5890.

[42] YANG B, LUO W, URTASUN R. PIXOR: Real-Time 3D Object Detection from Point Clouds[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 7652-7660.

[43] SIMON M, MILZ S, AMENDE K, et al. Complex-YOLO: An Euler-Region-Proposal for Real-Time 3D Object Detection on Point Clouds[C]// Proceedings of the European Conference on Computer Vision Workshops. Munich: Springer, 2018: 197-209.

[44] ALI W, ABDELKARIM S, ZIDAN M, et al. Yolo3D: End-to-End Real-Time 3D Oriented Object Bounding Box Detection from LiDAR Point Cloud[C]// Proceedings of the European Conference on Computer Vision Workshops. Munich: Springer, 2018: 716-728.

[45] LANG A H, VORA S, CAESAR H, et al. PointPillars: Fast Encoders for Object Detection from Point Clouds[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 12697-12705.

[46] LUO W, YANG B, URTASUN R. Fast and Furious: Real Time End-to-End 3D Detection, Tracking and Motion Forecasting with a Single Convolutional Net[C]// Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 3569-3577.

[47] 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.

[48] DENG J, SHI S, LI P, et al. Voxel R-CNN: Towards High Performance Voxel-based 3D Object Detection[C]// Proceedings of the AAAI Conference on Artificial Intelligence. AAAI, 2021: 1201-1209.

[49] HE C, ZENG H, HUANG J, et al. Structure Aware Single-stage 3D Object Detection from Point Cloud[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020: 11873-11882.

[50] YE M，XU S，CAO T. HVNet：Hybrid Voxel Network for LiDAR based 3D Object Detection[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Seattle：IEEE，2020：1631-1640.

[51] DU L，YE X，TAN X，et al. Associate-3Ddet：Perceptual-to-Conceptual Association for 3D Point Cloud Object Detection[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Seattle：IEEE，2020：13329-13338.

[52] ZHENG W，TANG W，CHEN S，et al. CIA-SSD：Confident IoU-aware Single-Stage Object Detector from Point Cloud[C]// Proceedings of the AAAI Conference on Artificial Intelligence. AAAI，2021：3555-3562.

[53] ZHENG W，TANG W，JIANG L，et al. SE-SSD：Self-Ensembling Single-Stage Object Detector from Point Cloud[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE，2021：14494-14503.

[54] MAO J，XUE Y，NIU M，et al. Voxel Transformer for 3D Object Detection[C]// Proceedings of the IEEE International Conference on Computer Vision. Montreal：IEEE，2021：3164-3173.

[55] YANG Z，SUN Y，LIU S，et al. 3DSSD：Point-based 3D Single Stage Object Detector[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Seattle：IEEE，2020：11040-11048.

[56] SHI W，RAJKUMAR R. Point-GNN：Graph Neural Network for 3D Object Detection in a Point Cloud[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Seattle：IEEE，2020：1711-1719.

[57] ZARZAR J，GIANCOLA S，GHANEM B. PointRGCN：Graph Convolution Networks for 3D Vehicles Detection Refinement [J]. arXiv：1911.12236，2019.

[58] CHEN Y，LIU S，SHEN X，et al. Fast Point R-CNN[C]// Proceedings of the IEEE International Conference on Computer Vision. Seoul：IEEE，2019：9775-9784.

[59] SHI S，JIANG L，DENG J，et al. PV-RCNN++：Point-Voxel Feature Set Abstraction with Local Vector Representation for 3D Object Detection[J]. arXiv：2102.00463，2021.

[60] LI J，SUN Y，LUO S，et al. P2V-RCNN：Point to Voxel Feature Learning for 3D Object Detection from Point Clouds[J]. IEEE Access，2021，9：98249-98260.

[61] LI J，DAI H，SHAO L，et al. From Voxel to Point：IoU-Guided 3D Object Detection for Point Cloud with Voxel-to-Point Decoder[C]// Proceedings of the 29th ACM International Conference on Multimedia. Chengdu：ACM，2021：4622-4631.

[62] SONG N，JIANG T，YAO J. JPV-Net：Joint Point-Voxel Representations for Accurate 3D Object Detection[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2022：2271-2279.

[63] REDMON J，FARHADI A. YOLO9000：Better，Faster，Stronger [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu：IEEE，2017：7263-7271.

[64] SHEN Q，CHEN Y L，LIU S，et al. 3D Object Detection Algorithm based on Two-Stage Network[J]. Computer Science，2020，47(10)：145-150.

[65] GRAHAM B. Spatially-Sparse Convolutional Neural Networks [J]. arXiv：1409.6070，2014.

[66] GRAHAM B. Sparse 3D Convolutional Neural Networks[J]. arXiv：1505.02890，2015.

[67] GRAHAM B，VAN DER MAATEN L. Submanifold Sparse Convolutional Networks[J]. arXiv：1706.01307，2017.

[68] VASWANI A，SHAZEER N，PARMAR N，et al. Attention is All You Need[J]. Advances in Neural Information Processing Systems，2017，30：5998-6008.

[69] HINTON G，VINYALS O，DEAN J，et al. Distilling the Knowledge in a Neural Network[J]. arXiv：1503.02531，2015.

[70] YIN T，ZHOU X，KRAHENBUHL P. Center-based 3D Object Detection and Tracking[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE，2021：11784-11793.

[71] ZHOU X，WANG D，KRÄHENBÜHL P. Objects as Points[J]. arXiv：1904.07850，2019.

[72] CHEN C，CHEN Z，ZHANG J，et al. SASA：Semantics-Augmented Set Abstraction for Point-based 3D Object Detection [C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2022：11873-11882.

[73] ZHANG Y，HU Q，XU G，et al. Not All Points Are Equal：Learning Highly Efficient Point-based Detectors for 3D LiDAR Point Clouds[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. New Orleans：IEEE，2022：18953-18962.

[74] SCARSELLI F，GORI M，TSOI A C，et al. The Graph Neural Network Model[J]. IEEE Transactions on Neural Networks，2008，20(1)：61-80.

[75] KIPF T N，WELLING M. Semi-Supervised Classification with Graph Convolutional Networks[J]. arXiv：1609.02907，2016.

[76] BHATTACHARYYA P，CZARNECKI K. Deformable PV-RCNN：Improving 3D Object Detection with Learned Deformations [J]. arXiv：2008.08766，2020.

[77] ZHAO L，HU J，AN Y P，et al. Deep Learning Based on Semantic Segmentation for Three-Dimensional Object Detection from Point Clouds[J]. Chinese Journal of Lasers，2021，48(17)：177-189.

[78] CHEN X，MA H，WAN J，et al. Multi-View 3D Object Detection Network for Autonomous Driving[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu：IEEE，2017：1907-1915.

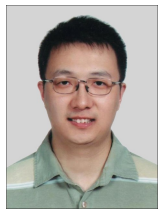
[79] KU J，MOZIFIAN M，LEE J，et al. Joint 3D Proposal Generation and Object Detection from View Aggregation[C]// Proceedings of the IEEE International Conference on Intelligent Robots and Systems. Madrid：IEEE，2018：1-8.

[80] LIANG M，YANG B，WANG S，et al. Deep Continuous Fusion for Multi-Sensor 3D Object Detection[C]// Proceedings of the European Conference on Computer Vision. Munich：Springer，2018：641-656.

[81] LIANG M，YANG B，CHEN Y，et al. Multi-Task Multi-Sensor Fusion for 3D Object Detection[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long

Beach: IEEE, 2019: 7345-7353.
[82] CUI H, WU J, ZHANG J, et al. 3D Detection and Tracking for On-road Vehicles with a Monovision Camera and Dual Low-cost 4D mmWave Radars[C]// Proceedings of the IEEE International Intelligent Transportation Systems Conference. Indianapolis: IEEE, 2021: 2931-2937.
[83] WANG Z, JIA K. Frustum Convnet: Sliding Frustums to Aggregate Local Point-wise Features for Amodal 3D Object Detection [C]// Proceedings of the IEEE International Conference on Intelligent Robots and Systems. Macau: IEEE, 2019: 1742-1749.
[84] LIU X H. 3D Object Detection Research Based on Image and Point Cloud Fusion[D]. Shanghai: Donghua University, 2020.
[85] XIE L, XIANG C, YU Z, et al. PI-RCNN: An Efficient Multi-Sensor 3D Object Detector with Point-based Attentive Cont-Conv Fusion Module[C]// Proceedings of the AAAI Conference on Artificial Intelligence. New York: AAAI, 2020: 12460-12467.
[86] VORA S, LANG A H, HELOU B, et al. PointPainting: Sequential Fusion for 3D Object Detection [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020: 4604-4612.
[87] ZHU H, DENG J, ZHANG Y, et al. VPFNet: Improving 3D Object Detection with Virtual Point based LiDAR and Stereo Data Fusion[J]. arXiv: 2111. 14382, 2021.
[88] WANG C, MA C, ZHU M, et al. PointAugmenting: Cross-Modal Augmentation for 3D Object Detection[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2021: 11794-11803.
[89] PANG S, MORRIS D, RADHA H. CLOCs: Camera-LiDAR Object Candidates Fusion for 3D Object Detection [C]// Proceedings of the IEEE International Conference on Intelligent Robots and Systems. Las Vegas: IEEE, 2020: 10386-10393.
[90] GEIGER A, LENZ P, URTASUN R. Are We Ready for Autonomous Driving? The KITTI Vision Benchmark Suite[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Providence: IEEE, 2012: 3354-3361.
[91] GEIGER A, LENZ P, STILLER C, et al. Vision Meets Robotics: The KITTI Dataset[J]. The International Journal of Robotics Research, 2013, 32(11): 1231-1237.
[92] CAESAR H, BANKITI V, LANG A H, et al. nuScenes: A Multimodal Dataset for Autonomous Driving[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020: 11621-11631.
[93] SUN P, KRETZSCHMAR H, DOTIWALLA X, et al. Scalability in Perception for Autonomous Driving: Waymo Open Dataset [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020: 2446-2454.
[94] PHILION J, KAR A, FIDLER S. Learning to Evaluate Perception Models Using Planner-Centric Metrics[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020: 14055-14064.
[95] GUO Y, CAESAR H, BEIJBOM O, et al. The Efficacy of Neural Planning Metrics: a Meta-Analysis of PKL on nuScenes[J]. arXiv: 2010. 09350, 2020.
[96] PAEK D, KONG S, WIJAYA, K T. K-Radar: 4D Radar Object Detection Dataset and Benchmark for Autonomous Driving in Various Weather Conditions[J]. arXiv: 2206. 08171, 2022.

HUO Weile, born in 1998, postgraduate. Her main research interests include artificial intelligence and virtual reality.



REN Shuang, born in 1981, Ph.D, associate professor, Ph. D supervisor, is a member of China Computer Federation. His main research interests include artificial intelligence and virtual reality.

（责任编辑：何杨）