自动驾驶中的3D目标检测研究进展

陈 建[1,2],苏思教[1],黄立勤[1],赵铁松[1,2*]
(1. 福州大学物理与信息工程学院,福建福州 350108;
2. 媒体信息智能处理与无线传输福建省重点实验室,福建福州 350108)

摘 要: 近年来,自动驾驶因其在提升道路安全、提高交通效率等方面展现出巨大的潜力而受到越来越多的关注. 在现代自动驾驶系统中,感知系统扮演着至关重要的角色,其目标是准确地估计周围环境的状态,并为预测和规划提供可靠的观测信息. 其中,3D目标检测作为感知系统的重要组成部分,旨在预测自动驾驶车辆周围物体的位置、大小和类别. 本文归纳了近年来自动驾驶领域中3D目标检测的研究进展,从单模态检测和多模态融合检测两个角度出发,介绍了使用不同传感器进行单模态方法和多模态融合方法的优势和不足. 此外,本文还对比了各种代表性算法在公共数据集上的性能,总结了当前常用训练策略,并讨论了该领域未来的发展趋势.

关键词: 自动驾驶;3D目标检测;单模态;多模态融合
基金项目: 国家自然科学基金(No.62171134,No.62271149);福州市科技项目(No.2023-P-001)
中图分类号: TP391.4 **文献标识码:** A **文章编号:** 0372-2112(2025)06-2131-26
电子学报URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20250043

Research Advances on 3D Object Detection in Autonomous Driving

CHEN Jian[1,2], SU Si-jiao[1], HUANG Li-qin[1], ZHAO Tie-song[1,2*]
(1. *College of Physics and Information Engineering, Fuzhou University, Fuzhou, Fujian* 350108, *China*;
2. *Fujian Key Laboratory for Intelligent Processing and Wireless Transmission of Media Information, Fuzhou, Fujian* 350108, *China*)

Abstract: In recent years, autonomous driving has gained increasing attention due to its significant potential in improving road safety and enhancing traffic efficiency. The perception system plays a crucial role in modern autonomous driving systems, aiming to accurately estimate the surrounding environment's state and provide reliable observations for prediction and planning. Among them, 3D object detection serves as an important component of the perception system for predicting the positions, sizes, and categories of objects surrounding the autonomous vehicle. This paper provides a comprehensive overview of the research advancements in 3D object detection for autonomous driving in recent years. It discusses the advantages and limitations of single-modal methods and multi-modal fusion methods using different sensors from the perspectives of single-modal detection and multi-modal fusion detection. Furthermore, the paper compares the performance of various representative algorithms on public datasets, summarizes the current commonly used training strategies, and discusses the future development directions in this field.

Key words: autonomous driving; 3D object detection; single-modal; multi-modal fusion
Foundation Item(s): National Natural Science Foundation of China (No.62171134, No.62271149); Science and Technology Projects of Fuzhou City (No.2023-P-001)

1 引言

自动驾驶技术在全球范围内正受到越来越广泛的关注,它在提升道路安全、提高交通效率、优化城市规划等方面展现出了巨大潜力,有望从根本上改变交通部门的运作方式. 人类在驾驶时利用视觉和听觉系统来感知真实的世界,那么自动驾驶汽车呢? 如果它们像人类一样开车,那么它们需要通过传感器来感知周围环境中的其他交通参与者和障碍物,识别道路上检测到的目标来规划行进路线[1]. 传感器就像汽车的眼睛,正是由于车载传感器能够精确地判断障碍物的位置和运动信息,才能为车辆预知危险、紧急制动及防撞

收稿日期:2025-01-10;修回日期:2025-02-28;责任编辑:覃怀银
*通讯作者:赵铁松

避险提供最直接可靠的信息. 感知任务涉及对环境进行分割、目标检测和跟踪等关键步骤. 其中,3D目标检测是自动驾驶感知领域的研究热点之一,旨在从传感器数据中准确地识别和定位3D空间中的目标物体[2],例如汽车、行人和自行车等,这对于自动驾驶系统来说至关重要.

根据传感器输入形式,可以将3D目标检测算法分为以下两种架构:基于单模态的方法和基于多模态融合的方法. 基于单模态[2,3]的方法通常使用单一传感器作为输入源. 在自动驾驶领域,常见的单一模态包括激光雷达、毫米波雷达和相机等传感器. 每种传感器都有着各自的特点和优势,比如,相机通过拍摄高分辨率的图像,可获取丰富的视觉信息,且硬件成本很低;激光雷达能够精确探测到目标的距离,随着扫描线的增加能生成高分辨率的点云数据,并具有良好的目标边界信息;毫米波雷达则能抵抗雨天、雪天等恶劣的天气条件,不受颜色、光照等影响,价格相对便宜. 在实际的自动驾驶情况下,通过单一类型的传感器进行目标检测是远远不够的. 首先,每种传感器都有其固有的缺点. 例如,相机易遭受光线强弱及物体遮挡影响,缺乏距离信息;激光雷达价格昂贵,易受天气影响等;毫米波雷达捕获的点云很稀疏,噪声多. 表1归纳了自动驾驶中几种典型传感器的优缺点. 其次,要实现真正的自动驾驶,我们需要应对多样化的天气、道路和交通条件,确保在不同环境下都能提供可靠的感知结果,依靠单一类型的传感器是难以满足的. 为了克服单一传感器的局限性,多模态融合成为一种重要的解决方案. 多模态融合[1,4~7]是指将多个传感器(如相机、激光雷达、毫米波雷达等)的数据进行融合,以综合利用它们各自的优势,从而提高目标检测的准确性、鲁棒性和可靠性.

近年来出现了不少关于3D目标检测的综述文章[1~5],与现有的综述相比,本文更全面地阐述了自动驾驶场景中的单模态和多模态目标检测算法,即对基于相机、激光雷达、传统毫米波雷达(简称雷达)以及新兴的4D毫米波雷达(简称4D雷达)的单个传感器检测和多传感器融合检测进行了分类归纳和典型算法分析,并对不同自动驾驶场景下各类检测算法的适用性进行归类. 另外,本文还在公共数据集上对比了不同传感器的经典算法,并分析了其性能差异. 表2总结了单模态方法和多模态方法的优缺点.

表1 各类传感器的优缺点

传感器	优点	缺点
相机	成本低,高分辨率、丰富的视觉信息.	易受光照和阴影等环境因素影响,缺乏目标的距离和深度信息.
激光雷达	精确的距离测量、高分辨率的点云数据、良好的目标边界信息.	价格昂贵,数据规模大,受雨天、雪天等天气影响.
毫米波雷达	价格便宜,抵抗雨天、雪天等恶劣天气条件,不受颜色、光照等影响.	低分辨率,点云稀疏,噪声多.
4D毫米波雷达	价格较便宜,高分辨率的点云信息,更准确的高度信息.	相比激光雷达分辨率仍较低,噪声仍较多.

表2 单模态与多模态对比

方法	定义	优点	缺点
单模态方法	使用单一传感器的数据作为输入源.	数据结构单一,处理方便.	难以克服每个传感器的固有缺点.
多模态方法	多个传感器的数据作为输入源进行融合.	综合利用每个传感器优缺点,特征丰富.	数据对齐困难,数据融合信息损失.

2 自动驾驶中的3D目标检测

2.1 单模态的3D目标检测

单一模态的3D目标检测是指在仅使用一种传感器或一种数据模态的情况下进行3D目标检测的任务[3]. 在本节中,我们对自动驾驶领域中经典的单模态3D目标检测算法进行分类阐述.

2.1.1 基于相机的3D目标检测算法

相机使用光学传感器捕获环境的2D图像信息,可以通过视差、纹理、颜色等特征来推断目标的深度和位置. 然而,相机在距离测量和遮挡物处理方面存在挑战,对光照和纹理变化敏感,并且在低光或高光照条件下性能下降. 尽管相机存在定位准确性欠佳和光照敏感的缺点,但因其能展示丰富的纹理和色彩,具有良好的可视效果以及成本低廉等优势,仍然是当前目标检测算法的研究热点之一.

基于相机的3D目标检测旨在给定RGB图像和相应的相机参数下,对感兴趣的对象进行分类和定位. RGB图像是2D数据,因此基于相机的3D目标检测的核心问题在于如何从2D数据当中生成3D结果. 根据该问题,我们可以将基于相机的3D目标检测方法进一步分为基于结果升维的方法、基于特征升维的方法和基于数据升维的方法.

(1)结果升维

基于结果升维的方法[8~14]利用传统的2D目标检测算法,在RGB图像中检测出目标的2D边界框,然后从这些结果中得到3D位置. 为了获得目标的3D位置,常用解决方案是使用卷积神经网络(Convolutional Neural

Network,CNN)估计深度值,CNN是一种强大的深度学习模型,具有良好的特征提取、学习及表示能力.通过将包含目标的图像块作为输入,CNN可以学习到输入图像与目标深度之间的非线性关系.然后应用公式(1)将2D坐标投影到3D空间:

$$\begin{cases} z=d \\ x=(u-C_u)\times z/f \\ y=(v-C_v)\times z/f \end{cases} \tag{1}$$

其中,d是通过CNN估计的深度值,(C_u,C_v)是图像的中心点,f是相机焦距,(u,v)是目标的2D坐标,(x,y,z)是目标的3D坐标.然而,仅通过CNN进行深度值估计来确定3D位置可能面临估计误差等问题.在这种情况下,需要设计中间模块来提取更具效用的特征,同时在本质上帮助网络建立与全局最优值之间的关联[2].

为了提高检测精度,Xu等人[8]提出基于端到端的多级融合框架.整个网络由两部分组成,一部分用于生成二维区域建议,另一部分用于同时预测目标的位置、方向、尺寸.作者引入全卷积网络(Fully Convolutional Network,FCN)估计每个像素的视差信息,利用摄像机标定文件得到近似的深度和伪点云,对估计信息进行叠加,实现3D定位.这一方法通过将不同参数预测解耦来实现3D目标检测,尽管在实践中能够提升3D定位的精确性,但依赖子网性能,增加了系统的复杂性,计算效率有待提高.

为了降低检测复杂度,Duan等人[9]提出了一种基于关键点估计的检测网络CenterNet.具体而言,将目标编码成一个关键点(文中取2D边界框中心点),将关键点视为图像平面上的3D投影位置,利用估计的深度和摄像机参数将其反向投影到3D空间.尽管这个检测器在架构上看起来非常简单,但在许多任务和数据集上都取得了不错的性能,很多工作都是在此模型的基础上进行改进的.Ma等人[10]基于CenterNet网络提出了单目定位误差(Delving into Localization Errors for Monocular,MonoDLE)的检测算法,通过大量实验定量分析发现定位误差是制约单目3D检测的重要因素,给出了三种定位误差的改进策略,包括估计2D边界框中心点和3D投影中心的偏差来确定关键点,调整训练样本,优化损失函数.Zhang等人[11]同样基于CenterNet网络,区分了完整目标和残缺目标,设计了灵活的单目检测(Flexible Monocular,MonoFlex)算法.对于完整物体,通过预测2D框中心点与3D框投影坐标之间的偏差来确定关键点;对于截断物体,该方法预测2D框中心和3D框中心投影点与图像边沿之间交点的偏差来确定关键点.但由于缺乏准确深度信息,两种算法难以精确定位物体.

基于CenterNet的方法在特征提取时聚焦于局部区域,可能导致深度预测误差符号一致,从而阻碍深度精度提升.因此,Yan等人[14]提出基于深度互补的3D目标检测方法(Monocular 3D object detection with Complementary Depths,MonoCD),利用全局深度线索来缓解局部相似性,并结合多个深度线索之间的几何关系来实现深度互补,提高了深度估计精度.然而,该方法在泛化能力方面有一定的局限性,对新场景的适应能力可能较差.

(2)特征升维

基于特征提升的方法[15~24]旨在将图像坐标系中的2D图像特征转换为世界坐标系中的3D特征.具体而言,首先从输入的2D图像中提取出一系列与目标物体相关的特征.这些特征可能包括颜色、纹理、边缘等图像特征.然后,通过将这些2D图像特征与相机的内外参数进行关联,将其映射到世界坐标系中.在世界坐标系中,每个目标物体都具有其特定的3D坐标,表示其在空间中的位置.

Roddick等人[15]提出基于正交特征变换法(Orthographic Feature Transform,OFT)的检测网络,旨在解决如何将基于图像的2D特征映射到3D空间,以实现特征升维.它通过将自定义体素的左上角(u_1,v_1)和右下角(u_2,v_2)投影到视图像特征的区域上,使用式(2)累积2D特征来获得体素特征:

$$V(x,y,z)=\frac{1}{(u_2-u_1)(v_2-v_1)}\sum_{u=u_1}^{u_2}\sum_{v=v_1}^{v_2}F(u,v) \tag{2}$$

其中,$V(x,y,z)$和$F(u,v)$表示3D体素(x,y,z)和2D像素(u,v)的特征.该方法在资源有限的情况下提供了一种更有效的方法来解释和理解3D环境,然而因其直接基于2D特征得到3D特征,检测性能很大程度上依赖于特征提取器.

为提高检测性能,Reading等人[17]提出了基于分类深度分布网络(Categorical Depth Distribution Network,CaDDN)的检测算法.CaDDN把连续深度空间离散化为多个深度区间,将深度估计视为分类任务.接着将图像特征与估计的深度关联起来,即使用式(3)将图像信息投影到3D空间中:

$$G(u,v)=D(u,v)\otimes F(u,v) \tag{3}$$

其中,(u,v)表示像素点坐标,$D(u,v)$表示预测的深度分布,$F(u,v)$为提取的图像特征,$G(u,v)$为视锥特征,$\otimes$代表张量积.进而,使用三线性插值法对视锥特征进行采样来填充体素,从而得到3D特征.然而,该方法使用预先训练过的深度估计器的方法,存在额外的计算成本,并且不准确的深度先验会影响性能提升[12].

Li等人[20]提出了一种基于时空Transformer的鸟瞰图表示方法BEVFormer,该方法不依赖深度信息,通过在跨视图视角中动态聚合空间和时间信息,生成准确

的鸟瞰图特征用于目标检测，提高了检测精度. Zhang等人[23]则沿用了CaDDN的分类深度方法思想，提出基于深度引导的3D目标检测(Depth-guided Transformer for Monocular，MonoDETR)算法，通过量化连续深度值来预测前景深度图，并利用Transformer结构，结合预测的深度信息来引导特征提取，从而使网络能够更加关注细节目标，并且无需进行任何后处理. 此外，作者将这种深度引导方法引入BEVFormer中，进一步提升了该算法的检测精度. 但由于单目视觉方法的固有限制，与采用激光雷达技术或多模态融合技术的方法相比，这两种算法在精度和性能方面仍存在一定差距[22].

(3)数据升维

基于数据升维的方法将2D图像转换为3D伪点云，然后从得到的伪点云中提取3D特征. 首先需要从图像中估计密集深度图，并根据坐标转换推导出像素的3D位置. 通过将所有像素投影到3D坐标中来生成伪点云. 之后使用伪点云作为输入，利用基于点云的检测方法检测目标. 其中，基于数据升维的模型的性能在很大程度上依赖于估计深度图的质量. Wang等人[25]基于视觉深度估计，提出从图像生成伪激光雷达(Pseudo-lidar)的检测方法. 该方法利用立体视差法估计深度，即以一对左右图像作为输入，然后进行特征提取得到左右特征图，拼接左右特征图构建视差成本容积，用于表示每个像素点在不同视差情况下的匹配成本. 然后再经过3D卷积和Softmax函数得到$Y(u,v)$，$Y(u,v)$记录每个像素的水平视差，利用式(4)估计每个像素点的深度$D(u,v)$：

$$D(u,v)=\frac{f_U\times b}{Y(u,v)} \tag{4}$$

其中，f_U为左视图焦距，b为左右视图的水平偏移量，$Y(u,v)$为像素点(u,v)的左右水平视差. 进而，将深度图反向投影到激光雷达点云坐标系上生成伪点云. 该方法将图像表示转换成伪点云表示，但图像视差中的小偏差会造成远距离深度换算的大误差，导致远处物体位置估计效果不佳[26].

You等人[26]改进了Pseudo-lidar深度估计方法，提出了Pseudo-lidar++算法，同样是以一对左右视图作为输入，对左右视图进行特征提取并拼接构建视差成本容积. 不同是该方法将视差成本容积转换成深度成本容积，深度成本容积存储了每个像素的深度可能性. 通过优化深度估计方法提高伪点云表示的准确性，但生成的伪点云与真实点云仍存在偏差. 此外，作者还探索了利用低成本的4线激光雷达传感器采集极其稀疏的点云作为辅助，来纠正深度估计偏差. 图1展示了不同方法产生的伪点云，可以看出由激光雷达辅助产生的伪点云(紫色)比起图像估计的伪点云(红色)，与真实目标(蓝色)的差距更小.

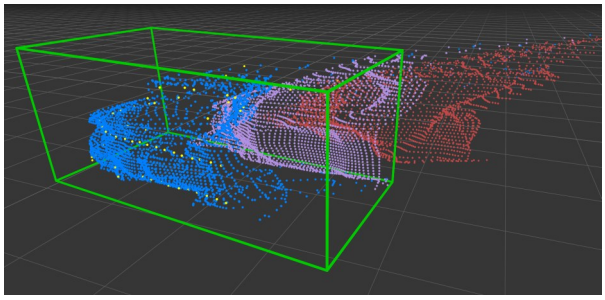

图1 点云可视化

(4)相机检测算法小结

基于结果升维的方法是一类在2D特征维度上进行处理的目标检测方法，能够充分利用已经发展成熟的2D目标检测方法，易于实现和部署，但这类方法无法充分利用3D信息，难以获取目标的精确3D位置和姿态. 基于特征升维和基于数据升维的方法在3D特征维度上进行处理. 其中，基于数据升维的方法可以增加数据表示形式，提升检测性能，但伪点云的生成很大程度依赖于深度估计的准确性. 基于特征升维的方法通过提取更丰富、更具信息量的3D特征来实现3D目标检测，但在特征提取的过程可能会受到数据质量和噪声的影响.

图2给出了近几年基于相机的部分检测算法及应用场景分类，表3总结了上述介绍的基于相机的部分检测算法特点，可以发现基于特征提升和结果提升的方法是目前比较主流的方法.

2.1.2 基于激光雷达的3D目标检测算法

激光雷达通过向周围发射激光束并测量其返回时间来感知环境. 尽管激光雷达的成本较高，对天气条件敏感(如雨雪、雾等)，且难以识别目标的细节特征，但因其能提供精确的距离测量、高分辨率的点云数据、良好的目标边界信息和准确的物体姿态估计[27]，是自动驾驶领域不可或缺的感知设备. 激光雷达生成的点云具有丰富的3D信息，根据点云的处理方法可以将基于激光雷达的3D物体检测方法分为以下三类：基于原始点集的方法、基于体素的方法和基于体素-点的方法.

(1)基于原始点集

基于原始点集的检测算法[28~34]是一种直接从原始点云数据中提取特征并进行目标检测的算法. 点云本质上是一长串点，并且在空间上具有无序性和旋转性的特点[28]. 无序性指点的顺序不影响它在空间中对整体形状的表示，例如，相同的点云可以由两个完全不同的矩阵表示. 旋转性指相同的点云在空间中经过一定的刚性变化后，坐标会发生变化. 基于原始点集的方法需要确保无论点云的顺序如何，都能获得相同的特征

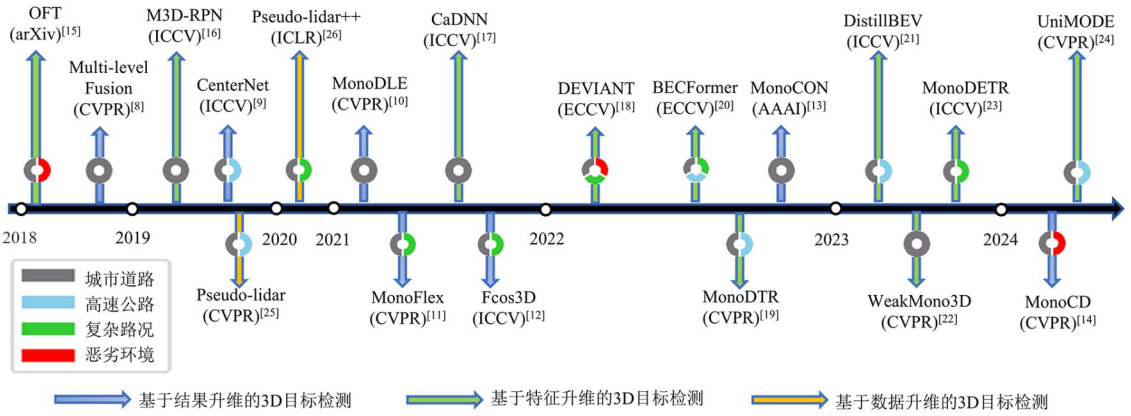

图2　基于相机的3D目标检测算法

表3　基于相机的部分目标检测算法总结

类型	每类方法的特征	作者	典型算法	典型算法的优缺点
结果升维	特点:先由图像检测出目标的2D边界框,然后从这些结果中得到3D位置. 优点:可以利用2D目标检测方法,较其他类方法简单易行. 缺点:无法充分利用3D信息,定位不准确.	Xu等人	Multi-level Fusion[8]	优点:引入全卷积网络估计视差,实现3D定位,提高检测精度. 缺点:依赖子网性能,增加了系统的复杂性.
		Duan等人	CenterNet[9]	优点:引入基于关键点的方法减少计算量,提高了检测效率. 缺点:结构简单,性能仍有改进空间.
		Ma等人	MonoDLE[10]	优点:分析了单目3D目标检测中的定位误差问题,提出了多任务学习等策略提高检测精度. 缺点:缺乏深度信息,难以准确定位物体.
		Zhang等人	MonoFlex[11]	优点:考虑物体的完整性差异,提高了检测的鲁棒性 缺点:缺乏深度信息,难以准确定位物体.
		Yan等人	MonoCD[14]	优点:利用全局深度线索来缓解局部相似性,结合多个深度线索之间的几何关系来实现深度互补,提高深度估计精度. 缺点:在泛化能力方面有一定的局限性,对新场景的适应能力可能较差.
特征升维	特点:将图像坐标系中的2D图像特征转换为世界坐标系中的3D特征. 优点:具有信息较丰富的3D特征. 缺点:特征提取受数据质量、噪声影响.	Roddick等人	OFT[15]	优点:引入正交特征变换提升了检测精度. 缺点:模型质量过于依赖特征提取器质量,容易影响检测精度.
		Reading等人	CaDNN[17]	优点:将深度估计任务视为分类任务,提高深度估计精度. 缺点:使用预先训练过的深度估计器的方法造成额外的计算成本,并且性能提升受限于深度先验的准确性.
		Li等人	BEVFormer[20]	优点:聚合跨视图时间和空间信息,提高了检测精度. 缺点:检测精度与激光雷达技术的方法相比仍有差距.
		Zhang等人	MonoDETR[23]	优点:利用深度引导特征提取,能捕捉输入图像中的细微线索. 缺点:由于单目视觉方法的固有限制,与采用传感器融合技术的方法相比,在精度和性能方面仍然存在显著差异.
数据升维	特点:将图像转换为伪点云,从得到的伪点云中提取3D特征. 优点:增加数据表示形式,提升检测性能. 缺点:依赖于深度估计的准确性,计算量较大.	Wang等人	Pseudo-lidar[25]	优点:提出基于立体视差法的伪激光雷达表示方法,可以沿用激光雷达的检测方法,提升检测精度. 缺点:依赖视差估计,导致远处目标检测精度低.
		You等人	Pseudo-lidar++[26]	优点:将立体视差法的视差成本容积转换成深度成本容积,提高伪点云表示准确性. 缺点:生成的伪点云与真实点云仍存在偏差,影响检测性能.

提取结果,并且,这些方法不受点云在不同坐标系下的表现方式的影响.

Qi等人[28]首先基于原始点集编码提出了PointNet,网络框架如图3所示.该框架将点云中的每个点作为独立同分布输入,通过一系列的神经网络将点云从低维表示映射到高维特征表示,作者在高维特

征后添加最大池化操作解决点云无序性问题，利用特征变换解决旋转性问题．然而，PointNet是对每一个点做从低维到高维的映射的学习，缺少局部的概念[29]．

为解决上述问题，该团队提出了PointNet网络的升级版PointNet++[29]．作者在PointNet的基础上，添加了集合抽象(Set Abstraction，SA)层来提取局部特征．SA层由采样层、分组层和特征提取层组成．其中，采样层利用最远点采样(Farthest Point Sampling，FPS)在点云中抽取部分点作为中心点，分组层在中心点周围寻找K个临近点组成局部区域点集，特征提取层提取每个局部区域点集特征，通过对局部区域进行特征迭代，提取丰富局部信息．但对于具有大量噪声的点云数据，PointNet++网络可能会受到影响．

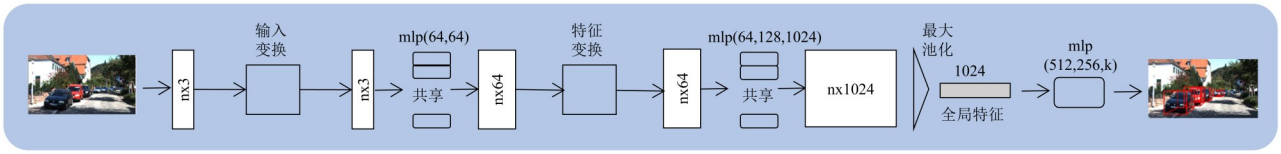

图3 PointNet网络架构[28]

Qian等人[33]为提高PointNet++的检测精度，改进了训练策略，提出了PointNeXt算法．通过叠加实验定量地研究每个数据集上每种数据增强方法和优化策略的效果，针对每种数据集提出改进的优化策略．此外，作者提出反向残差多层感知机(Inverted Residual MLP，InvResMLP)模块，该模块在SA模块的基础上引入了残差连接以缓解梯度消失问题，并利用了可分离的MLP结构来降低计算复杂度．然而，该方法的改进仅限于现有模块，没有引入新的体系结构[33]．

(2)基于体素

体素是一种将3D空间划分为规则小块的表示方式，类似于像素在2D图像中的作用．基于体素的方法[35~43]通过将点云数据映射到相应的体素中，可以将连续且无规则的点云信息转化为离散且规则的体素表示，并且可以在每个体素中记录点的存在与否．由于点云分布稀疏，3D空间中的大部分体素都是空的，不包含点．在实际应用中，只有那些非空体素被存储并用于特征提取．

Zhou等人[35]首先提出了基于点云体素特征的编码网络VoxelNet．该方法设计了一种体素特征编码(Voxel Feature Encoding，VFE)层，通过将逐点特征与局部聚合特征相结合，可以实现体素内点之间的交互，有效处理大规模点云．但该方法受限于3D卷积的运算量，推理速度十分缓慢[36]．为解决该问题，Yan等人[36]基于该框架引入稀疏卷积，提出了稀疏嵌入式卷积检测(Sparsely Embedded CONvolutional Detection，SECOND)方法，它在卷积操作中只考虑非零值的体素，从而减少了计算量，并且能够有效地提取点云数据中的特征．然而该方法对于鸟瞰图特征检测，以及针对行人和自行车的检测精度较低．

尽管体素化可以实现更快的处理速度，却存在局部信息丢失，导致精度下降的问题，为提升细节信息检测精度，Deng等人[37]引入了体素感兴趣区域(Region Of Interest，ROI)池化操作，提出了基于体素区域卷积神经网络(Voxel Region-based Convolutional Neural Network，Voxel RCNN)的目标检测算法，将感兴趣区域划分为多个规则子体素，以中心点代表各个体素，利用体素查询操作来得到局部邻近信息，利用池化操作来聚合局部信息，从而提高感兴趣区域的检测性能．

稀疏体素特征需要通过密集预测头进行特征密集化处理，这不可避免地引入了额外的计算量．为了降低计算复杂度，Chen等人[42]提出基于完全稀疏体素网络(Fully Sparse Voxel Network)的VoxelNeXt算法．该方法去掉锚点、稀疏密集转换、区域建议网络和其他复杂组件，直接对稀疏卷积体素特征进行下采样，以扩大感受野，并将稀疏特征放到统一鸟瞰图平面上进行叠加求和，进而对压缩后的特征进行预测，加快了检测速度．但该方法依赖单中心体素特征进行检测，准确性较低．Zhang等人[43]同样采用完全稀疏的方法，提出了稀疏自适应特征扩散网络(Sparse Adaptive Feature Diffusion Network，SAFDNet)．该自适应特征扩散选择性地将物体边界框内的特征扩展到相邻区域，并根据体素位置动态调整扩散范围，利用扩展后的特征进行预测，解决中心特征缺失问题，提高了检测效率．但该方法可能会产生类似噪声的特征区域，会影响检测性能．

(3)基于体素-点

基于体素的方法在计算上具有高效性，但其数据量化过程可能导致细节信息的丢失．相比之下，基于点的方法保留了点云的不规则性和局部性，但容易受到点云的噪声影响．因此，一些学者采用混合模式，集成两种方法的优势进行目标检测[44~47]．

He等人[45]通过在原始点云上应用辅助网络，以赋予体素特征的结构感知能力，提出了基于结构感知的单阶段检测器(Structure-Aware Single-Stage Detector，SA-SSD)．具体而言，辅助网络首先将基于SECOND的骨干网络生成的特征转化为点级表征信息，然后分配

前景分割和点中心估计两个辅助任务,前者使得特征对物体边界更为敏感,而后者则促使特征学习物体内部的关系. 该方法在保持推理速度的同时,提高了检测准确性,然而增加了训练复杂度.

为了进一步提高检测精度,Shi等人[46]致力于双阶段检测算法,提出了基于点-体素的区域卷积神经网络(Point-Voxel Region-based Convolutional Neural Network,PV-RCNN),集成了多尺度三维体素特性和关键点特性,聚合到ROI网格点以捕获更丰富的上下文信息,从而得到细粒度的区域建议,然而增加了算法复杂度. 为了降低计算开销,该团队在PV-RCNN的基础上提出了改进的PV-RCNN++[47],设计了扇区化建议中心策略来提高关键点采样的效率,并利用聚合模块获取局部特征,加快处理速度且提升了性能. 但是,此类双阶段检测方法的速度与单阶段方法相比仍有差距.

(4)激光雷达检测算法小结

基于原始点集的方法充分利用点云的形状细节信息,具有排序和旋转不变性. 然而,由于点云的稀疏性,该方法容易受到噪声的影响,导致误检和漏检的问题. 此外,处理大规模点云数据所需的计算量也相对较大. 基于体素的方法可以有效处理噪声问题和计算量问题,但体素网格的固定分辨率使其容易失去目标的精细形状和固定分辨率,影响检测精度. 基于体素-点的方法虽然能结合基于点和基于体素的方法的优点,但融合的方法会增加算法的复杂性和计算成本.

图4展示了近年来基于激光雷达的检测算法及应用场景分类,表4总结了上述介绍的基于激光雷达的目标检测算法特点. 从中可以观察到基于体素的方法展现出巨大的潜力. 这一趋势源于激光雷达技术的发展,包括高精度激光雷达(如64线激光雷达)的出现等,这些新技术有效地解决了体素网格在细节损失方面的问题.

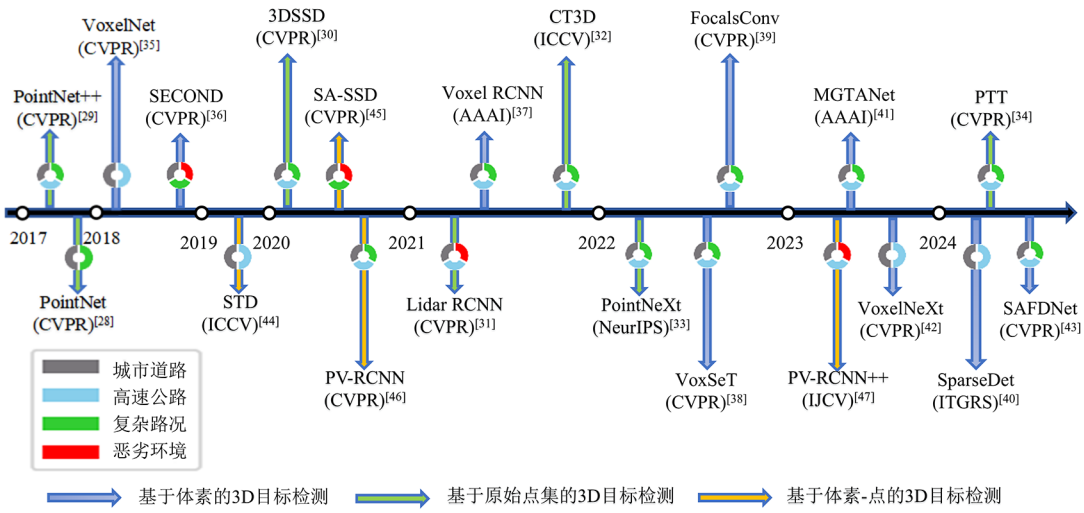

图4　基于激光雷达的3D目标检测算法

2.1.3　基于3D雷达/4D雷达的3D目标检测算法

雷达是一种工作于毫米波频段的有源传感器,通过测量反射波来确定物体的位置和速度,无需引入额外的时间信息. 相较于激光雷达,雷达具有更低的成本,并且具备抵抗恶劣天气条件(如雾、烟和灰尘)的能力,且对于照明变化不敏感. 然而,由于雷达的分辨率低、噪声大,难以通过稀疏的雷达点云直接检测物体的形状. 为了改善该问题,4D雷达作为3D雷达的改进版本出现,具有更高的分辨率,并且能够测量距离、方位角、仰角和速度等四种目标信息. 然而,在现有的信号处理流程中,由4D雷达生成的点云仍然存在稀疏性和噪声问题,点云质量与激光雷达点云相比仍有一定差距,并且缺乏大规模数据集.

(1)基于3D雷达的目标检测

雷达点云的稀疏性问题对于传统的检测方法(例如基于体素的方法)构成了重大挑战. 这些方法倾向于通过降低分辨率来处理稀疏点云数据,导致提取的细节信息进一步丢失. 为了解决这一挑战,一些研究提出应用原始点云特征提取方法[48~51]从输入数据中获得具有更大的感受野的特征.

Bansal等人[48]基于PointNet网络提出了Pointillism算法,通过空间和时间的一致性分析来区分噪声点和真实点,提高了输入数据质量. 然而由于雷达点云的稀疏性,该方法可能无法利用点间关系中的语义信息,从而影响3D边界框预测的准确性. 为了提高检测精度,Svenningsson等人[51]提出基于图神经网络的雷达点云(Radar-Point Graph Neural Network,Radar-PointGNN)检测方法,单独提取每个点的特征,并将相邻的点连接起来生成图结构,进而利用图卷积提取点的邻域信息以

表4 基于激光雷达的部分目标检测算法总结

类型	每类方法的特征	作者	典型算法	典型算法的优缺点
基于原始点集的方法	特点:直接从原始点云数据中提取特征并进行目标检测. 优点:充分利用点云的形状细节信息,具有排序和旋转不变性. 缺点:容易受到噪声影响导致误检漏检,处理大规模点云的计算量较大.	Qi等人	PointNet[28]	优点:解决点云无序性和旋转性问题. 缺点:没有考虑点与点之间的局部关系,限制了细节检测的性能.
		Qi等人	PointNet++[29]	优点:添加集合抽象层,增强了其对局部结构的捕捉能力. 缺点:对于大量噪声点云数据处理效果一般.
		Qian等人	PointNeXt[33]	优点:改进训练策略以提高检测精度,引入残差链接缓解梯度消失问题,并利用可分离的MLP结构降低计算复杂度. 缺点:改进仅限于现有的模块,没有引入新的体系结构.
基于体素的方法	特点:通过将点云数据映射到相应的体素中,可以将连续且无规则的点云转化为离散且规则的体素表示,并且可以在每个体素中记录点的存在与否. 优点:有效处理点云噪声问题,相对原始点集方法计算量较小. 缺点:容易失去目标的精细形状和固定分辨率,导致检测精度欠佳.	Zhou等人	VoxelNet[35]	优点:设计了一种逐点特征与局部聚合特征相结合的体素特征编码,有效地处理大规模的点云数据. 缺点:受限于3D卷积的运算量,推理速度缓慢.
		Yan等人	SECOND[36]	优点:引入稀疏卷积神经网络,提高了计算速度. 缺点:对于鸟瞰图特征检测,以及针对行人和自行车的检测精度较低.
		Deng等人	Voxel RCNN[37]	优点:引入了体素感兴趣区域聚合局部信息,提高了检测性能. 缺点:增加了计算量.
		Chen等人	VoxelNeXt[42]	优点:直接在稀疏体素特征上进行预测,降低了计算复杂度,提高了推理速度. 缺点:依赖单中心体素特征进行检测,准确性较低.
		Zhang等人	SAFDNet[43]	优点:提出了一种自适应特征扩散策略,解决中心特征缺失问题,提高了检测效率. 缺点:可能会产生类似噪声的特征区域,影响检测性能.
基于体素-点的方法	特点:在体素点云上添加点的结构特性,或者在原始点上添加体素的结构特性. 优点:能结合基于点和基于体素的方法的优点. 缺点:增加算法的复杂性和计算成本.	He等人	SA-SSD[45]	优点:原始点云上应用辅助网络以赋予体素特征的结构感知能力,保持推理速度的同时,提高了检测准确性. 缺点:增加了训练复杂度.
		Shi等人	PV-RCNN[46]	优点:集成了多尺度三维体素特性和关键点特性,捕获更丰富的上下文信息,得到细粒度的区域建议,提高了检测性能. 缺点:增加了计算复杂度.
		Shi等人	PV-RCNN++[47]	优点:设计扇区化提议中心策略来提高关键点采样的效率,加快处理速度并提升性能,设计聚合模块获取局部特征,节省内存. 缺点:属于双阶段检测方法,检测速度与单阶段方法相比仍有差距.

生成上下文信息,充分利用了点间局部特征,提高了检测性能. 然而,该方法需要对每个点云单独处理,计算量较大.

为了减少计算开销,一些学者利用雷达的距离-方位-多普勒(Range-Azimuth-Doppler,RAD)张量作为数据输入[52~55]. Zhang等人[52]结合雷达信号信息,提出了基于距离-方位角-多普勒的检测方法(Range-Azimuth-Doppler based Detection,RADDet)算法,将多普勒维度视为距离-方位图(Range-Azimut,RA)的通道,然后利用YOLO(You Only Look Once)算法对RA图进行目标检

测. 该方法牺牲多普勒速度的空间分布特性来换取较小的计算和内存开销,可能会影响检测精度. Decourt等人[54]则提出基于距离-多普勒图(Range-Doppler,RD)的深度汽车雷达目标检测器(Deep Automotive Radar Object Detector,DAROD),通过卷积神经网络提取特征,并利用区域建议网络在RD图中生成区域建议,提高了检测精度. 然而,该方法忽略了雷达的时序信息,预测性能仍有提升空间.

(2)基于4D雷达的目标检测

与雷达类似,基于4D雷达的目标检测也可以利用原始点云特征提取方法来捕获特征[56~59]. 研究[56,57]证明基于注意力机制的Transformer在处理这些具有排列不变性的点集时具有先天的优越性. Bai等人[56]基于原始点集的方法,提出一个由自注意力模块构成的分类网络Radar Transformer,利用向量注意力模块来提取局部特征和全局特征,实现深度特征融合,提高检测精度. 然而,随着4D雷达的发展,点云密度逐渐增加,直接对点集进行处理会带来较大的计算量. 为了降低复杂度,Xu等人[57]将点云划分为体柱,利用自注意力机制提出了雷达体柱特征注意力网络(Radar Pillar Feature Attention Network,RPFA-Net),通过从点云中提取全局特征,有效地捕获了远距离信息并提高了航向角回归能力,但该方法容易受点云数据噪声影响. Liu等人[58]提出空间多表示融合(Spatial Multi-Representation Fusion,SMURF)的检测算法,通过点云体柱化分支和核密度估计(Kernel Density Estim-ation,KDE)分支提取体柱化和KDE特征,得到多尺度融合特征. 其中,KDE有效地减轻了点云中的固有噪声和稀疏性的影响,然而也会增加一定的计算量. 一些研究借鉴3D雷达RAD张量算法,通过增加高度张量维度将其扩展到4D张量,以提高检测效率. Paek等人[60]提出了一种直接使用4D张量作为输入的3D目标检测方法K-Radar,并验证了4D张量的高度信息对于3D目标检测至关重要. 但该方法仅提供了一个基线网络,检测性能仍有提升空间.

图5展示了基于雷达和基于4D雷达的检测算法及应用场景分类,表5总结了上述介绍的基于3D雷达和4D雷达的目标检测算法特点. 4D雷达是一种新兴传感器,不仅适用于恶劣环境,而且能取得比3D雷达更高的性能,因而具有较大的研究潜力.

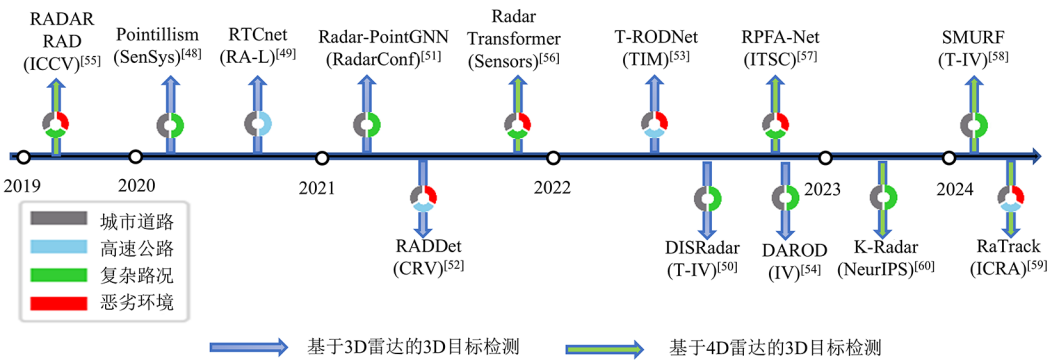

图5　基于3D/4D雷达的3D目标检测算法

2.1.4　单模态检测算法小结

根据不同传感器的输入模态,可以将其分为RGB图像、激光雷达及3D/4D雷达模态. 图像模态具备丰富的颜色信息,并且在数据处理方面具有较低的计算复杂度,而激光雷达和毫米波雷达模态则包含着丰富的3D信息. 这三种模态在自动驾驶系统中扮演着重要的角色,并且均具备着潜在的研究价值. 图6展示了不同模态的总体流程框架.

2.2　多模态融合的3D目标检测

3D目标检测的多模态融合是指将多种传感器的数据进行融合,以提高目标检测系统在3D环境中的性能和鲁棒性[4]. 多模态融合的目标是综合利用各个传感器的优势,并弥补单一传感器的局限性,从而得到更准确、可靠且全面的目标检测结果. 如图7所示,本节从融合传感器、融合位置以及融合输入类别三个角度出发,阐述了现有多模态融合的3D目标检测方法.

2.2.1　融合传感器及融合输入类别

融合的分类方式有多种,其中融合传感器和融合输入类别最具多样性,代表了每个设计的独特理念. 融合传感器根据所使用的感知设备类型进行划分,包括相机、雷达和激光雷达的两两组合方案;融合输入类别根据输入点云的类型进行分类,可以选择将体素表示、原始点云或者其在鸟瞰图或前视图上进行投影的点云作为融合输入. 本节中,我们根据融合传感器类别结合融合输入类型对融合算法进行了分类.

(1)激光雷达与相机融合的3D目标检测方法

①鸟瞰图投影与图像融合

在3D目标检测发展之前,基于图像的2D物体检测

表5 基于3D/4D雷达的部分目标检测算法总结

类型	方法及特征	作者	典型算法	典型算法的优缺点
3D雷达	基于原始点集:应用原始点云特征提取方法从输入数据中获得更大感受野的特征.	Bansal等人	Pointillism[48]	优点:通过空间和时间的一致性分析来区分噪声点和真实点,提高了输入数据质量. 缺点:无法利用点间关系中的语义信息,影响检测准确性.
		Svenningsson等人	Radar-Point GNN[51]	优点:利用图卷积提取点的邻域信息,充分利用点间局部特征,提高检测性能. 缺点:需要对每个点云单独处理,计算量较大.
	基于RAD张量:利用雷达的距离-方位-多普勒张量作为输入以减小计算开销.	Zhang等人	RADDet[52]	优点:计算复杂度低,内存开销小. 缺点:牺牲多普勒速度的空间分布特性,影响检测效果.
		Decourt等人	DAROD[54]	优点:利用区域建议网络在RD图中生成区域建议,提高了检测精度. 缺点:忽略了雷达的时序信息,预测性能仍有提升空间.
4D雷达	基于原始点集:除了直接利用原始点云特征提取方法来捕获特征外,还可以转换为体柱化表示方法来减小计算复杂度.	Bai等人	Radar Transformer[56]	优点:利用向量注意力模块来提取局部特征和全局特征,实现深度特征融合,提高检测精度. 缺点:采用原始点集的方法会带来较大的计算量.
		Xu等人	RPFA-Net[57]	优点:将点云划分为体柱,利用自注意力机制从体柱化点云中提取全局特征,有效捕获远距离信息,提高了航向角回归能力. 缺点:该方法容易受点云数据噪声影响.
		Liu等人	SMURF[58]	优点:将点云体柱化,并通过核密度估计减轻点云中的固有噪声和稀疏性的影响. 缺点:增加了计算量.
	基于4D张量:相较3D雷达的RAD张量,增加了高度张量.	Paek等人	K-Radar[60]	优点:引入4D张量,提高了检测效率. 缺点:仅提供了一个基线网络,检测性能仍有提升空间.

技术已经相当成熟. 因此在针对点云数据的处理时,一些学者提出将3D点云投影到2D平面上[61~64],即鸟瞰图投影. 在鸟瞰图投影中将3D点云数据投影到垂直于高度方向的平面上,通过这种方式,点云数据的3D信息被转化为2D平面上的像素值,提供了全局的视角,并且相比于直接在3D点云数据上进行目标检测,鸟瞰图投影可以降低计算复杂度,提高算法的效率. 然而,由于高度信息的压缩,会损失部分数据深度信息.

Chen等人[61]从多个视角出发,设计了一个多视角3D目标检测网络(Multi-View 3D object detection network, MV3D). 该方法利用点云的鸟瞰图生成3D候选框,再将候选框投影到图像、点云上,在图像、点云、鸟瞰图多个模态上获取区域特征,融合不同模态的区域特征用于目标分类和边界框回归,有效降低了数据维度的同时保留了关键信息. 但是该方法融合的信息较少,对于小目标来说检测精度较低. 为解决该问题,Ku等人[62]提出基于聚合视图的检测网络(Aggregate View Object Detection network,AVOD),利用RGB图与鸟瞰图的特征图融合来进行边界框的预测,再将预测框投影到图像、鸟瞰图上获取区域特征. 相较于MV3D,该方法在预测边界框时具有更丰富的特征信息,有效提高了小目标的检测精度.

然而,上述粗粒度融合方式,在ROI矩形区域往往包含大量的背景噪声,使得融合效果欠佳[63]. Lu等人[64]引入空间注意力机制来关注不同尺度的特征信息,引入通道注意力机制以关注不同通道的全局信息,设计了用于3D目标检测的空间-通道注意网络(Spatial-Channel Attention Network,SCANet). SCANet兼具空间和通道注意力,可以有效区分多层次特征并抑制不明显的特征,有助于网络更加专注于关键信息,提升检测性能,但空间-通道注意力模块也额外增加了模型计算复杂度.

②点云体素化与图像融合

点云体素化是一种将不规则的点云数据转换为规则的3D体素表示的方法. 相对于将点云投影到鸟瞰图,体素化可以有效地保留高度信息,并且离散的体素表示使得数据处理更加方便和高效. 通过体素化能够将稀疏的点云数据转换为结构化的三维数据,再与二维图像进行多模态融合,使得深度学习模型能够同时利用几何信息和视觉信息来提高3D目标检测或场景理解的精度. 随着点云体素化处理技术的成熟发展,一些学者[65~72]将体素化作为融合输入的一种策略,以利用其在多模态融合中的潜力.

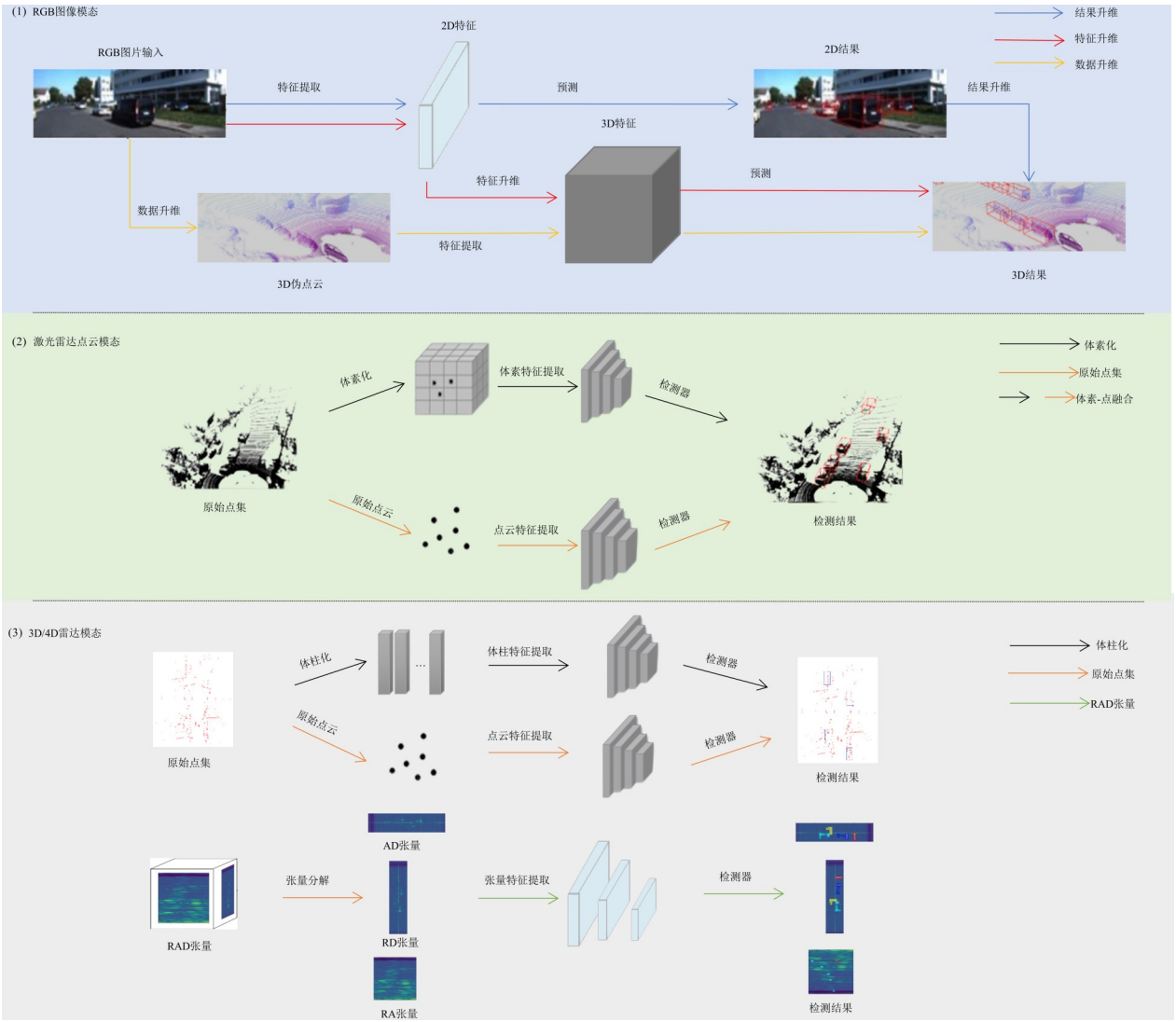

图6　单模态3D目标检测总体框架

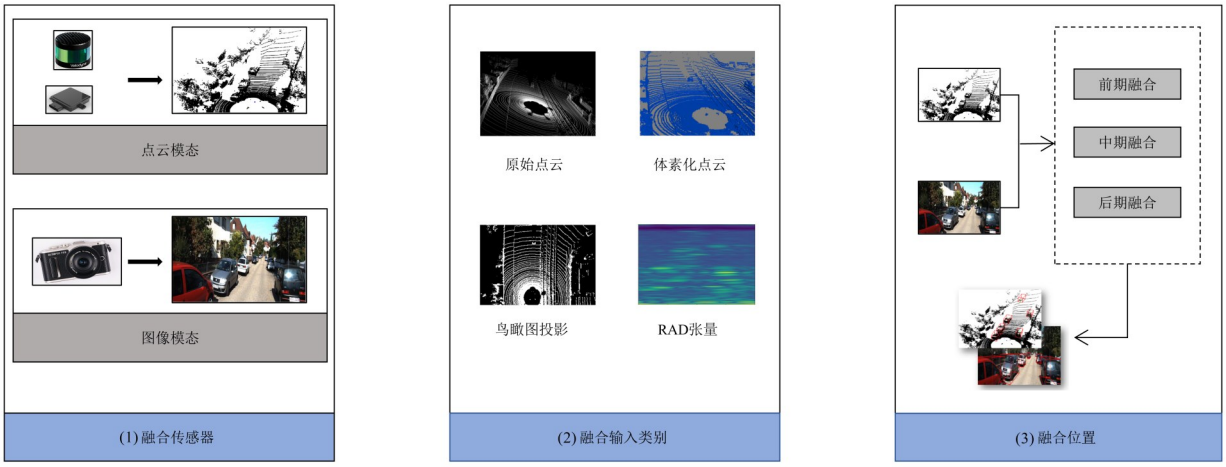

图7　多模态3D目标检测融合方法

Sindagi等人[65]提出基于体素投影的多模态融合检测网络(Multimodal VoxelNet,MVX-Net). 该方法将非空体素特征投影到图像平面来确定2D感兴趣区域,通过池化操作得到定长的2D特征,并将2D特征与体素特征

级联得到融合特征. 利用体素投影方法在体素级别进行特征融合,网络可以同时学习3D几何信息和2D语义信息,但可能出现多个体素投影到图像平面上的同一个区域的情况,会降低对不同体素的区分能力.

与体素投影不同,Li等人[68]利用深度估计,提出基于体素表示的Transformer检测网络(Unifying Voxel-based Representation with Transformer, UVTR). 该方法从图像中构建出体素特征,利用体素特征的空间交互特性将点云与图像整合到同一体素空间中进行目标预测. 该方法保留了体素空间的高度信息,避免了高度压缩带来的语义模糊性,有利于空间交互. 但由于仅在独立的主干网络中处理图像和点云,容易丢失细节信息.

为了全面挖掘点云和图像特征,Li等人[70]提出基于全局-局部的融合网络(Local-to-Global fusion Network, LoGoNet). 在全局模块使用体素内点的质心来表示非空体素特征的空间位置,并利用可变形交叉注意力模块自适应地融合点云特征与图像特征. 局部模块在区域级别动态融合点云特征和图像特征,用于提供更多局部和细粒度的几何信息. 然而该模型的复杂度较高,训练时间长.

为了在有效提取细节特征的同时兼顾复杂度因素,Song等人[71]提出多模态的体素融合网络VoxelNext-Fusion,通过融合粗粒度和细粒度的多模态特征,保持图像的连续性和语义性,并区分前景特征和背景特征,以消除背景像素特征的潜在影响. 该方法利用图像语义信息和背景信息,有效提高了远处目标的检测性能. 因其专注于远距离稀疏点云场景的性能提升,整体检测精度仍有待提高.

③原始点集与图像融合

尽管点云体素化可以保留点云的高度信息,但仍会造成一定的信息丢失. 较大的体素会导致细节信息丢失,而较小的体素可能会增加计算复杂度,并引入噪声. PointNet的出现使得可以直接处理原始点云,因此一些学者将原始点集与相机特征进行融合[73~77].

Xu等人[73]对图像数据和原始点云数据进行独立处理,提出了PointFusion算法. 通过级联方式融合图像特征和点云特征,可以保留原始特征的所有信息. 级联的融合方法操作简单,但缺乏对特征信息的局部处理,容易丢失细节信息. 为了提供更精细和个性化的表示,Huang等人[74]提出点云增强网络(Enhancing Point Network, EPNet). 作者设计了激光雷达引导融合(Lidar-guide fusion, Li-fusion)模块,建立点特征和相机图像特征之间的映射,以利用语义图像特征增强点特征,然而未考虑两种特征的双向交互[77],可能会影响检测性能. Liu等人[77]基于上述方法进行改进,提出基于级联双向多模态融合(Cascaded Bidirectional multi-modal Fusion, CB-Fusion)的EPNet++. 该模块在Li-Fusion的基础上增加了一个逆向模块,先通过图像对原始点云进行筛选,再用筛选后的点云提取更精准的图像特征. 利用多个CB-Fusion模块在多个尺度上建立图像和点云之间的双向信息交换路径,可以获得更全面和鲁棒的特征表示. 但该方法是基于级联思路的特征融合,是在单一层级上使用单一模态数据进行处理,可能无法充分利用不同模态数据的互补信息[3].

④激光雷达与相机融合目标检测方法小结

鸟瞰图投影与图像融合方法通过将3D点云转换为2D视角,有效提高了计算效率,然而由于高度信息的丢失,局部信息不完整,进而影响了目标检测精度. 点云体素化与图像融合方法在保留高度信息的同时简化了点云处理方式,具有较高的计算效率,但体素化网格可能导致边缘细节模糊. 原始点集与图像融合方法包含丰富的几何细节信息,但同时也带来了更大的计算量. 图8展示了近年来激光雷达与相机融合在目标检测领域的一些算法研究及应用场景分类,表6总结了上述激光雷达与相机融合的目标检测方法特点. 这些算法通过将激光雷达点云与相机图像进行融合,旨在提高检测性能和鲁棒性.

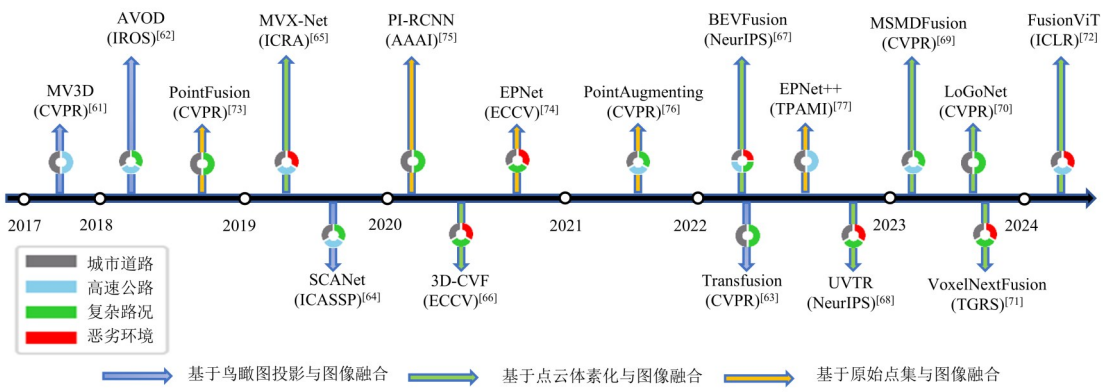

图8 激光雷达与相机融合的3D目标检测方法

表6　激光雷达与相机融合的部分目标检测方法总结

类型	每类方法的特征	作者	典型算法	典型算法的优缺点
鸟瞰图投影与图像融合	特点:将点云投影到鸟瞰图,利用二维鸟瞰图特征与图像特征进行融合. 优点:计算复杂度低. 缺点:压缩高度信息,损失部分深度信息.	Chen等人	MV3D[61]	优点:利用点云的鸟瞰图生成3D候选框,降低数据维度的同时保留了关键信息. 缺点:融合信息较少,小目标检测精度低.
		Ku等人	AVOD[62]	优点:RGB图与鸟瞰图的特征图融合,提升了特征表示的丰富性,提高了小目标的检测精度. 缺点:包含大量的背景噪声,使得融合效果欠佳.
		Lu等人	SCANet[64]	优点:引入空间和通道注意模块,有效区分多层次特征并抑制不明显的特征,专注于关键信息,提升检测性能. 缺点:增加了额外的复杂度.
点云体素化与图像融合	特点:将不规则点云数据转化为规则体素网格,再与二维图像特征进行融合. 优点:保留高度信息,简化点云处理方式,计算效率较高. 缺点:存在边缘细节模糊问题.	Sindagi等人	MVX-Net[65]	优点:将非空体素特征投影到图像平面,网络可以同时学习3D几何信息和2D语义信息,提升检测精度. 缺点:投影时体素区分能力低,影响融合效果.
		Li等人	UVTR[68]	优点:保留了体素空间的高度信息,避免高度压缩带来的语义模糊性,有利于空间交互. 缺点:独立的主干网络中处理图像和点云,容易丢失细节信息.
		Li等人	LoGoNet[70]	优点:在全局层面和局部区域分别融合,充分利用特征信息. 缺点:模型复杂度较高,训练时间长.
		Song等人	VoxelNeXtFusion[71]	优点:融合粗粒度和细粒度的多模态特征,利用图像语义信息和背景信息,有效提高了远距离目标的检测性能. 缺点:整体检测精度仍有待提高.
原始点集与图像融合	特点:原始点集与图像融合直接利用点云细节特征与图像特征融合. 优点:能提取较为丰富的细节信息. 缺点:相对其他类算法,计算量较大.	Xu等人	PointFusion[73]	优点:通过级联方式融合特征,操作简单,可以保留原始特征的所有信息. 缺点:缺乏对特征信息的局部处理,容易丢失细节信息.
		Huang等人	EPNet[74]	优点:通过图像语义特征增强点特征,以获得更精细和个性化的表示. 缺点:未考虑语义特征和点特征的双向交互,可能影响检测精度.
		Liu等人	EPNet++[77]	优点:建立图像和点云之间的双向信息交换路径,获得更全面和鲁棒的特征表示. 缺点:级联思路的特征融合难以充分利用不同模态数据的互补信息,可能会影响检测精度.

(2)雷达与相机融合的3D目标检测方法

雷达生成的点云数据常呈现出稀疏性,图像具有丰富的语义信息,许多研究者[78~87]采用了雷达与相机融合的方法,利用相机的高分辨率和雷达的深度信息,从而提高了对环境的感知能力,以改善雷达数据的不足之处.

为了弥补雷达数据的稀疏性,Bansal等人[80]提出了图像和雷达融合网络RadSegNet,使用了一种独立信息提取的设计理念,将图像的语义信息作为额外的特征通道与雷达特征拼接来增强雷达特征,在恶劣条件下仍具有可靠性,但是噪声增加依然会影响性能.Kim等人[81]提出了相机雷达网络(Camera Radar Net,CRN),利用稀疏的雷达点将图像特征转换为鸟瞰图特征,克服图像中空间信息的不足;同时,设计多模态可变形注意力模块,以解决输入之间的空间失调.然而,该方法在传感器故障情况下性能降低明显,鲁棒性需进一步增强.Long等人[82]则是利用图像增强雷达特征,提出了雷达图像关联网络(RADar-Image Association Network,RADIANT),对雷达深度与图像预测深度进行加权,用于得到更为准确的深度信息.该方法改进了深度估计,但没有考虑物体大小等其他参数的优化.一些学者利用雷达额外的输入信息,如距离、速度和雷达截面积(Radar Cross-Section,RCS)值等,以增强雷达特征.Lin等人[83]提出了基于雷达-图像融合的鸟瞰图检测器(Radar-Camera Fusion in Bird's Eye View for 3D Object Detection,RCBEVDet),利用RCS作为目标大小的先验信息,将雷达点特征散布到鸟瞰图空间中的多个像素上,以增加雷达鸟瞰图特征的密度.然而该模型对图像分支的处理较简单,可能会影响检测精度.

除了稀疏性外,雷达提供的方位角分辨率较低,具有较多噪声点,使得雷达点云与图像的关联可能存在偏差.为解决该问题,Nabati等人[84]提出一种基于中心

点的3D目标检测方法CenterFusion. 该方法使用目标的2D框中心点及其估计的深度和尺寸,为目标创建一个3D感兴趣区域视锥体,过滤掉与该目标无关的雷达点云,取最近的点作为该目标对应的雷达关联点. 但当两个目标中心点重合时,该网络不能有效区分. Kim等人[85]基于软极坐标关联方法,提出相机雷达Transformer融合的检测算法(Camera RAdar Fusion Transformer,CRAFT),将图像建议边界框坐标和雷达点坐标转换为极坐标,利用自适应阈值来关联雷达点云. 然而,该方法的检测精度与激光雷达相比仍有差距.

(3)雷达与激光雷达融合的3D目标检测算法

激光雷达和雷达是两种动态捕获点云3D信息的传感器,两者可以提供互补的传感信息,例如激光雷达专门捕捉物体的3D形状,而雷达提供更大的检测范围以及速度. 一些学者[88~90]尝试利用雷达与激光雷达融合进行目标检测,但由于雷达的点云数据稀疏,且噪声多,如何融合两者特征仍是挑战.

Yang等人[88]提出了一种激光雷达和雷达数据融合的检测网络RadarNet,该方法的关键是充分利用了雷达数据的几何和动态信息. 在几何方面,将雷达点云与激光雷达点云在体素层次上进行融合;在动态信息方面,聚合每个目标框与雷达点云的速度信息实现对目标速度的精确估计. 该方法为融合雷达数据中的几何和动态信息提供了一种有效方案,但没有解决雷达固有的高度信息不准确问题和稀疏性问题. Wang等人[89]提出基于激光雷达-雷达双向融合的检测方法(Bi-directional LiDAR-Radar Fusion,Bi-LRFusion). 具体而言,通过从激光雷达分支学习重要细节来丰富雷达的局部特征,以缓解高度信息缺失和极端稀疏性带来的问题,但雷达原本的噪声特征可能会影响融合性能.

上述方法旨在通过利用从雷达上获得的数据特征来弥补激光雷达的局限性,Bang等人[90]提出基于雷达蒸馏(Radar Distillation, RadarDistill)的检测算法. RadarDistill通过知识蒸馏技术,利用激光雷达数据指导雷达编码网络,生成与激光雷达特征相似的雷达特征,实现雷达数据增强. 该方法仅在训练阶段引入激光雷达数据,推理阶段只依赖于雷达数据,可能会限制模型在多样化数据场景下的适应能力.

(4)4D雷达与相机/激光雷达融合的3D目标检测算法

4D雷达在3D雷达的基础上增加了高度信息,具备独特的速度测量和全天候传感能力,因此具有很高的研究价值. 然而,4D雷达点云的稀疏性和噪声对性能的提升构成了阻碍. 通过将4D雷达与相机/激光雷达进行融合,有望实现自动驾驶中高精度性能. 目前,4D雷达在与其他传感器模态的深度融合方面缺乏深入的研究,并且缺乏大型数据集. 当前的研究关注将不同模态的特征转换到统一的鸟瞰图特征下进行融合.

Wang等人[91]提出4D雷达和激光雷达的多模态多尺度融合方法(Multi-Modal and Multi-Scale Fusion,M^2-Fusion),包括基于交互的多模态融合(Interaction-based Multi-Modal Fusion,IMMF)和基于中心点的多尺度融合(Center-based Multi-Scale Fusion,CMSF)技术. IMMF模块利用自注意力机制从不同模态中学习特征并交换中间层信息,CMSF模块将预测目标的中心点周围划分成不同尺度后进行叠加,从而提取出多尺度特征,系统能够在不同距离和分辨率下对目标进行有效检测,提高检测的可靠性和稳定性. 然而,该模型复杂度高,计算量较大.

Zheng等人[92]提出基于4D雷达和相机融合的3D目标检测方法(Radar Camera Fusion,RCFusion). 该方法使用正交特征变换对图像像素进行采样,将2D图像投影到3D体素中,并利用共享注意力编码器生成鸟瞰图特征,与4D雷达点云的鸟瞰图特征融合后送入检测器. 然而,该模型采用基于锚点的检测头对小目标检测精度欠佳. Xiong等人[93]提出无需激光雷达的精简模型(LiDAR eXcluded Lean,LXL),利用预测图像深度分布图和雷达3D占用网格来辅助图像视图变换,将2D图像特征转换为3D特征,并与4D雷达特征融合用于检测,显著提高了检测性能. 然而,该方法模态间的自适应交互能力有待提高.

图9展示了近年来雷达/4D雷达与相机/激光雷达融合的3D目标检测方法在目标检测领域的部分算法研究及应用场景分类,表7总结了文中所提到的目标检测方法特点.

2.2.2 融合位置

融合位置是指在信息融合过程中,将来自不同阶段的数据进行整合的过程. 这一过程通常包括输入阶段、特征阶段和预测阶段. 如图10所示,根据融合的时间点不同,可以将融合位置划分为前期融合、中期融合和后期融合,表8总结了不同融合位置方式的特点,以及上述融合方式涉及的典型算法优缺点.

(1)前期融合

前期融合是指在传感器数据进入目标检测算法前,将多个传感器或多个模态的信息融合成同一个模态. 这种融合方法将不同传感器或模态的数据进行组合,生成一个多模态的输入,然后将其输入到目标检测网络中. 点云的投影输入形式通常属于前期融合的范畴,通过投影将多模态的数据融合生成单一模态的输入. 前期融合可以更好地利用模态间的交互信息,但可能会导致数据维度增加和信息冗余. Chen等人[61]提出将点云投影到鸟瞰图和前视图,将其作为融合输入进行特征提取和特征交互的方法,以综合考虑不同视角

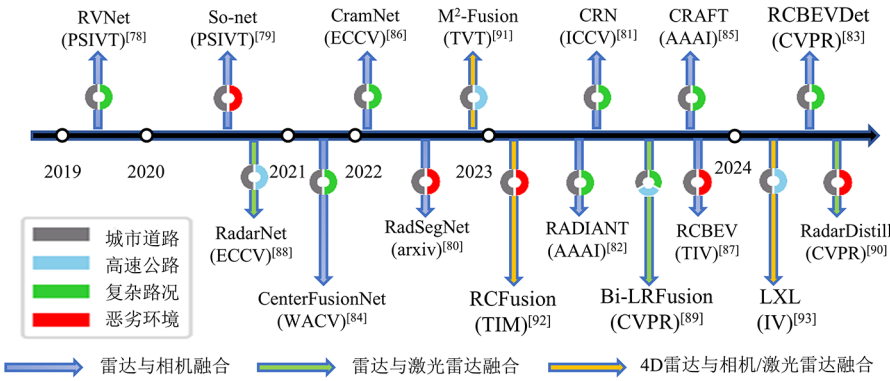

图9　雷达/4D雷达与相机/激光雷达融合的3D目标检测方法

表7　雷达/4D雷达与相机/激光雷达融合的部分目标检测算法总结

类型	每类方法的特征	作者	典型算法	典型算法的优缺点
雷达与相机融合	特点:结合雷达的距离与速度信息和相机的丰富视觉特征进行融合. 优点:成本低,充分利用雷达的深度信息和相机的彩色信息. 缺点:雷达与图像的分辨率不一致,具有较多噪声,使其与图像的关联性存在偏差,影响检测精度.	Bansal 等人	RadSegNet[80]	优点:将图像的语义信息作为额外的特征通道来增强雷达特征,在恶劣条件下仍具有可靠性. 缺点:噪声会影响性能.
		Kim 等人	CRN[81]	优点:利用稀疏的雷达点克服图像中空间信息的不足,设计多模态可变形注意力模块解决空间失调. 缺点:传感器故障情况下性能降低明显,鲁棒性需进一步增强.
		Long 等人	RADIANT[82]	优点:对雷达深度与图像预测深度进行关联加权,提高深度估计准确性. 缺点:未考虑物体大小和运动参数影响,性能有待优化.
		Lin 等人	RCBEVDet[83]	优点:利用RCS作为目标大小的先验信息增加雷达鸟瞰图特征的密度,丰富雷达特征. 缺点:图像分支的处理较简单,可能会影响检测精度.
		Nabati 等人	CenterFusion[84]	优点:创建3D感兴趣区域视锥体过滤掉无关的雷达点云,提高数据关联准确性. 缺点:两个目标中心点重合时,该网络不能有效区分.
		Kim 等人	CRAFT[85]	优点:利用软极坐标关联方法将图像建议与雷达点关联,提高数据关联准确性. 缺点:检测精度与激光雷达相比仍有差距.
雷达与激光雷达融合	特点:结合激光雷达高精度点云和雷达点云以及速度信息进行融合. 优点:结合激光雷达的密度点云与雷达的高度信息,提高检测精度. 缺点:数据噪声、目标形状及运动会影响性能.	Yang 等人	RadarNet[88]	优点:将雷达的几何信息和速度信息融入激光雷达,提高了检测精度. 缺点:没有解决雷达固有的高度信息不准确问题和稀疏性问题.
		Wang 等人	Bi-LRFusion[89]	优点:利用激光雷达数据增强雷达数据,缓解高度信息缺失和极端稀疏性带来的问题. 缺点:雷达原本的噪声特征可能会影响融合性能.
		Bang 等人	RadarDistill[90]	优点:引入蒸馏技术,通过激光雷达指导增强雷达数据. 缺点:推理阶段只依赖于雷达数据,可能会限制模型在多样化数据场景下的适应能力.
4D雷达与相机/激光雷达融合	特点:结合4D雷达的点云信息以及额外的高度信息,与相机2D特征或激光雷达点云特征融合. 优点:引入图像的语义信息或激光雷达的几何信息,丰富4D雷达点云特征,提高检测性能. 缺点:缺乏大规模数据集.	Wang 等人	M²-Fusion[91]	优点:多尺度融合,系统能够在不同距离和分辨率下对目标进行有效检测,提高检测的可靠性和稳定性. 缺点:模型复杂度高,影响推理速度.
		Zheng 等人	RCFusion[92]	优点:直接融合预测的图像鸟瞰图特征和雷达鸟瞰图特征,模型简单高效. 缺点:基于锚点的检测头对小目标检测精度欠佳.
		Xiong 等人	LXL[93]	优点:引入雷达3D占用网格作为辅助信息,提高图像的视图转换精度,增强融合特征的性能. 缺点:模态间的自适应交互能力有待提高.

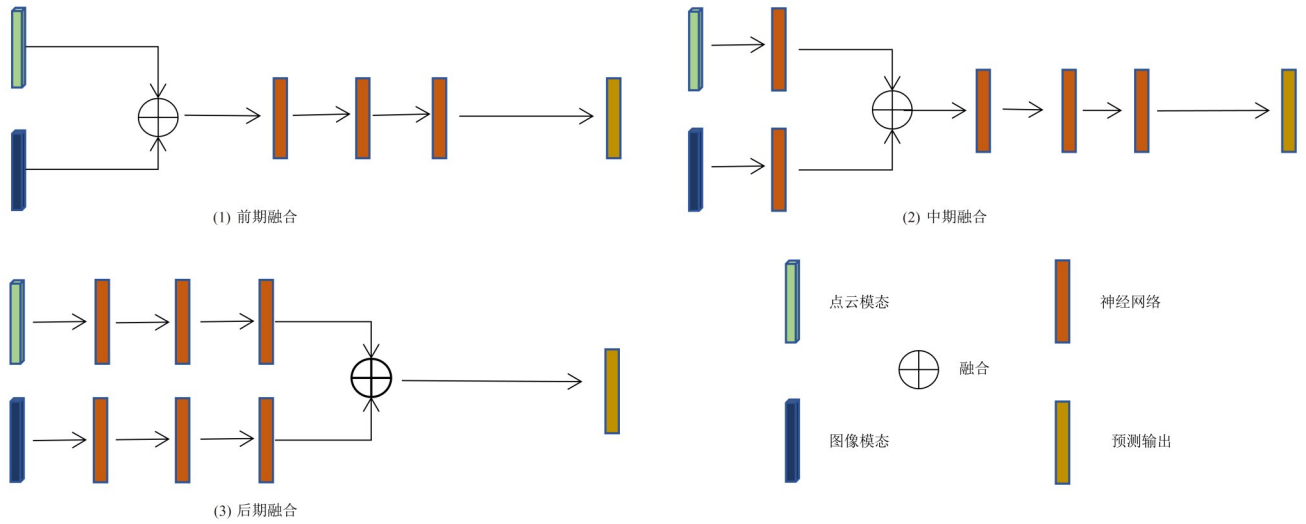

图 10 融合位置类别

表 8 不同融合方式的优缺点以及文献方法特点

类型	该类方法的特征	作者	典型算法	典型算法的优缺点
前期融合	特点:将多个传感器或多个模态的数据信息融合成同一个模态. 优点:具有丰富的模态交互信息. 缺点:数据维度增加和信息冗余.	Chen 等人	MV3D[61]	优点:利用点云投影的方法,在降低数据维度的同时保留了关键信息. 缺点:融合信息较少,小目标检测精度低.
		Wang 等人	PointAugmenting[76]	优点:图像信息与点云信息叠加,增强点云信息的表征能力. 缺点:检测性能受点云图像对齐影响.
中期融合	特点:在目标检测算法的中间阶段将多个传感器或模态的信息进行融合. 优点:保留丰富的上下文信息和语义关联,融合检测效果相对较好. 缺点:不同模态间依赖性大,缺少灵活性.	Liang 等人	BEVFusion[67]	优点:各模态单独处理生成鸟瞰图特征,保证了系统稳定. 缺点:鸟瞰图存在语义模糊的问题,可能会影响精度.
		Li 等人	UVTR[68]	优点:将各模态的特征转换为统一体素特征表示进行融合,消除鸟瞰图带来的语义不确定性. 缺点:利用独立主干网络处理图像和点云,容易丢失细节信息.
后期融合	特点:在目标检测算法的预测阶段将多个传感器或模态的检测结果进行融合. 优点:灵活性强,复杂度低,处理速度快. 缺点:无法利用中间特征,检测性能欠佳.	Pang 等人	CLOCs[94]	优点:结合几何信息和语义信息融合预测框,综合利用不同维度的信息,提高检测精度. 缺点:性能受数据关联质量影响.
		Dong 等人	Association-Net[95]	优点:通过将雷达针和图像预测框投影到图像平面生成伪图像,提高数据关联准确性,具有低成本和高可靠性. 缺点:在远距离目标关联准确性有待提高.

的信息,在降低数据维度的同时保留了关键信息,但该方法的融合信息较少,小目标检测精度低. 与此不同,Wang 等人[76]基于跨模态增强的 PointAugmenting 算法,直接将图像数据信息与点云信息叠加,从而增强了点云信息的表征能力,但点云与图像的对齐准确性很大程度上影响了模型整体检测性能.

(2)中期融合

中期融合是指在目标检测算法的中间阶段将多个传感器或模态的信息进行融合,中间阶段包括特征提取阶段和预测阶段前. 中期融合可以在保留模态特性的同时,充分利用多模态信息,使得中间层的多模态特征之间能够进行更直接的交互,并且可以适应不同模态数据的特点,保留丰富的上下文信息和语义关联,是目前应用最广泛的融合阶段. Liang 等人[67]对每个模态单独处理提出了 BEVFusion 算法,针对点云主干网络和图像主干网络得到点云特征和图像特征,在统一的鸟瞰图下进行融合. 该方法通过独立处理和融合摄像头与激光雷达数据,即使在激光雷达或摄像头失效的情况下也能确保系统的稳健性,但鸟瞰图存在语义模糊的问题. Li 等人[68]将图像数据升级至体素空间,提出了 UVTR 将点云和图像的特征映射到统一的体素特征空间中实现特征融合,以消除鸟瞰图带来的语义不确定性. 然而,该方法利用独立的主干网络处理图像和点云,容易丢失细节信息.

(3)后期融合

后期融合是指在目标检测算法的预测阶段将多个

传感器或模态的检测结果进行融合. 这种融合方法为每个模态采用单独的分支,不需要处理不同模态的数据异构问题,可以更好地利用每种模态的现有网络. 后期融合对每个模态独立进行完成任务,当某一模态缺失时,不会对整体产生破坏性结果,并且结构复杂度低,具有较快处理速度,但由于无法利用丰富的中间特征,性能相较中期融合有一定差距.

Pang等人[94]提出融合相机和激光雷达候选目标的检测算法(Camera-LiDAR Object Candidates, CLOCs),该方法结合激光雷达和相机两种预测结果的几何特性和语义特性获得最终的目标检测框,以综合利用不同维度的信息,但其检测结果可能会受到数据关联影响. Dong等人[95]提出基于变换矩阵的AssociationNet方法,通过将雷达针和图像预测框投影到图像平面生成伪图像,输入神经网络来学习高级语义表示,解决数据关联问题,具有低成本和高可靠性. 然而,该方法在远距离目标关联准确性有待提高.

2.2.3　多模态检测算法小结

本节详细介绍了多模态检测中的传感器融合、输入类别融合以及位置融合. 传感器融合涉及相机、激光雷达和3D/4D雷达等设备,每种传感器均具有其固有的优缺点,通过结合不同的传感器,可以综合利用各自的优势. 输入类别的融合包括鸟瞰图投影点云、原始点集和体素化点云. 位置融合则分为前期融合、中期融合和后期融合. 不同的融合策略各有其优缺点,选择适当的融合方式应根据具体的应用场景进行权衡.

3　3D目标检测的常用数据集及训练策略

3.1　自动驾驶中常用数据集及评价指标

3.1.1　KITTI

KITTI数据集[96]的数据采集平台装配有2个灰度摄像机,2个彩色摄像机,1个64线激光雷达,4个光学镜头,以及1个GPS导航系统. KITTI数据集针对3D目标检测任务提供了14 999组图像以及对应的点云数据,其中7 481组用于训练,7 518组用于测试,并针对场景中的汽车、行人、自行车3类目标进行标注,共计80 256个标记对象.

KITTI数据集使用平均精度(Average Precision, AP)作为评价指标,并根据检测难易程度分为简单、适度和困难3个级别. 其官方评估方法设置了40个相等间隔的召回率点$r\in\{1/40, 2/40, \cdots, 1\}$,利用$\mathrm{AP}|_R$插值方法计算AP值,公式如下:

$$\mathrm{AP}|_R=\frac{1}{|R|}\sum_{r\in R}\rho\mathrm{interp}(r) \tag{5}$$

其中,$|R|$代表采样点的总数,$\rho\mathrm{inter}(r)$遍历每个召回率点r,取大于等于当前召回率点的所有召回率点对应的精度值.

3.1.2　nuScenes

nuScenes数据集[97]收集了1 000个驾驶场景,采集平台装配有1个激光雷达,5个雷达传感器和6个相机. 完整的数据集涵盖约140万张相机图像、39万次激光雷达扫描、140万次雷达扫描和4万个关键帧中的140万个对象边界框.

nuScenes数据集使用nuScenes检测评分(Nuscenes Detection Score, NDS)作为评价测定的指标:

$$\mathrm{NDS}=\frac{1}{10}\left[5\mathrm{mAP}+\sum_{\mathrm{mTP}\in\mathrm{TP}}\left(1-\min(1,\mathrm{mTP})\right)\right] \tag{6}$$

其中,mAP代表平均精密度,其公式如下:

$$\mathrm{mAP}=\frac{1}{|C||D|}\sum_{c\in C}\sum_{d\in D}\mathrm{AP}_{c,d} \tag{7}$$

其中,C表示目标类集合,D表示中心距离集合,$|C|$和$|D|$分别表示该集合元素的数量,$\mathrm{AP}_{c,d}$表示类别c和距离d下的平均精度.

mTP受到5个TP指标影响:(1)平均平移误差(Average Translation Error, ATE);(2)平均尺度误差(Average Scale Error, ASE);(3)平均方向误差(Average Orientation Error, AOE);(4)平均速度误差(Average Velocity Error, AVE);(5)平均属性误差(Average Attribute Error, AAE). mTP公式如下:

$$\mathrm{mTP}=\frac{1}{|C|}\sum_{c\in C}\mathrm{TP}_c \tag{8}$$

其中,TP_c表示类别c下的真阳性(True Positives, TP)数量. 前文的TP指标通过影响TP数量,间接影响了mTP的结果.

根据式(6)的表达式,NDS被构建为一个综合性的指标,其计算结果一半基于检测性能,另一半基于位置、大小、方向、属性和速度度量的检测质量.

3.1.3　Waymo

Waymo数据集[98]的数据采集平台装配有1个中程激光雷达,4个短程激光雷达和5个摄像头. 总共有798个场景用于训练,202个场景用于验证,场景由5个激光雷达传感器和5个针孔相机收集. Waymo数据集采用航向加权平均精度(Average Precision weighted by Heading, APH)用作检测度量:

$$\mathrm{APH}=100\int_0^1\max\left\{h(r')|r'\geq r\right\}\mathrm{d}r \tag{9}$$

其中,$h(r)$的计算类似于P-R曲线,在P-R曲线的基础上引入航向信息,使用的真阳性(TP)值由航向加权得到的. Waymo数据集分为2个难度级别:LEVEL_1和LEVEL_2,LEVEL_1适用于至少包含有5个激光雷达信号的锚框,LEVEL_2适用于所有非空的锚框.

3.1.4　4D毫米波雷达数据集

View-of-Delft(VoD)数据集[99]和TJ4DRadSet数据

集[100]是两个新颖的车载数据集,涵盖了4D雷达数据.VoD是在复杂的城市交通场景中采集的,包含8 693帧时间同步且校准好的64线激光雷达、双目相机和4D雷达数据,标注了123 106个运动和静态目标的3D边界框,涉及26 587个行人、10 800个骑自行车的人和26 949辆汽车.TJ4DRadSet囊括44个序列中的7 757帧激光雷达、摄像机和4D雷达数据.这两个数据集采用了和KITTI数据集相同的评估指标.

3.1.5 自动驾驶数据集小结

随着自动驾驶技术的快速发展,数据集的规模和多样性将持续增加.通过表9的数据,我们可以看出目前自动驾驶数据集的规模已从15 000帧扩展到超过200 000帧,为3D目标检测模型的开发提供了丰富的支持.

3.2 训练策略

训练3D目标检测模型需要经历多个阶段,包括数据准备、网络架构选择、数据增强、损失函数定义、优化器选择、学习率调度、模型优化、验证集调优、测试集评估等步骤.本节重点介绍数据增强和优化技术,它们提供了训练过程中优化模型性能的关键信息,表10总结了文中所提到的训练策略.此外,本节深入对比了多种训练策略在ScanObjectNN数据集上的性能展现.ScanObjectNN数据集不专用于车载场景,但相较于前述的自动驾驶数据集,它提供了更为真实且多样化的物体扫描数据.这种特性有助于模型在真实世界的复杂场景中实现更好的适应性.此外,该数据集包含丰富的标注信息,能够赋予模型更细粒度的理解能力.因此,利用该数据集进行不同训练策略的对比,能够更直观地观察到精度变化的细微差别.表11在Pointnet++算法[29]上验证了部分训练策略的可行性[33].

表9 3D目标检测常用数据集

数据集	年份	激光雷达数量	激光雷达通道数量	雷达	4D雷达	相机数量	帧数
KITTI	2012	1	64	无	无	4	15K
nuScenes	2019	1	32	有	无	6	40K
Waymo	2019	5	64	无	无	5	200K
VoD	2022	1	64	无	有	1	8 693
TJ4DRadSet	2022	1	64	无	有	1	7 757

表10 训练策略总结

	方法	特点
数据增强	随机抽取颜色	增加数据的多样性
	公共点重采样	减少计算复杂度并保持信息的完整性
	场景分割	考虑场景上下文信息
	随机抖动	增强点云数据多样性
	附加高度	丰富数据信息
损失函数	平滑L1损失	用于目标检测中的边界框回归
	交叉熵损失	用于衡量模型预测的概率分布与真实标签分布之间的差异
	IoU损失	用于度量预测边界框与真实边界框之间的重叠程度
优化器	标签平滑	将硬标签转化为软标签,减少模型对硬标签的过度依赖
	Adam	结合动量梯度下降自适应地调整学习率
	RMSprop	结合平滑梯度的平方值调整学习率
	AdamW	在Adam算法上增加权重衰减项
学习率调度策略	学习率衰减	在训练过程中逐渐降低学习率
	学习率周期调度	训练过程分为多个阶段,每个阶段内保持学习率不变
	阶梯学习率衰减	在训练的特定时间点或迭代次数,学习率乘以预设的衰减因子
	余弦退火	通过余弦函数调整学习率
	余弦学习率衰减	将学习率按照余弦函数的形式进行衰减

表11 部分训练策略在Pointnet++上的精度 单位:%

方法	Pointnet++(原始网络)	增加公共点重采样	去除随机抖动	增加附加高度参数	增加随机缩放	标签平滑	Adam→AdamW	步骤衰减→余弦衰减
精度	77.9	81.4 (↑2.5)	82.5 (↑1.1)	83.6 (↑1.1)	83.7 (↑0.1)	85.0 (↑1.3)	85.6 (↑0.6)	86.1 (↑0.5)

3.2.1 数据增强

数据增强是提高模型鲁棒性的重要手段,常用的数据增强方法包括旋转、平移、缩放、翻转、抖动等.在一些较新的研究中,Thomas等人[101]在训练过程中通过随机抽取颜色来增加数据的多样性,以提高模型的鲁棒性.Yu等人[102]则采用了公共点重采样策略,从原始

点云中随机选择1 024个点进行数据缩放,以减少计算复杂度并保持信息的完整性. 相比之下,Zhao等人[103]则将整个场景作为分割任务的输入,以更全面地考虑场景的上下文信息. 此外,还可以通过点云随机抖动、附加高度参数等[33]来丰富数据信息.

3.2.2　优化技术

优化技术包括损失函数、优化器、学习率调度器等. 本节列举了一些常用的优化技术,在进行选择和调整时,应根据具体的任务需求和模型特性进行合理的决策,以取得优异的检测结果.

常用的损失函数包括平滑L1损失[104]、交叉熵损失[105]和IoU(Intersection over Union)损失[106]等. 其中,平滑L1损失结合了L1损失和L2损失优点,在误差较小时表现为L2损失(平方损失),有助于模型稳定收敛;在误差较大时表现为L1损失(线性损失),避免了梯度爆炸问题,被广泛用于边界框位置的回归任务. 交叉熵损失用于衡量模型预测的概率分布与真实标签分布之间的差异,通过最小化预测分布与真实分布之间的交叉熵来优化模型,鼓励模型输出更"确定"的分类结果,适用于多分类和二分类任务. IoU损失则用于度量预测边界框与真实边界框之间的重叠程度,常用于检测任务中以提高边界框的预测准确性.

常用优化器包括标签平滑(Label Smoothing)[107]、Adam(Adaptive moment estimation)[108]、RMSprop(Root Mean Square propagation)[109]、AdamW(Adam with Weight decay)[110]等. 其中,标签平滑是一种正则化技术,通过将硬标签转化为软标签,减少模型对硬标签的过度依赖,提高模型的泛化能力. Adam和RMSprop是自适应学习率的优化算法. RMSprop通过平滑梯度的平方值来调整学习率,可以有效地处理不同参数的尺度问题. Adam结合了动量梯度下降,可以自适应地调整每个参数的学习率,具有较快的收敛速度和较好的泛化性能. AdamW是对Adam算法的改进版本,通过增加权重衰减项来解决Adam算法在权重衰减方面的缺陷.

常用学习率调度器包括学习率衰减(Learning Rate Decay)、学习率周期调度(Learning Rate Scheduling with Periods)、阶梯学习率衰减(Step Learning Rate Decay)、余弦退火(Cosine Annealing)、余弦学习率衰减(Cosine Learning Rate Decay)等. 学习率衰减是一种基本的学习率调度策略,通过在训练过程中逐渐降低学习率来帮助模型收敛. 学习率周期调度将训练过程分为多个阶段,在每个阶段内保持学习率不变,但在不同阶段之间调整学习率,帮助模型在不同阶段学习到不同的特征. 阶梯学习率衰减是一种定期调整学习率的方法. 在训练的特定时间点或迭代次数,学习率会乘以一个预设的衰减因子,从而实现学习率的阶梯式下降. 余弦退火是一种通过余弦函数调整学习率的策略,学习率在训练过程中先快速下降,然后缓慢下降寻求最优解. 余弦学习率衰减结合了余弦函数的特性和学习率衰减的思想,将学习率按照余弦函数的形式进行衰减,在训练后期能够更加平滑地调整学习率.

3.3　代表性算法的性能对比

nuScenes数据集是当前自动驾驶数据集中场景丰富、传感器较为齐全的大规模数据集之一. 我们选取了部分代表性的算法在该数据集中进行测试和比较. 表12列出了近几年来在不同传感器输入下的典型算法的性能,大部分算法都在一台搭载NVIDIA GeForce RTX 3090 GPU的工作站上实现. 观察结果显示,单模态检测算法中激光雷达性能最佳,相机次之,毫米波雷达表现最差,这主要归因于雷达点云的稀疏性和噪声对目标检测的显著影响. 当相机与雷达进行融合后,相机能够提供更丰富的视觉信息,弥补了雷达的不足,从而使算法性能得到一定的提升,但由于雷达点云的稀疏性和噪声,与基于激光雷达的检测算法仍有差距. 基于激光雷达与相机融合的算法融合了激光雷达精确的几何信息和相机丰富的语义信息,性能是目前最佳的多模态融合算法. 可见,基于融合的方法在点云相较稀疏时具有较大的感知能力提升.

表12　3D目标检测代表性算法客观指标对比

模型	传感器	NDS	mAP	源码链接
Radar-PointGNN[51]	雷达	0.034	0.005	未提供源码,直接引用论文数据
BEVFormer[20]	相机	0.515	0.412	https://github.com/ZrrSkywalker/MonoDETR
VoxelNeXt[42]	激光雷达	0.666	0.605	https://github.com/dvlab-research/VoxelNeXt
CenterFusion[84]	雷达+相机	0.452	0.331	https://github.com/mrnabati/CenterFusion
BEVFusion[67]	激光雷达+相机	0.714	0.685	https://github.com/mit-han-lab/bevfusion

4　思考与展望

综上所述,传统的单模态方法由于仅依赖于单一传感器提供的数据,在目标检测任务中存在一定的限制. 除了基于高线程激光雷达(例如64线激光雷达)的方法外,其他常见传感器的检测精度往往不及多模态方法. 多模态方法的优势在于能够综合利用不同传感

器的特点和数据,从而更好地解决单模态方法所面临的问题. 近年来,自动驾驶中的3D目标检测研究取得了显著进展,但仍存在一些挑战值得关注,未来该领域将面临的一些新问题也有待思考.

4.1 当前的研究热点

4.1.1 多传感器融合的数据对齐

不同传感器具有不同的工作原理、感知范围和数据表示方式,其精度、噪声和采样率等差异使得数据具有异构性. 这种数据异构性可能导致数据在空间、时间和特征空间上不一致,影响融合语义对齐的效果,减低融合检测精度. 目前常用的方法是使用校准矩阵来对齐传感器之间的数据,并通过单个传感器与车辆坐标系进行校准以提高一致性. 然而,车辆在行驶过程中会产生颠簸抖动等情况,导致传感器的外部参数发生变化. 如果不及时纠正,这些误差可能会不断积累,最终影响检测结果. 因此,如何实现不同数据源之间的语义统一,以提高多模态融合结果的质量和可信度,是当前的研究热点之一.

4.1.2 低成本3D目标检测

相机和雷达是常见且成本较低的传感器,但相机缺乏3D几何信息,而雷达点云又过于稀疏,以致无法获得与激光雷达相当的检测结果,这限制了它们在精确3D检测中的应用. 相比之下,4D雷达提供了更密集的点云,并增加了高度信息. 此外,4D雷达相对于激光雷达具有更低的成本,受外部环境影响较小,并且具备独特的速度测量和全天候传感能力. 因此,低成本的4D毫米波雷达结合视觉传感器的检测方式将是值得关注的热点之一,具有广阔的应用前景.

4.1.3 降低融合信息损失

在多模态数据融合的过程中,由于投影、量化等操作的存在,不可避免地会导致一定程度的信息丢失. 例如,后期融合可能会缺乏早期阶段的特征信息,点云体素化会丢失细节信息,点云的鸟瞰图投影则会丢失高度信息,而前视图投影则可能引起尺度问题. 为了解决这些问题,可以采用多种输入形式进行融合,并仔细考量融合操作的切入时机,以尽可能减少信息的损失. 然而,这种方法会增加算法的复杂性和计算成本. 因此,如何在受限的算力资源下最小化融合过程中的信息损失,仍然是一个值得深入研究的方向.

4.1.4 跨模态数据增强

数据增强是3D目标检测中的重要环节,目前数据增强主要应用于单模态方法,对多模态场景的考虑相对较少. 由于点云和图像是两种异构数据,在实现跨模态同步增强时可能导致严重的跨模态错位问题和失真问题. 现有方法主要有三种:一种方法仅对点云部分进行增强,而忽略图像部分;第二种方法保持原始图像不变,在点云中进行逆变换,以实现图像点云对应;第三种方法则通过重建,将异构数据转换为统一的表示,以实现同步数据增强. 然而,这些方法都不能很好地解决问题,如何在点云和相机数据上同步应用数据增强而不引入错位和失真仍是一项具有挑战性的任务.

4.1.5 半监督学习的高效应用

标注数据稀缺是限制3D目标检测发展的重要因素. 传统的全监督学习需要大量标注数据,而无监督方法虽然降低了数据标注成本,但在实际应用中效果仍不理想. 相比之下,半监督学习能够充分利用有限标注数据,并结合未标注数据生成伪标签,有效平衡标注成本和检测性能. 目前研究热点主要集中在伪标签生成的可靠性、知识蒸馏、领域自适应等方面.

4.2 未来面临的新挑战

4.2.1 国内自动驾驶数据集

当前,主流的大型自动驾驶数据集主要基于国外的道路环境进行采集. 例如,nuScenes数据集是在新加坡和波士顿采集的,KITTI数据集是在德国采集的,而Waymo数据集则是在美国的多个城市采集. 这些数据集在国际上为自动驾驶技术的研究和开发提供了重要的基础. 然而,由于国内道路环境与国外存在显著差异,包括交通规则、驾驶习惯、道路基础设施等方面的不同,这些国外数据集在国内应用时可能面临适用性和有效性的问题. 因此,针对国内特定的交通环境和需求,开发和构建本土化的大型自动驾驶数据集显得尤为重要. 这不仅能够提升自动驾驶技术在国内的适应性和安全性,还能推动相关技术的本地化创新和发展,将成为未来研究和应用的重要方向之一.

4.2.2 动态场景中的实时多模态融合

动态场景下,传感器的时间戳偏移、数据稀疏性和运动误差显著增加了多模态数据融合的难度. 例如,高速行驶的车辆可能导致相机捕捉到的画面和激光雷达的点云数据存在时间上的不同步,而颠簸的路面会引入外参漂移,从而影响传感器数据的对齐和融合. 目前的研究大多集中在静态场景下的传感器对齐与融合,而动态场景中的实时多模态融合方法仍未成熟. 如何开发能够动态调整对齐和融合策略的算法,将是未来研究的重要方向之一.

4.2.3 复杂小目标检测

在远距离或复杂背景下,小目标(如宠物、儿童等)的检测精度仍是一个重大挑战. 稀疏点云和有限的几何细节限制了现有算法对小目标的检测性能,同时复杂背景可能增加误检的概率. 目前已有研究尝试通过增加样本数据量或改进网络结构来提升小目标检测精度,但仍存在一些难点,例如稀疏数据的补全、遮挡处理以及多尺度融合,将是未来值得探讨的热点方向.

4.2.4　道路静态目标检测

在自动驾驶道路场景中,除了车辆、行人等动态目标外,还存在一些可能造成误检的静态目标,如广告牌. 例如,2024年5月,新浪财经报道了一辆理想汽车在高速上开启了自动辅助驾驶系统,因误检到广告牌上的小货车图片突然急刹而导致追尾事故. 该案例表明,前向毫米波雷达可能对广告牌上的金属材质过于敏感,也可能是由于感知设备过多导致感知系统接收到干扰项而产生误检. 目前的解决方法包括引入鸟瞰图技术或者高度检测,但这些方法仍然存在一定改进空间,值得深入研究.

4.2.5　复杂环境下的精确检测

在自动驾驶领域,不同环境场景的复杂性对3D目标检测提出了更高要求. 目前主流算法在晴朗天气下的道路场景中表现良好,但在特殊场景(如雨天或复杂道路)条件下,缺乏有效处理策略,容易造成误检. 例如,据新浪财经报道,2023年1月,一位车主在雨夜驾车出行时,中控屏显示有人在后方奔跑追车,实际后方并没有行人. 同年5月,另一位车主在陵园祭祖时发现车内雷达显示车辆周边有人影,但实际上车外并没有行人经过. 目前的解决方法包括提升传感器的感知性能,或者通过引入更多负样本数据进行算法训练. 然而,仍存在许多未知场景,如何有效应对这一问题还需要进一步研究.

5　结论

3D目标检测在自动驾驶等应用中的重要性日益增加,本文回顾了近几年3D目标检测领域的研究进展,主要从单模态检测和多模态融合检测两个方面展开讨论. 在单模态检测方面,综合考量了不同传感器及其检测算法的优劣之处. 在多模态融合检测方面,则从融合传感器、融合类别以及融合位置等多个角度对多模态融合进行了探讨. 此外,本文深入总结了自动驾驶常用的几个常用数据集,在nuScenes公共数据集上对部分算法的性能进行了比较分析,并从数据增强和优化技术两个角度出发,总结当前常用的训练策略. 最后,探讨了3D目标检测领域面临的挑战以及未来的研究趋势,涉及多模态融合、半监督检测、复杂环境检测自动驾驶数据集等.

参考文献

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

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

[3] MA X Z, OUYANG W L, SIMONELLI A, et al. 3D object detection from images for autonomous driving: A survey[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 46(5): 3537-3556.

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

[5] WANG K, ZHOU T Q, LI X C, et al. Performance and challenges of 3D object detection methods in complex scenes for autonomous driving[J]. IEEE Transactions on Intelligent Vehicles, 2023, 8(2): 1699-1716.

[6] 葛同澳, 李辉, 郭颖, 等. 基于双融合框架的多模态3D目标检测算法[J]. 电子学报, 2023, 51(11): 3100-3110.
GE T A, LI H, GUO Y, et al. A multimodal 3D object detection method based on double-fusion framework[J]. Acta Electronica Sinica, 2023, 51(11): 3100-3110. (in Chinese)

[7] 周治国, 马文浩. 一种多层多模态融合3D目标检测方法[J]. 电子学报, 2024, 52(3): 696-708.
ZHOU Z G, MA W H. 3D object detection based on multilayer multimodal fusion[J]. Acta Electronica Sinica, 2024, 52(3): 696-708. (in Chinese)

[8] XU B, CHEN Z Z. Multi-level fusion based 3D object detection from monocular images[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 2345-2353.

[9] DUAN K W, BAI S, XIE L X, et al. CenterNet: Keypoint triplets for object detection[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2019: 6569-6578.

[10] MA X Z, ZHANG Y M, XU D, et al. Delving into localization errors for monocular 3D object detection[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2021: 4719-4728.

[11] ZHANG Y P, LU J W, ZHOU J. Objects are different: Flexible monocular 3D object detection[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2021: 3288-3297.

[12] WANG T, ZHU X G, PANG J M, et al. FCOS3D: Fully convolutional one-stage monocular 3D object detection[C]//2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW). IEEE, 2021: 913-922.

[13] LIU X P, XUE N, WU T F. Learning auxiliary monocular contexts helps monocular 3D object detection[J]. Proceedings of the AAAI Conference on Artificial Intelligence,

2022, 36(2): 1810-1818.

[14] YAN L F, YAN P, XIONG S Z, et al. MonoCD: Monocular 3D object detection with complementary depths[C]//2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2024: 10248-10257.

[15] RODDICK T, KENDALL A, CIPOLLA R. Orthographic feature transform for monocular 3D object detection[EB/OL]. (2018-11-20)[2025-02-26]. https://arxiv.org/pdf/1811.08188.

[16] BRAZIL G, LIU X M. M3D-RPN: Monocular 3D region proposal network for object detection[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE, 2019: 9287-9296.

[17] READING C, HARAKEH A, CHAE J L, et al. Categorical depth distribution network for monocular 3D object detection[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2021: 8555-8564.

[18] KUMAR A, BRAZIL G, CORONA E, et al. DEVIANT: Depth equivariant network for monocular 3D object detection[C]// Computer Vision-ECCV 2022. Cham: Springer Nature Switzerland, 2022: 664-683.

[19] HUANG K, WU T, SU H, et al. MonoDTR: Monocular 3D object detection with depth-aware transformer[C]// 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2022: 4002-4011.

[20] LI Z Q, WANG W H, LI H Y, et al. BEVFormer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers[M]//Computer Vision-ECCV 2022. Cham: Springer Nature Switzerland, 2022: 1-18.

[21] WANG Z Y, LI D W, LUO C X, et al. DistillBEV: Boosting multi-camera 3D object detection with cross-modal knowledge distillation[C]//2023 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2023: 8603-8612.

[22] TAO R Z, HAN W C, QIU Z Y, et al. Weakly supervised monocular 3D object detection using multi-view projection and direction consistency[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2023: 17482-17492.

[23] ZHANG R R, QIU H, WANG T, et al. MonoDETR: Depth-guided transformer for monocular 3D object detection[C]// 2023 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2023: 9121-9132.

[24] LI Z L, XU X G, LIM S, et al. UniMODE: Unified monocular 3D object detection[C]//2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2024: 16561-16570.

[25] 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). Piscataway: IEEE, 2019: 8437-8445.

[26] YOU Y, WANG Y, CHAO W L, et al. Pseudo-lidar++: Accurate depth for 3D object detection in autonomous driving[C]//International Conference on Learning Representations. Britain: ML Research Press, 2020: 1-22.

[27] PANDHARIPANDE A, CHENG C H, DAUWELS J, et al. Sensing and machine learning for automotive perception: A review[J]. IEEE Sensors Journal, 2023, 23(11): 11097-11115.

[28] CHARLES R Q, 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). Piscataway: IEEE, 2017: 77-85.

[29] QI C R, YI L, SU H, et al. Pointnet++: Deep hierarchical feature learning on point sets in a metric space[C]//Advances in Neural Information Processing Systems. Virtual: PMLR, 2017: 5099-5108.

[30] 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). Piscataway: IEEE, 2020: 11040-11048.

[31] 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). Piscataway: IEEE, 2021: 7546-7555.

[32] SHENGA 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). Piscataway: IEEE, 2021: 2723-2732.

[33] QIAN G, LI Y, PENG H, et al. PointNext: Revisiting pointnet++ with improved training and scaling strategies[C]//Advances in Neural Information Processing Systems. Virtual: PMLR, 2022: 23192-23204.

[34] HUANG K, LYU W J, YANG M, et al. PTT: Point-trajectory transformer for efficient temporal 3D object detection[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2024: 14938-14947.

[35] 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. Piscataway: IEEE, 2018: 4490-4499.

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

[37] 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.

[38] 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). Piscataway: IEEE, 2022: 8407-8417.

[39] CHEN Y K, LI Y W, ZHANG X Y, et al. Focal sparse convolutional networks for 3D object detection[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2022: 5418-5427.

[40] LIU L, SONG Z Y, XIA Q M, et al. SparseDet: A simple and effective framework for fully sparse LiDAR-based 3-D object detection[J]. IEEE Transactions on Geoscience and Remote Sensing, 2024, 62: 5707114.

[41] KOH J, LEE J, LEE Y, et al. MGTANet: Encoding sequential LiDAR points using long short-term motion-guided temporal attention for 3D object detection[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2023, 37(1): 1179-1187.

[42] CHEN Y K, LIU J H, ZHANG X Y, et al. VoxelNeXt: Fully sparse VoxelNet for 3D object detection and tracking[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2023: 21674-21683.

[43] ZHANG G, CHEN J N, GAO G H, et al. SAFDNet: A simple and effective network for fully sparse 3D object detection[C]//2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2024: 14477-14486.

[44] 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). Piscataway: IEEE, 2019: 1951-1960.

[45] 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). Piscataway: IEEE, 2020: 11870-11879.

[46] 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). Piscataway: IEEE, 2020: 10529-10538.

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

[48] BANSAL K, RUNGTA K, ZHU S Y, et al. Pointillism: Accurate 3D bounding box estimation with multi-radars[C]//Proceedings of the 18th Conference on Embedded Networked Sensor Systems. New York: ACM, 2020: 340-353.

[49] PALFFY A, DONG J A, KOOIJ J F P, et al. CNN based road user detection using the 3D radar cube[J]. IEEE Robotics and Automation Letters, 2020, 5(2): 1263-1270.

[50] LIU J N, XIONG W Y, BAI L P, et al. Deep instance segmentation with automotive radar detection points[J]. IEEE Transactions on Intelligent Vehicles, 2023, 8(1): 84-94.

[51] SVENNINGSSON P, FIORANELLI F, YAROVOY A. Radar-PointGNN: Graph based object recognition for unstructured radar point-cloud data[C]//2021 IEEE Radar Conference. Piscataway: IEEE, 2021: 1-6.

[52] ZHANG A, NOWRUZI F E, LAGANIERE R. RADDet: Range-azimuth-Doppler based radar object detection for dynamic road users[C]//2021 18th Conference on Robots and Vision (CRV). Piscataway: IEEE, 2021: 95-102.

[53] JIANG T Z, ZHUANG L, AN Q, et al. T-RODNet: Transformer for vehicular millimeter-wave radar object detection[J]. IEEE Transactions on Instrumentation and Measurement, 2022, 72: 5003912.

[54] DECOURT C, VANRULLEN R, SALLE D, et al. DAROD: A deep automotive radar object detector on range-Doppler maps[C]//2022 IEEE Intelligent Vehicles Symposium (IV). Piscataway: IEEE, 2022: 112-118.

[55] MAJOR B, FONTIJNE D, ANSARI A, et al. Vehicle detection with automotive radar using deep learning on range-azimuth-Doppler tensors[C]//2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW). Piscataway: IEEE, 2019: 924-932.

[56] BAI J, ZHENG L Q, LI S, et al. Radar Transformer: An object classification network based on 4D MMW imaging radar[J]. Sensors, 2021, 21(11): 3854.

[57] XU B W, ZHANG X Y, WANG L, et al. RPFA-Net: A 4D RaDAR pillar feature attention network for 3D object detection[C]//2021 IEEE International Intelligent Transportation Systems Conference (ITSC). Piscataway: IEEE, 2021: 3061-3066.

[58] LIU J N, ZHAO Q C, XIONG W Y, et al. SMURF: Spatial multi-representation fusion for 3D object detection with 4D imaging radar[C]//2024 IEEE Intelligent Vehicles Symposium (IV). Piscataway: IEEE, 2024: 3141.

[59] PAN Z J, DING F Q, ZHONG H T, et al. RaTrack: Moving object detection and tracking with 4D radar point cloud[C]//2024 IEEE International Conference on Robotics and Automation (ICRA). Piscataway: IEEE, 2024: 4480-4487.

[60] PAEK D H, KONG S H, KEVIN T W. K-Radar: 4D radar object detection for autonomous driving in various weather conditions[EB/OL]. (2023-11-07) [2025-03-11]. https://arxiv.org/abs/2206.08171.

[61] 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). Piscataway: IEEE, 2017: 6526-6534.

[62] 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). New York: ACM, 2018: 1-8.

[63] 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). Piscataway: IEEE, 2022: 1080-1089.

[64] LU H H, CHEN X S, ZHANG G Y, et al. Scanet: Spatial-channel attention network for 3D object detection[C]//ICASSP 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE, 2019: 1992-1996.

[65] SINDAGI V A, ZHOU Y, TUZEL O. MVX-net: Multimodal VoxelNet for 3D object detection[C]//2019 International Conference on Robotics and Automation (ICRA). Piscataway: IEEE, 2019: 7276-7282.

[66] YOO J H, KIM Y, KIM J, et al. 3D-CVF: Generating Joint Camera and LiDAR Features Using Cross-View Spatial Feature Fusion for 3D Object Detection[M]//Computer Vision-ECCV 2020. Cham: Springer International Publishing, 2020: 720-736.

[67] LIANG T, XIE H, YU K, et al. BEVfusion: A simple and robust lidar-camera fusion framework[C]//Advances in Neural Information Processing Systems. Virtual: PMLR, 2022: 10421-10434.

[68] LI Y W, CHEN Y L, QI X J, et al. Unifying voxel-based representation with transformer for 3D object detection[C]//Proceedings of the 36th International Conference on Neural Information Processing Systems. New York: ACM, 2022: 18442-18455.

[69] JIAO Y, JIE Z Q, CHEN S X, et al. MSMDFusion: Fusing LiDAR and camera at multiple scales with multi-depth seeds for 3D object detection[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2023: 21643-21652.

[70] LI X, MA T, HOU Y N, et al. LoGoNet: Towards accurate 3D object detection with local-to-global cross- modal fusion[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2023: 17524-17534.

[71] SONG Z Y, ZHANG G X, XIE J, et al. VoxelNextFusion: A simple, unified, and effective voxel fusion framework for multimodal 3-D object detection[J]. IEEE Transactions on Geoscience and Remote Sensing, 2023, 61: 5705412.

[72] XIANG X, ZHANG J. FusionViT: hierarchical 3D object detection via lidar-camera vision transformer fusion[C]//International Conference on Learning Representations. Britain: ML Research Press, 2024: 1-16.

[73] XU D F, ANGUELOV D, JAIN A. PointFusion: Deep sensor fusion for 3D bounding box estimation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 244-253.

[74] HUANG T T, LIU Z, CHEN X W, et al. EPNet: Enhancing point features with image semantics for 3D object detection[M]//Computer Vision-ECCV 2020. Cham: Springer International Publishing, 2020: 35-52.

[75] XIE L, XIANG C, YU Z X, et al. PI-RCNN: An efficient multi-sensor 3D object detector with point-based attentive cont-conv fusion module[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(7): 12460-12467.

[76] WANG C W, MA C, ZHU M, et al. PointAugmenting: Cross-modal augmentation for 3D object detection[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2021: 11789-11798.

[77] LIU Z, HUANG T T, LI B L, et al. EPNet++: Cascade bi-directional fusion for multi-modal 3D object detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(7): 8324-8341.

[78] JOHN V, MITA S. RVNet: Deep sensor fusion of monocular camera and radar for image-based obstacle detection in challenging environments[M]//Image and Video Technology. Cham: Springer International Publishing, 2019: 351-364.

[79] JOHN V, NITHILAN M K, MITA S, et al. SO-Net: Joint semantic segmentation and obstacle detection using deep fusion of monocular camera and radar[M]//Image and Video Technology. Cham: Springer International Publishing, 2020: 138-148.

[80] BANSAL K, RUNGTA K, BHARADIA D. RadSegNet: A reliable approach to radar camera fusion[EB/OL]. (2022-08-08)[2025-02-26]. https://arxiv.org/pdf/2208.03849.

[81] KIM Y, SHIN J, KIM S, et al. CRN: Camera radar net for accurate, robust, efficient 3D perception[C]//2023 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2023: 17569-17580.

[82] LONG Y F, KUMAR A, MORRIS D, et al. RADIANT: Radar-image association network for 3D object detection[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2023, 37(2): 1808-1816.

[83] LLIN Z W, LIU Z, XIA Z Y, et al. RCBEVDet: Radar-camera fusion in bird's eye view for 3D object detection[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2024: 14928-14937.

[84] NABATI R, QI H R. CenterFusion: Center-based radar and camera fusion for 3D object detection[C]//2021 IEEE Winter Conference on Applications of Computer Vision (WACV). Piscataway: IEEE, 2021: 1526-1535.

[85] KIM Y, KIM S, CHOI J W, et al. CRAFT: Camera-radar 3D object detection with spatio-contextual fusion transformer[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2023, 37(1): 1160-1168.

[86] HWANG J J, KRETZSCHMAR H, MANELA J, et al. CramNet: Camera-radar fusion with ray-constrained cross-attention for robust 3D object detection[M]//Computer Vision-ECCV 2022. Cham: Springer Nature Switzerland, 2022: 388-405.

[87] ZHOU T H, CHEN J J, SHI Y N, et al. Bridging the view disparity between radar and camera features for multi-modal fusion 3D object detection[J]. IEEE Transactions on Intelligent Vehicles, 2023, 8(2): 1523-1535.

[88] YANG B, GUO R S, LIANG M, et al. RadarNet: Exploiting radar for robust perception of dynamic objects[M]// Computer Vision-ECCV 2020. Cham: Springer International Publishing, 2020: 496-512.

[89] WANG Y J, DENG J J, LI Y, et al. Bi-LRFusion: Bi-directional LiDAR-radar fusion for 3D dynamic object detection[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2023: 13394-13403.

[90] BANG G, CHOI K, KIM J, et al. RadarDistill: Boosting radar-based object detection performance via knowledge distillation from LiDAR features[C]//2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). Piscataway: IEEE, 2024: 15491-15500.

[91] WANG L, ZHANG X Y, LI J, et al. Multi-modal and multi-scale fusion 3D object detection of 4D radar and LiDAR for autonomous driving[J]. IEEE Transactions on Vehicular Technology, 2023, 72(5): 5628-5641.

[92] ZHENG L Q, LI S, TAN B, et al. RCFusion: Fusing 4-D radar and camera with bird's-eye view features for 3-D object detection[J]. IEEE Transactions on Instrumentation and Measurement, 2023, 72: 8503814.

[93] XIONG W Y, LIU J N, HUANG T, et al. LXL: LiDAR excluded lean 3D object detection with 4D imaging radar and camera fusion[C]//2024 IEEE Intelligent Vehicles Symposium (IV). Piscataway: IEEE, 2024: 79-92.

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

[95] DONG X, ZHUANG B N, MAO Y X, et al. Radar camera fusion via representation learning in autonomous driving[C]// 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Piscataway: IEEE, 2021: 1672-1681.

[96] 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. Piscataway: IEEE, 2012: 3354-3361.

[97] 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). IEEE, 2020: 11621-11631.

[98] 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). IEEE, 2020: 2446-2454.

[99] PALFFY A, POOL E, BARATAM S, et al. Multi-class road user detection with 3+1D radar in the view-of-delft dataset[J]. IEEE Robotics and Automation Letters, 2022, 7(2): 4961-4968.

[100] ZHENG L Q, MA Z X, ZHU X C, et al. TJ4DRadSet: A 4D radar dataset for autonomous driving[C]//2022 IEEE

25th International Conference on Intelligent Transportation Systems (ITSC). Piscataway: IEEE, 2022: 493-498.
[101] THOMAS H, QI C R, DESCHAUD J E, et al. KPConv: Flexible and deformable convolution for point clouds[C]// 2019 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE, 2019: 6411-6420.
[102] YU X M, TANG L L, RAO Y M, et al. Point-BERT: Pre-training 3D point cloud transformers with masked point modeling[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2022: 19291-19300.
[103] ZHAO H S, JIANG L, JIA J Y, et al. Point transformer[C]// 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2021: 16239-16248.
[104] REN S Q, HE K M, GIRSHICK R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137-1149.
[105] BENGIO Y, GOODFELLOW I, COURVILLE A. Deep Learning[M]. Cambridge: MIT press, 2017: 239.
[106] ZHOU D F, FANG J, SONG X B, et al. IoU loss for 2D/3D object detection[C]//2019 International Conference on 3D Vision (3DV). Piscataway: IEEE, 2019: 85-94.
[107] SZEGEDY C, VANHOUCKE V, IOFFE S, et al. Rethinking the inception architecture for computer vision[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2016: 2818-2826.
[108] KINGMA DIEDERIK P., JIMMY BA. Adam: A method for stochastic optimization[C]//International Conference for Learning Represe-ntations. California: ML Research Press, 2014: 1-15.
[109] TIELEMAN T, HINTON G. Divide the gradient by a running average of its recent magnitude. Neural networks for machine learning[EB/OL]. (2012-09-14)[2025-02-26]. https://www.coursera.org/learn/neural-networks-deep-learning.
[110] LOSHCHILOV I. Decoupled weight decay regularization[C]//In International Conference on Learning Representations. UK: ML Research Press, 2019: 1-15.

作者简介

陈 建 女,1981年生. 博士,福州大学物理与信息工程学院副教授、硕士生导师. 主要研究方向为视频编码、压缩感知、点云压缩和目标跟踪.
E-mail: chenjian-fzu@163.com

黄立勤 男,1973年生. 博士,福州大学物理与信息工程学院教授、博士生导师. 主要研究方向为高性能计算、人工智能与机器学习、医学图像处理等. 中国电子学会会员编号:E190055471M.
E-mail: hlq@fzu.edu.cn

苏思教 男,2000年生. 福州大学物理与信息工程学院硕士研究生. 主要研究方向为3D目标检测.
E-mail: 867225883@qq.com

赵铁松 男,1984年生. 博士,福州大学物理与信息工程学院教授、博士生导师. 主要研究方向为多媒体通信系统、人工智能和视频编码等. 中国电子学会会员编号:E190014840S.
E-mail: t.zhao@fzu.edu.cn