自动驾驶3D目标检测研究综述

任柯燕†, 谷美颖, 袁正谦, 袁 帅

(北京工业大学 信息学部，北京 100124)

摘 要: 精确实时地进行目标检测是自动驾驶车辆能够准确感知周围复杂环境的重要功能之一,如何对周围物体的尺寸、距离、位置、姿态等3D信息进行精准判断是自动驾驶3D目标检测的经典难题. 服务于自动驾驶的3D目标检测已成为近年来炙手可热的研究领域,鉴于此,对该领域主要研究进展进行综述. 首先,介绍自动驾驶感知周围环境各相关传感器的特点;其次,介绍3D目标检测算法并按照传感器获取数据类型将其分为:基于单目/立体图像的算法、基于点云的算法以及图像与点云融合的算法;然后,对每类3D目标检测的经典算法以及改进算法进行详细综述、分析、比较,梳理了当前主流自动驾驶数据集及其3D目标检测算法的评估标准,并对现有文献广泛采用的KITTI和NuScenes数据集实验结果进行对比及分析,归纳了现有算法存在的难点和问题;最后,提出自动驾驶3D目标检测在数据处理、特征提取策略、多传感器融合和数据集分布问题方面可能遇到的机遇及挑战,并对全文进行总结及展望.

关键词: 机器视觉；深度学习；目标检测；3D目标检测；自动驾驶

中图分类号: U463.6；TP391.41 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2022.0618

引用格式: 任柯燕,谷美颖,袁正谦,等. 自动驾驶3D目标检测研究综述[J]. 控制与决策, 2023, 38(4): 865-889.

3D object detection algorithms in autonomous driving: A review

REN Ke-yan†, GU Mei-ying, YUAN Zheng-qian, YUAN Shuai

(Faculty of Information Technology，Beijing University of Technology，Beijing 100124，China)

Abstract: Accurate and real-time object detection is one of the important functions for autonomous vehicles to accurately perceive the surrounding complex environment. Nevertheless, how to get the accurate size, distance, position, posture and other 3D information of surrounding objects is a classic problem. 3D object detection for autonomous driving has become a popular research field in recent years. Main research progress in this field is reviewed. Firstly, the characteristics of relevant sensors in the surrounding environment of autonomous driving is introduced. Then, the development of object detection from 2D to 3D is introduced and the loss functions is applied for optimization. According to the type of data acquired by the sensor, 3D object detection algorithms is categorized into three types, which are algorithms based on monocular/stereo images, point clouds, image and point cloud fusion. Futhermore, the classic and improved algorithms for each type of 3D object detection are reviewed, analyzed, and compared in detail. Simultaneously, the mainstream autonomous driving datasets and the evaluation criteria of their 3D object detection algorithms are summarized. Extensive experiment results of KITTI and NuScenes datasets are also compared and analyzed, which is widely used inpresent literature, summarizing the difficulties and problems of the existing algorithms. Besides, the opportunities and challenges of 3D object detection in data processing, feature extraction strategy, multi-sensor fusion and data distribution problems are proposed in hope of inspiring more future work.

Keywords: computer vision；deep learning；object detection；3D object detection；autonomous driving

0 引 言

自动驾驶汽车依靠人工智能、视觉计算、雷达、监控装置和全球定位系统协同合作,让电脑可以在没有人类主动的操作下自动安全地操控机动车辆. 自动驾驶车能够更好地适应人群,缓解交通堵塞,提高公路安全性,解放劳动力等[1]. 在18世纪初期,自动驾驶主要针对军事或概念性畅想. 美国国防高级研究计划署(DARPA)[2]自1984年起与陆军展开合作,启

收稿日期: 2022-04-14；**录用日期:** 2022-09-03.

基金项目: 国家重点基础研究发展计划项目(2019YFC1511000)；国家自然科学基金项目(61803004).

责任编委: 侯忠生.

†通讯作者. E-mail: keyanren@bjut.edu.cn.

*本文附带电子附录文件,可登录本刊官网该文“资源附件”区自行下载阅览.

动了自主地面车辆计划,并于2004、2005、2007年分别举办了3届自动驾驶挑战赛[3-4]. 自2009年,谷歌开始针对民用市场做研发,开展了自动驾驶汽车项目. 2015年后,特斯拉、蔚来、百度、华为等国内外多家企业及高校如雨后春笋,纷纷开启了自动驾驶研发及测试.

美国汽车工程师学会(society of automotive engineers, SAE)[5]于2014年发布了一项名为J3016“驾驶自动化水平”的标准,旨在促进技术和政策领域之间沟通与合作. 该标准根据驾驶员干预和注意力需求将自动驾驶定义为6个级别,如表1所示. 可见,自动驾驶车的自主驾驶功能分界发生在第3级. 在1~3级中,高级驾驶辅助系统(advanced driver assistance system, ADAS),根据传感器[6]感知到的周围环境信息及本车运行状态信息,进行一定的决策规划后,提醒驾驶员采取某些动作或代替驾驶员进行一部分操控,从而达到减轻驾驶员操控负担,提高车辆驾乘安全性和舒适性的目的. 在4~5级中,已经达到了超高度的自动驾驶级别,是完全的自动驾驶系统,无需驾驶员的操作.

表 1 自动驾驶6级定义

汽车智能化分级	SAE定义
第0级人工驾驶	驾驶员负责所有的驾驶操作
第1级辅助驾驶	车辆控制转向和加减速中的一项操作,驾驶员负责操作其余驾驶动作
第2级半自动驾驶	车辆控制转向和加减速中的多项操作,驾驶员负责操作其余驾驶动作
第3级高度自动驾驶	在限定道路和环境条件下,车辆完成绝大部分驾驶操作,驾驶员在必要时对车辆进行操作
第4级超高度自动驾驶	在限定道路和环境条件下,车辆能完成所有驾驶操作,驾驶员无需操作
第5级全自动驾驶	任何条件下,车辆能完成所有驾驶操作,驾驶员无需操作

目前,4级超高度自动驾驶汽车还未普及,主要原因在于车辆对周围环境感知的准确性和精确性要求更高,甚至超过人类的认知水平. 作为环境感知的核心,3D目标检测算法[7-10]在2D目标检测[11]的基础上,在定位和尺寸回归上引入了带有深度信息的第3个维度. 自动驾驶3D目标检测的主要任务是通过对不同传感器输入数据的处理,输出目标物体的类别和精确位置,从而帮助车辆躲避障碍、规划路线. 因此,能够准确地对周围环境进行感知,实现精确可靠的3D目标检测对于自动驾驶至关重要. 然而,由于获取数据的传感器容易受到外界环境因素的干扰、场景视野变化大、目标过小或互相遮挡等原因,自动驾驶场景下的3D目标检测任务更具挑战性. 如何更高效快速地提高3D目标检测算法的识别定位精度,如何降低外界环境对3D目标检测算法的影响是一项艰巨的技术挑战.

为了进一步提升3D目标检测算法的鲁棒性和准确度,使自动驾驶车辆能够达到4级甚至5级,本文系统地总结了近年来自动驾驶3D目标检测方法,并将相关工作分为4个部分. 第1部分介绍应用于自动驾驶的3D目标检测的基础知识,包括数据采集传感器、常用的数据集、基础检测模型、3D输出框的表示和损失函数的设计;第2部分详细归纳了针对不同传感器数据输入的典型3D目标检测算法,将其分类为基于单目/立体图像的方法、基于点云的方法以及基于图像和点云融合的方法,分析框架如图1所示;第3部分介绍各类算法在自动驾驶的3D目标检测的主流数据集上的相关评估指标和大量实验对比分析;第4部分提出一些在未来可能遇到的挑战和机遇.

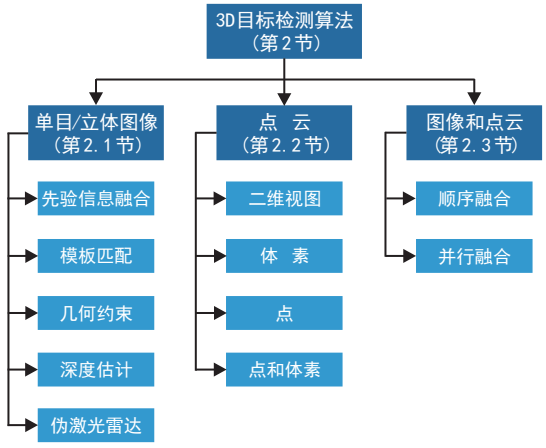


图 1 自动驾驶3D目标检测算法分类

1 基础知识

为了正确理解深度学习框架如何解决3D目标检测问题,本节主要介绍应用于自动驾驶的3D目标检测算法相关的基础知识,包括自动驾驶车辆所使用的数据采集传感器,自动驾驶3D目标检测所使用的数据集,3D目标检测的基础模型,3D目标检测算法输出的3D边框定义,3D目标检测模型损失函数的设计.

1.1 数据采集传感器的分类

自动驾驶汽车看到和感知道路上的一切、收集安全驾驶所需的信息都是通过车载传感器来完成的. 这些车载传感器包括摄像头、激光雷达、超声波雷达和毫米波雷达等,其在自动驾驶汽车上的分布如

图2所示. 本节主要对不同车载传感器的传感方式和特点进行详细总结与归纳,并在表2中列出了他们的优缺点.

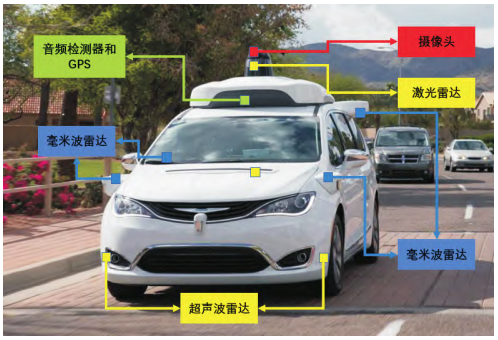


图 2　自动驾驶汽车相关传感器布置

表 2　车载传感器对比

类型	优点	缺点
单目相机	成本低廉,提供颜色和纹理信息	无深度信息,受环境光照等因素影响
立体相机	提供深度信息以及颜色和纹理信息	计算消耗高,实时性较差,受环境光照等因素影响
机械式旋转激光雷达	360° 视野,运行稳定,精度高	高稀疏性和不规则性,成本较高且难部署,无颜色和纹理信息
固态激光雷达	体积小,成本低	视角有限,不同视角点实时合并时会增大误差
混合固态激光雷达	体积小,成本较低,360° 视野	不稳定,对光源功率要求过高
超声波雷达	成本低,功能简单	只能做近距离探测,易受天气影响
毫米波雷达	探测距离远,不受天气影响	目标识别的准确性较低

1.1.1　摄像头

车载摄像头是自动驾驶汽车最常见的传感器,已被广泛部署于车上. 摄像头相当于车的“眼睛”,是唯一能分辨所见物体颜色、形状和位置的传感器. 利用获取的物体信息数据,自动驾驶汽车可以观察和分析道路上的物体,对环境进行理解与判断. 摄像头有着显著的视觉性能,所采集的丰富图像信息是其他传感器无法比拟的,比如红绿灯监测和交通标志识别只能通过摄像头实现. 但作为自动驾驶中一个独立的感知系统还远远不够,主要原因在于摄像头容易受到光线和天气等条件的影响,例如在夜间亮度较低或下雨的天气条件下会导致精度降低. 因此,自动驾驶汽车还需配备辅助传感器,以防摄像头系统出现故障或断开. 按照镜头和布置方式的不同可将摄像头划分为以下几种.

1) 单目相机.

单目相机通过传感器采样和量化,将3D世界中的物体变换到2D空间,用单个或多个通道的二维图像来描绘物体的形状、颜色、纹理和轮廓等信息,这些信息可用于检测物体类别、交通标志和车道线等. 但是,单目相机不直接提供深度信息,需要先将感知到的目标障碍物与模型数据库样本建立起对应关系,再通过样本库所识别出的对应物体与车辆进行距离估算.

2) 立体相机.

立体相机在视觉感知识别方面拥有单目相机的全部识别功能. 在定位测距方面,由于立体相机含有两个摄像头,可以使用匹配算法对左右图像的对应位置进行深度恢复,得到稠密的深度图,帮助车辆对物体距离进行更好地把控. 缺点是需要相机标定,计算量大,实时性较差.

3) 其　他.

除了单目相机和立体相机外,三目摄像头定位测距方面感知范围更大,全景摄像头可以获得更大视野范围,深度摄像头可以利用单个相机的运动轨迹形成虚拟的双目相机成像序列以获取环境的深度和颜色信息. 但这些摄像头成本和算力投入更高,目前在自动驾驶中还没有获得大规模的应用.

1.1.2　雷　达

雷达(radio detection and ranging, radar)的含义是无线电探测和测距,即用无线电的方法发现目标并测定它们的空间位置. 因此,雷达也被称为“无线电定位”,是利用电磁波探测目标的电子设备. 雷达发射电磁波对目标进行照射并接收其回波,从而获得目标到电磁波发射点的距离、距离变化率(径向速度)、方位、高度等信息. 按雷达频段分类,可分为超视距雷达、激光雷达、毫米波雷达以及超声波雷达. 其中:超视距雷达多用于军事,在自动驾驶领域中通常使用后3种. 接下来,主要从定义、分类、原理、优缺点对后3种雷达进行介绍.

1) 激光雷达.

激光雷达(light detection and ranging, lidar)是自动驾驶车辆中最重要的传感器之一,绝大多数自动驾驶方案中都选择配备激光雷达. 激光雷达由发射系统、接收系统和信息处理3部分组成,其工作原理是利用可见和近红外光波(多为950纳米波段附近的红外光)向目标发射探测信号,然后测量反射或散射信号的到达时间、强弱程度等参数,进而确定目标的距离、方位、运动状态以及表面光学特性. 激光雷达可

按照有无机械旋转部件分为以下3类.

① 机械式激光雷达. 机械式激光雷达是指其发射系统和接收系统存在宏观意义上的转动,通过不断旋转发射头,将速度更快、发射更准的激光从“线”变成“面”,并在竖直方向上排布多束激光,形成多个面,达到360°动态扫描并动态接受信息的目的. 机械式激光雷达技术是最早最成熟的,因为带有机械旋转结构,所以具有扫描速度快、视场范围大、可承受高的激光功率等优点. 但其信号稳定性易受温度变化的影响,并且结构笨重、体积和重量较大、部署成本高,比如百度和谷歌无人驾驶汽车车身上的Velodyne 64线机械式激光雷达,售价最高时达到70万元人民币. 由于通过复杂的机械结构实现高频准确的转动,平均失效时间仅1 000～3 000 h,无法满足乘用车车规级要求的13 000 h,现在已沦落为测试车专用了. 机械式LiDAR的代表企业有Velodyne、Valeo、Ouster、Waymo、速腾聚创、禾赛科技、镭神智能、北科天绘等.

② 固态激光雷达. 相比于机械式激光雷达,固态激光雷达内部没有宏观与微观运动部件,在结构和尺寸上可以大大减小,成本也可以大幅降低. 由于没有过多的机械结构,耐久性和可靠性较高,符合自动驾驶对雷达的要求. 从使用的技术上看,固态激光雷达分为OPA固态激光雷达和Flash固态激光雷达. OPA激光雷达是使用光学相控阵技术,主要采用多个光源组成阵列,通过控制各光源发光时间差,合成具有特定方向的主光束,对主光束加以控制便可实现对不同方向的扫描,具有扫描速度快、精度高、可控性高、体积小(Quanergy激光雷达只有90×60×60 mm大小)的优点. 但是,每台的调节角度只有60°,要想实现360°全方位扫描,需要在不同方向部署多台,还容易形成旁瓣影响光束作用距离和角分辨率,同时生产难度较高,市场上有部分企业正在研发探索中,如Analog Photonics、力策科技、万集科技、洛微科技等,预计该技术的真正落地还需5年以上. Flash固态激光雷达是在一段时间直接发射出一大片覆盖探测区域的激光,加上高度灵敏的接收器,快速记录整个场景,没有转动与镜片磨损,相对更为稳定. 但是缺陷也很明显,比如Flash固态激光雷达的探测距离仅能达到100 m左右,对处理器要求较高,难以克服的发热问题,还有成本较高的问题,代表厂商有LeddarTech、Sense Photonics、大陆、IBEO、北醒光子、北科天绘、Xenomatix、Ouster等.

③ 混合固态激光雷达. 传统的机械式激光雷达必须使激光发射器转动,而混合固态激光雷达利用半导体微动器件(如MEMS扫描镜)驱动转镜或棱镜达到机械式激光雷达360°扫描的功能. MEMS扫描镜是一种硅基半导体元器件,属于固态电子元件,但其内部集成了可动的微型镜面,因此,混合固态激光雷达具有固态和运动两种形式. 混合固态激光雷达本身不用大幅度地进行旋转,因此,可有效降低整个系统在行车环境出现问题的几率,具有较高的可靠性. 此外,由于主要部件采用芯片工艺,其生产能力也大幅度提高,有利于降低激光雷达的成本. 探测点可以任意分布,例如在高速公路主要扫描前方远处,但是不能完全忽略侧面,可以对侧面进行稀疏扫描,在十字路口则加强侧面扫描. 这是只能匀速旋转的机械式激光雷达所无法执行的精细操作. 探测距离也符合当前自动驾驶要求,使其得到广泛应用. 代表厂商有Innoviz、Innovusion、Pioneer、Blickfeld、华为、速腾聚创、万集科技、禾赛科技、一径科技、镭神智能等.

2) 毫米波雷达.

毫米波雷达[12]是在30～300 GHz毫米波波段探测的雷达,其频率高于无线电,低于可见光和红外线,通过发射和接受电磁波来确定物体的距离和速度. 常见的车载毫米波雷达可按照感测距离划分为以下3类.

① 短程毫米波雷达(24 GHz频段),感测距离在0.15～30 m左右,安装在车辆的后保险杠内,用于汽车盲点监测和变道辅助. 目前价格大约为45～60美元/只.

② 中程毫米波雷达(76～77 GHz频段),感测距离在1～100 m左右,装配在车辆前保险杠,用于探测与前车之间的距离及车速,以确保紧急制动和自动跟车等主动安全功能的实施. 目前价格大约为45美元/只.

③ 长程毫米波雷达(77 GHz频段),感测距离可达到250 m左右,相较于短程毫米波雷达,分辨准确率提高2～4倍,测速和测距精确度也提高3～5倍,主要用于主动巡航系统和汽车前向碰撞报警系统,让车主有足够的时间来刹车或闪避. 目前价格大约为80～90美元/只.

毫米波雷达具有体积小、重量轻的优点,安装之后对车辆外观影响不大,虽然不具备产生高分辨率图像的能力,但是,能够同时检测多个物体的距离、角度和相对速度,特别是高速移动的物体. 毫米波雷达测量距离最大可达250 m,穿透雾、烟、灰尘的能力强,能够适应各种不同的天气,这是昂贵的激光雷达做不到的. 但是,毫米波雷达也有明显的缺点,其测量角

度受限,特别是垂直角度检测效果较差. 毫米波雷达采样的点比较稀疏,分辨率较低,难以识别小尺寸物体. 近年来,毫米波雷达逐渐从3D演进到4D成像. 传统雷达输出3个维度的信息,分别是方位角、速度和距离,4D高精成像技术增加了雷达对目标俯仰高度数据的探测和解析,可实现俯仰角、水平角、速度和距离的信息感知. 加上对目标的高度分析,将第4个维度整合到传统毫米波雷达中,更好地了解和绘制环境地图,使测到的交通数据更为精准,这将弥补传统毫米波雷达的许多问题,能够全方位提升毫米波雷达性能,有望使毫米波雷达成为ADAS系统中的核心传感器之一,是毫米波雷达未来发展的重要方向.

3) 超声波雷达.

超声波雷达通过发射超声波来测算与障碍物的距离,但探测距离非常短,一般在10 m以内. 超声波雷达一般分为两种,分别是UPA和APA,它们的探测范围和探测区域都不太相同. UPA超声波雷达的探测距离一般在15～250 cm之间,感测距离较短,频率为58 kHz比较高,精度也高,主要安装在汽车前后保险杠上,用于测量汽车前后方的障碍物; APA超声波雷达的探测距离一般在30～500 cm之间,感测距离较长,但频率为40 kHz比较低,精度一般,主要安装在汽车侧面,用于测量侧方障碍物距离. 超声波雷达的特点是频率高、波长短和绕射现象小,对液体、固态穿透性较强,但传输速度容易受到温度影响. 超声波雷达价格低,单个超声波雷达售价大约为数十元. 目前,主要应用在自动泊车、倒车雷达等配置中,用于检测障碍物,避免碰撞和擦蹭.

1.2　数据集

自动驾驶是对安全性要求极高的应用领域,所以在进行感知算法研究时,需要尽可能地覆盖更复杂多样的路况. 大规模数据集对于数据驱动[13]的深度学习算法的成功至关重要,还能够为不同算法之间的比较提供平台以及统一的评估标准. 因此,本节重点介绍现有公开的与3D目标检测相关的自动驾驶数据集,包括数据集大小、多样性和附加数据等,这些数据集的出现极大地促进了3D检测的发展. 表3对比了这些数据集的特点.

表3　3D目标检测公开数据集对比

数据集	年份	场景	图像	点云/k	视角范围/(°)	3D框	类别	拍摄环境
KITTI	2012	22	15 k	15	90	200 k	8	白天
NuScenes	2019	1 k	1.4 M	400	360	1.4 M	23	白天＋夜晚＋雨天
Waymo Open	2019	1 k	1 M	200	360	12 M	4	白天＋夜晚＋雨天
ApolloScape	2018	–	144 k	–	360	70 k	25	白天＋夜晚

1) KITTI数据集.

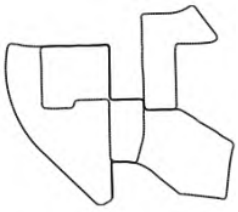
KITTI[14]数据集由德国卡尔斯鲁厄理工学院和丰田美国技术研究院于2012年联合创办,是针对自动驾驶环境感知算法研究中最常用的公开数据集之一,如图3所示. 该数据集具有丰富的激光点云数据和图像数据,可用于视觉测距、2D/3D目标检测、目标跟踪、语义分割和光流等计算机视觉算法的研究. 针对3D目标检测任务,KITTI数据集由7 481张图片组成训练集和验证集,7 518张图片组成测试集,共计包含超过20万的3D目标标注信息. 该数据集将3D目标分为汽车、行人、骑行的人等8种类别,标注信息包括类别、2D检测框坐标、3D中心点坐标、3D尺寸、遮挡、截断以及航向角等信息. 数据集从市区、乡村和高速公路等场景采集真实图像数据,每张图像中最多包含30个行人和15辆车. 根据每个场景下不同程度的遮挡与截断将其分类为简单、中等和困难3类.

该数据集尽管已被广泛采用,但仍存在一些局限性,特别是在传感器配置和光照条件方面: 传感器仅采集了面向驾驶方向的90° 范围内数据,并且所有测量值都是由同一组传感器在白天且大多在阳光充足的条件下获得的. 另外,数据集中目标出现的频率非常不平衡: 汽车占比75%,骑行的人占比4%和行人占比15%. 多样性的缺乏也降低了其在实际应用中的可靠性.

2) NuScenes数据集.

NuScenes[15]数据集是由nuTonomy与Scale联合发布的,主要针对3D目标检测任务,共标注了1 000个道路场景. 其中: 850个场景作为训练验证集,另外150个场景作为测试集. 每个场景包含20 s的视频,有40个关键帧,并对每个关键帧道路场景下汽车、行人、骑行的人、卡车、公交车和交通路标等23类目标进行手工标注,并且标注信息可实现与KITTI数据集标注格式之间的转换. 相比KITTI所采集的90° 范围数据,NuScenes通过6个相机同时采集图像数据,相机覆盖了车体的360° 方向,使图像数据所包含的视野更具多样性. 如图4所示,图4(a)展示了6种不同的相机视图,图4(b)展示了激光雷达和雷达数据,以及人类注释语义地图.

(a) 带有传感器的记录平台

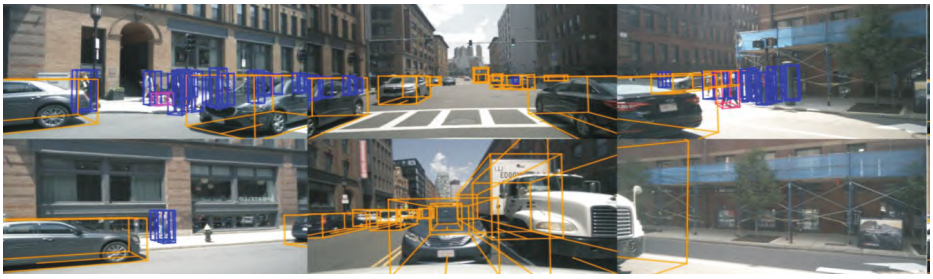


(b) 视觉里程基准轨迹

(c) 视差和光流图

(d) 相机前视图及其 3D 目标标签

图 3 KITTI 数据集示例



(a) 6 种不同相机视图

(b) 雷达和激光雷达数据, 以及人类注释语义地图

图 4 NuScenes 数据集示例

相比于 KITTI 数据集, NuScenes 的数据规模更大, 还包含了白天和夜晚以及不同天气、光照等场景状况的应用. 图 5 是 NuScenes 数据集的前置摄像头获取得到的 4 种场景下的图像示例: 晴天 (图 5(a))、夜间 (图 5(b))、雨天 (图 5(c))、施工区域 (图 5(d)).

(a) 晴天 (b) 夜间 (c) 雨天 (d) 施工区域

图 5 NuScenes 数据集不同场景下的图像示例

3) Waymo Open数据集.

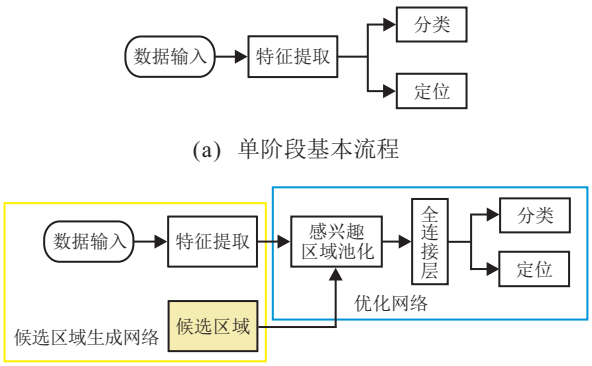
Waymo[16]数据集是由Waymo公司发布的自动驾驶数据集. 数据集使用5个激光雷达传感器和5个高分辨率针孔摄像机进行数据收集,选取了一天中不同时间段以及不同天气的郊区和城市地区的场景,包含798个用于训练的场景和202个用于验证的场景,以及150个用于测试的场景,每个场景的时间跨度为20 s. 数据集对车辆、行人、标志和自行车4类目标一共标注了约1 200万个3D标签和1 000万个2D标签.

4) ApolloScape数据集.

ApolloScape[17]是由百度公司发布的大规模自动驾驶数据集,收集了中国4个不同区域白天和夜晚条件下的驾驶场景,包含143 906帧带有像素注释、2D和3D标签的图像. 数据集中共定义了25种类别,例如汽车、自行车、行人、建筑、摩托、卡车和交通标志等,以及28类车道线的精细标记. 根据每幅图像中车辆和行人的数量级将数据集划分为容易、适中和困难3个子集,并计划在未来获取和注释约100万帧图像和对应的3D点云.

1.3　3D目标检测基础模型

作为重要的自动驾驶感知算法,3D目标检测算法以传感器数据为输入对3D空间中的目标进行分类和定位. 每个目标由一个带有概率分数的类标签(K个预定义类中的一个)和一个3D框表示(表示方法见1. 4节). 同2D目标检测算法[18-23]一样,3D目标检测可分为单阶段方法和两阶段方法,通用的3D目标检测模型流程如图6所示.



(a)　单阶段基本流程

(b)　两阶段基本流程

图6　单阶段和两阶段方法的基础流程

1) 单阶段方法.

单阶段方法是端到端训练,直接回归目标的类别概率和位置坐标. 因为是端到端训练,所以计算速度快,缺点是对小目标的检测效果不太好,平均检测精度不如两阶段方法. 经典的单阶段算法是SSD (single shot detector)系列工作,如3D-SSD[24]、SA-SSD[25]、SE-SSD[26]等.

2) 两阶段方法.

两阶段方法是先定位、后识别,先从场景中提取可能包含目标的区域候选框,再回归目标的类别概率和目标精准位置坐标. 优点是平均检测精度高,缺点是普遍训练时间较长. 经典的两阶段算法是区域卷积神经网络(region-based convolutional neural network, R-CNN)系列工作,如3D-RCNN[27]、Voxel R-cnn[28]、PointRcnn[29]等. 具体的算法介绍和分类可见第2节.

1.4　3D边框输出表示

作为3D目标检测算法的输出,3D边框表示目标的位置、大小和方向. 无论物体是否被遮挡、截断或具有不规则的形状,都用一个紧密边界的立方体包围住被检测到的目标.

3D边框编码方式主要有3种,分别是8角点法、4角2高法、7坐标参数法,如图7所示. 目标方向对于目标跟踪和轨迹预测等任务起着关键作用,由于自动驾驶汽车都是在地面上行驶,本文主要对目标沿z轴偏航角θ预测的方法进行综述.

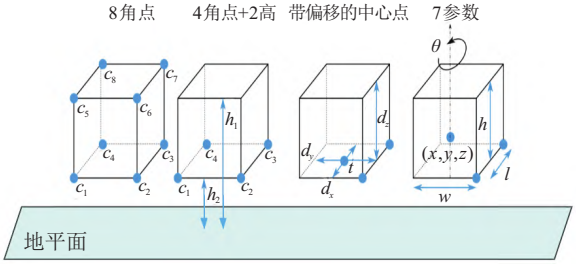


图7　3种不同的3D框表示法

1) 8角点法.

Chen等[30]提出8角点法将3D边框通过连接8个角点$(c_1, c_2, \ldots, c_8)$来形成. 每一个角点由三维坐标(x, y, z)表示,总计24维向量. 目标的方向通常被认为是3D边框较长边的方向.

2) 4角2高法.

为了保持地面目标的物理约束,3D框的上角需要保持与下角对齐,Ku等[31]提出了一种4角2高编码的方法. 4个角点(c_1, c_2, c_3, c_4)表示3D边框底面的4个顶点,每个角点用2D坐标(x, y)表示. 两个高度值(h_1, h_2)表示从地平面到底部和顶部角的偏移量. 根据4个角点计算出4个可能的方向,并选择最近的一个作为方向向量.

3) 7坐标参数法.

此类方法[29,32-33]的3D边框由7个坐标参数来表示. 它包括边框的中心位置(x, y, z),边框在三维空间

中的尺寸 (l,w,h) 以及表示角度的偏航角 θ,即垂直于重力方向的平面内旋转角度.

1.5 损失函数的设计

在深度学习网络训练中,损失函数用来表现预测与实际数据的差距程度,可以衡量模型预测的好坏. 在3D目标检测模型中,主要使用两类损失函数:分类损失和位置损失. 这两类损失函数通常用于检测模型的最后部分,并根据模型输出的类别和位置以及实际标注框的类别和位置分别计算分类损失和位置损失. 两类损失函数可通用表达为

$$\text{Loss} = L_{\text{cls}}(p_i, p_i^*) + L_{\text{loc}}(t_i, t_i^*). \tag{1}$$

其中: $L_{\text{cls}}(p_i, p_i^*)$ 表示分类损失, p_i 是预测网络输出的 N_{cls} 维的向量, N_{cls} 表示所需要分类的类型, p_i^* 表示真实数据集中的类型,同样为 N_{cls} 维; $L_{\text{loc}}(t_i, t_i^*)$ 是位置回归损失, t_i 是一个表示预测边界框的坐标向量, t_i^* 表示真实框(ground truth)的坐标向量.

1) 分类损失.

通常使用交叉熵损失函数和Focal损失函数. 交叉熵损失函数的计算公式为

$$sL_{\text{cls}}(p_i, p_i^*) = -\sum_{j=1}^{c} y_{i,j}\log(p_{i,j}). \tag{2}$$

其中: $y_{i,j}$ 当第 i 个样本属于类别 j 时等于1,其他情况等于0; $p_{i,j}$ 表示第 i 个样本属于类别的概率,通常采用SoftMax函数计算样本属于每一个类别的概率.

Focal损失函数是对交叉熵损失函数的改进,主要用于样本分类不平衡问题,计算公式为

$$p_t = \begin{cases} p, & y = 1; \\ 1-p, & y = 0. \end{cases} \tag{3}$$

$$FL(p_t) = -(1-p_t)^{\gamma}\log(p_t). \tag{4}$$

其中: p_t 表示被预测为对应的正确类别的置信度,值越大说明分类越准确; γ 为可调节因子.

2) 回归损失.

通常使用 L_1 损失、 L_2 损失、平滑 L_1 损失或IoU损失等. 其中, L_1 损失即平均绝对误差(mean absolute error, MAE),指预测值与真实值之差的平均值. 记 t 为预测值, t^* 为真实值,计算公式为

$$\text{MAE} = |t - t^*|. \tag{5}$$

L_2 损失即均方误差损失(mean square error, MSE),指预测值与真实值之差的平方,计算公式为

$$\text{MSE} = (t - t^*)^2. \tag{6}$$

平滑 L_1 损失是基于 L_1 损失修改得到的,计算公式为

$$L_{\text{loc}}(t_i, t_i^*) = \begin{cases} 0.5(t-t^*)^2, & |t-t^*| < 1; \\ |t-t^*| - 0.5, & \text{otherwise}. \end{cases} \tag{7}$$

IoU损失是基于预测框与标注框之间的交并比,记预测框为 P,标注框为 T,则IoU损失计算公式为

$$L = 1 - \frac{P \bigcap T}{P \bigcup T}. \tag{8}$$

2 基于不同传感器的自动驾驶3D目标检测算法

本节按照不同传感器的数据输入将3D目标检测算法分为3大类别:基于单目/立体图像的方法(2.1节)、基于点云的方法(2.2节)、基于图像和点云融合的方法(2.3节). 图8按时间顺序,梳理了近几年经典3D目标检测方法,并分为单阶段和两阶段.

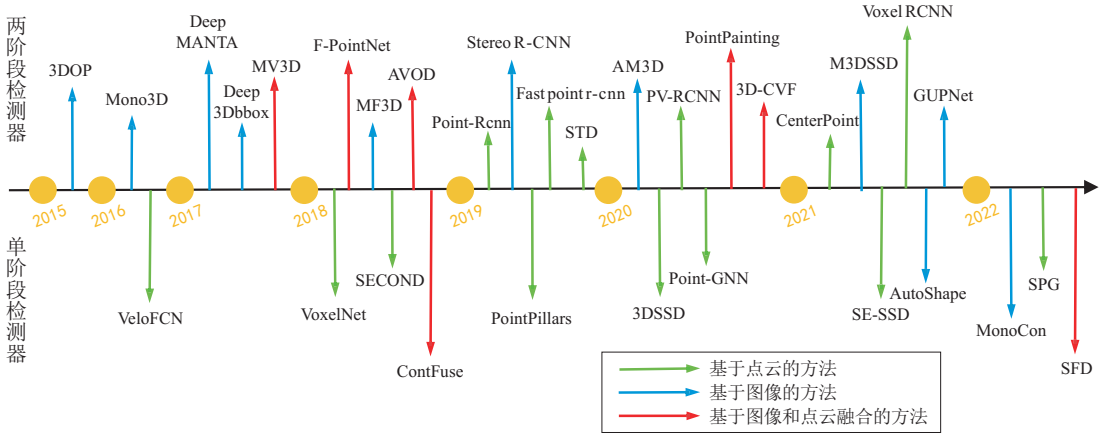


图 8 按时间顺序展示应用于自动驾驶的相关3D目标检测算法

2.1 基于单目/立体图像的3D目标检测

这些方法仅以单目/立体图像作为输入来预测物体的类别和3D位置. 常用的方法有模板匹配、几何约束、深度估计和生成伪激光点云.

1) 模板匹配.

由于车辆、行人等物体具有固定的尺寸,可以通

过穷举采样或对代表性模板进行匹配来提取目标3D信息. 此类方法通常需要建立一个相关数据模板库,并通过网络将图像中的目标与模板库中的最佳模型进行匹配.

早期的算法是Chen等[34]提出的3DOP算法,以立体图像对作为输入,利用Fast RCNN管道来联合回归目标位置,并提出了3D目标多特征先验的能量损失函数,其中先验特征包含物体的语义信息、点云密度、上下文信息等. 考虑到汽车只配备单目摄像头的情况,Chen等[35]提出了Mono3D算法,通过只使用单目摄像头而不是立体摄像头来达到相同的性能. 与3DOP不同的是,Mono3D在不计算深度信息的情况下,通过滑动窗口直接从3D空间中对3D目标候选对象进行采样,并假设目标所在的地平面与图像平面正交,以减少搜索力度. 结合语义信息、上下文信息、位置先验信息以及目标形状先验信息等,对候选区域进行评分并选取最有希望的候选区域,通过Fast RCNN管道实现最终的目标回归和定位. 由于Mono3D并没有取得很好的准确性和速度,He等[36]提出了Mono3D++算法对其进行性能改善. 他们采用EM-Gaussian算法实现遮挡或截断目标的关键点检测与补全,并结合单目深度、地面平面约束和车辆形状的先验信息计算回归3D包围框. 然而,为了建立2D先验信息与3D空间中目标之间的对应关系,这些算法过度依赖精细的工程计算和大量的领域专业知识,严重限制了这些模型在复杂场景中的普遍性.

另一种简单有效的方法是Xiang等[37]提出的3DVP模型,对RGB强度、体素和遮挡罩的3D形状外观进行建模,通过对数据上观察到的模式进行聚类并为给定车辆的2D图像片段的每个特定模式训练分类器以获得3DVP字典. 虽然它允许对遮挡和零件外观进行建模,但3DVP字典的获取方式是训练集中常见的现有可见性模型,因此,可能无法推广到与现有模式不同的任意车辆姿态. Chabot等[38]提出了Deep MANTA算法,定义了一组车辆关键点来表征车辆的外部形状,例如车灯、后视镜等,再利用2D检测来回归2D边界框、分类和关键点,与人工建立的3D标准模板库匹配出最佳的3D CAD模型,从而得到完整精确的3D位置和方向. Kundu等[27]提出了3D-RCNN算法,用一组CAD模型进行PCA建模,并利用一组基向量来表征物体的3D形状和姿态. Sun等[39]提出的实例视差估计网络Disp r-cnn,仅预测感兴趣目标物体的像素视差,并学习特定类别的形状先验,以获得更精确的视差估计. Liu等[40]提出的AutoShape算法,引入关键点以建模物体的形状信息,进而利用形状信息提升单目3D检测性能. 通过模板匹配结合的方法是目前解决遮挡、截断目标检测的有效方案,但是,受限于所选模型所覆盖的形状空间,不容易扩展到没有模型的应用场景,并且模板数据的获取较为困难,不利于多目标检测,因此,仍然存在一定的局限性.

2) 几何约束.

这类方法不需要大量的3D候选框来实现高召回率,而是直接从准确的2D边界框开始,利用几何属性估计3D框.

Mousavian等[41]提出了Deep3Dbbox算法,利用2D检测框和几何投影估计物体3D位姿和尺寸,通过求解目标中心到相机中心的平移矩阵,使预测的3D检测框重投影中心坐标与2D检测框中心坐标的误差最小. Brazil等[42]提出了M3D-RPN算法,利用深度感知卷积层提取全局和局部特征,并将两个分支的特征按照一定权重进行结合;然后采用位姿优化算法进行方位估计,并同时预测2D和3D检测框,验证了单阶段网络的有效性. Simonelli等[43]提出了MonoDIS算法,利用解耦的回归损失代替之前同时回归中心点、尺寸和角度的损失函数,该损失函数将回归部分分成K组,通过单独回归参数组来解决不同参数之间的依赖关系,有效避免了各参数间误差传递的干扰,使得训练更加稳定. Luo等[44]提出了M3DSSD算法,先根据预定义锚点的分类置信度得分得到目标区域,再利用2D/3D中心的预测结果来计算特征偏移量,以减小预测结果与其对应的特征图之间的差距. Li等[45]提出了Stereo R-CNN算法,同时对左右两侧图像进行目标检测并生成目标关联对,通过结合Mask RCNN[46]的关键点和左右目标感兴趣区域特征,利用几何约束,即3D角点与2D包围框的投影关系以及预测的关键点,恢复精确的3D检测框. Qin等[47]提出的TLNet利用三维锚定来显式地构造立体图像感兴趣区域之间的对象级对应关系,学习在三维空间中检测和三角化目标. Li等[48]提出的RTS3D设计了一种四维特征一致嵌入空间作为三维场景的中间表示,没有深度监督,通过探索立体对变形后的多尺度特征一致性来对目标的结构和语义信息进行编码. Guo等[49]提出的Liga-stereo在高级几何感知表示的指导下学习基于立体的三维检测器,附加一个辅助2D检测头来提供直接的2D语义监督.

之前的方法都预先对全部场景给出了各类目标的锚框,即Anchor-based. 这种方法在一定程度上能够解决目标尺度不一和遮挡问题,提高检测精度,但

缺乏效率性且很难枚举所有的方向,或为旋转的目标拟合一个轴对齐的包围框. 2D目标检测的Anchor-free方法[50-53]抛弃了需要生成的复杂锚框,而是通过直接预测目标的角点或中心点等方法来形成检测框. 受此启发,Li等[54]提出了RTM3D算法,直接预测3D框的8个顶点和1个中心点,然后通过使用透视投影的几何约束估计3D边框. Liu等[55]提出了SMOKE算法,同样也舍弃了对2D边界框的回归,通过将单个关键点估计与回归的三维变量相结合来预测每个检测目标的3D框. Ma等[56]发现子任务位置误差是影响检测效果的一个重要因素,因此,提出了MonoDLE算法,重新审视2D检测框中心和3D目标投影中心的偏差. Liu等[57]在MonoDLE算法的基础上提出了MonoCon算法,添加了辅助学习模块,提升了模型的泛化能力. Lu等[58]提出了GUPNet,利用几何关系衡量深度估计的不确定度. Wang等[59]提出了一种新的轻量级方法PCT算法,以便于学习坐标表示,引入了一种具有置信度感知损失的本地化增强机制,以逐步细化预测.

此外,Chen等[60]提出了MonoPair,尝试利用配对约束来建立车与车之间的距离关系,通过全局优化来解决遮挡可能带来的预测问题. Chen等[61]结合不确定性理论提出了使用自监督学习的MonoRUn,对2D检测模块后的兴趣区域增加类似的3D学习分支,使其能够学习出稠密的物体坐标,从而建立几何与2D-3D的一致性,通过采用改进的PnP求解器来恢复3D位置.

3) 深度估计.

在基于深度学习的图像深度估计[62]的基础上,许多3D目标检测算法将这些深度估计算法视为其自身网络的子模块. 深度估计可以弥补单目视觉的不足,更准确地检测物体的三维信息. Xu等[63]提出了MF3D算法,通过子网络生成深度图,并将目标感兴趣区域与深度图进行融合以回归目标3D位置信息. Qin等[64]提出了MonoGRNet,引入一种全新的实例深度估计算法,利用稀疏监督预测目标3D边框中心的深度. 不同于MF3D生成整个输入图像的深度图方法,该方法只对目标区域进行深度估计,避免了额外的计算量. Ku等[65]提出的MonoPSR则是利用相机成像原理计算图像中像素尺寸与3D空间之间的比例关系,估计目标的深度位置信息. Zhang等[66]提出的MonoFlex将通过高度比预测物体深度的思想与不确定性理论相结合,构建了一种集成学习的方式,实现对物体中心位置的估计. Peng等[67]提出了IDA-3D算法,不依赖于将深度图像作为额外通道输入,而是引入实例深度感知模块来预测3D框中心深度,从而提高3D目标检测的准确性. Chen等[68]提出的Dsgn从双目图像对中提取像素特征用作立体匹配,使用高级特征用作物体识别,提出的网络同时估计深度并检测3D物体. Ding等[69]提出的D4LCN,从基于图像的深度图中自动学习卷积核和视觉感受野,并且单独使用在对应图片对应通道的对应像素上,可以处理不同大小的目标. Reading等[70]提出的CaDDN算法不同于传统的连续深度预测模型,提出了离散化深度概念,将每个像素的深度概率离散化分配在不同的深度桶中,避免了网络过度依赖于深度的准确检测. Lian等[71]提出的MonoJSG算法将单目对象深度估计重新表述为渐近细化问题,并提出了联合语义和几何代价量来对深度误差进行建模.

考虑到基于图像方法的3D检测只从前置摄像头检测物体而忽略了车辆周围的物体,Payen等[72]选择了比激光雷达角度分辨率更高、成本更低且能提供丰富的场景色彩和纹理信息的车顶全景相机来获取360°环境的图像数据,并提出了EBS3D算法,通过估计360°全景图像的密集深度图实现3D目标检测. 由于缺乏用于自动驾驶的全景标记数据集,他们使用投影转换来适应KITTI数据集,并提供了合成数据集上的基准检测结果.

4) 伪激光点云.

虽然深度信息有助于3D场景的理解,但简单地将其作为RGB图像的额外通道并不能弥补基于单目/立体图像的方法和基于点云的方法之间的性能差异. 相比之下,点云数据要比估计的深度精确得多,于是,Wang等[73]提出了伪激光点云算法Pseudo-lidar,分别采用单目深度估计算法DORN[74]以及3种立体深度估计算法PSMNet[75]、 DispNet[76]和SPS-stereo[77]进行深度估计,将得到的像素深度反投影为3D点云,从而形成了伪激光点云数据,如图9所示. 最后,利用已有的基于点云的检测算法Frustum PointNets[78]进行3D框检测. You等[79]在其基础上提出了Pseudo-lidar++,在初始深度估计的指导下,将测量数据分散到整个深度图中以提高检测精度,并利用更加便宜的4线激光雷达来代替64线激光雷达以微调检测结果. Ma等[80]提出了AM3D,引入注意力机制进一步增强了用于描述三维物体的特征识别能力. 考虑到来自立体匹配获取到的点云保留了物体边界的条纹伪影,从而影响边框估计的结果,Pon等[81]提出了OC-stereo,试图通过仅在相关的2D边框

区域上估算视差来解决这个问题. 然而,该方法需要在左右图像中都成功地进行2D检测,这对于在图像边界上被截断或在同一视图上被遮挡的物体而言是很困难的,并且完全忽略了为3D场景提供上下文的背景像素信息. 于是,Li等[82]提出了CG-Stereo,在立体匹配网络中对前景像素和背景像素分别使用深度解码器来提高前景像素的深度估计精度,并使用来自立体匹配算法的置信度估计作为一种软注意机制,引导目标检测网络更多地关注具有更高质量深度信息的点,从而进一步提高检测精度.

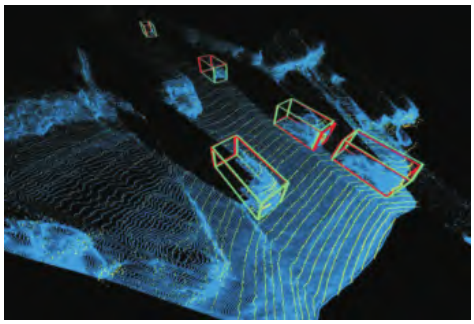
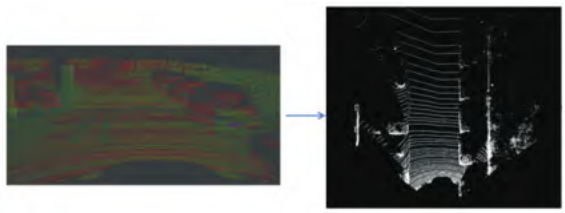


图9　激光雷达(蓝色)与激光雷达(黄色)[73]

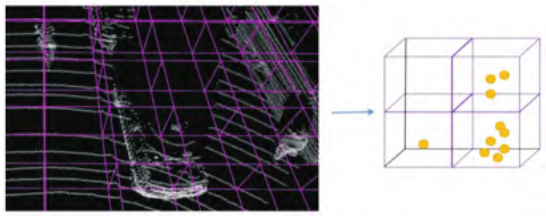
综上所述,由于2D相机比复杂的3D采集传感器更便宜且更灵活,基于单目/立体图像的方法已得到广泛研究. 二维图像以像素的形式提供了对象丰富的颜色和纹理信息,然而,2D图像的缺点是缺乏深度信息,这对于准确的物体大小和位置估计(尤其是在弱光条件下)以及检测远处和被遮挡的物体是必需的. 尽管有这个限制,使用基于2D图像的方法进行3D目标检测,对于经济的3D目标检测系统而言变得很重要. 由于图像缺少深度信息,研究深度估计算法是提高检测精度一种可行的解决方法. 除此之外,近几年伪激光点云的方法随着纯点云方法的发展越来越流行,但同样依赖于对图像进行深度估计的准确性.

2.2　基于点云的3D目标检测

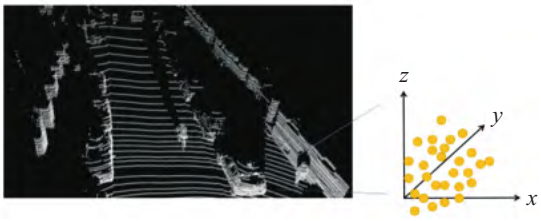
激光雷达获取的点云数据是三维坐标系中点的集合,通常由x、y、z坐标和反射强度定义. 点云能够提供比图像更加精确的深度信息,能够有效缓解图像中常见的遮挡问题. 点云具有无序性、稀疏性以及空间转换不变性等特征,为了能够有效提取其特征信息,通常将其处理为二维视图、体素,或直接以原始点的形式输入到网络中,表示形式如图10所示. 本节根据对点云数据的处理形式将基于点云的3D目标检测方法分为4种类型:基于二维视图的方法、基于体素的方法、基于空间点的方法和基于点与体素相结合的方法.



(a)　二维视图(鸟瞰图)表示



(b)　体素表示



(c)　点表示

图10　点云数据表示形式

1) 基于二维视图的方法.

此类方法将三维点云压缩到二维视图中,以减少处理三维点云数据的高计算成本,并提出了几种投影方案,如前视图(font view,FV)、距离视图(range view,RV)和鸟瞰图(bird's eye view,BEV). BEV是一个自顶向下的场景视角,这意味着数据是沿着y轴压缩的,如图10中(a)所示. FV和RV分别沿着z轴和x轴压缩.

Li等[83]提出了VeloFCN算法,将点云转换为2D前视图,利用现成的2D检测器实现了3D检测. 为了缓解遮挡问题,Yang等[84]提出了PIXOR算法,将等距点云进行离散化,并对反射率进行编码,得到了鸟瞰图表示,再使用全卷积网络[85]来估计目标的位置和航向角. Yang等[86]提出了HDNET算法,利用高画质地图映射提供的几何和语义先验信息来提高PIXOR的鲁棒性和检测性能. 由于高清地图并非随处可用,他们提出了在线地图预测模块,以便从单个点云中估计地图先验,不过这种方法在指向不同密度的点云时泛化性能较差. Beltrán等[87]提出了BirdNet算法,将输入点云投影得到三通道鸟瞰图,三通道依次为高度、强度和密度,其中密度进行归一化预处理,对每个单元中包含的最大点数进行编码,然后使用Faster R-CNN在鸟瞰图上实现目标检测. Simon等[88]提出了C-YOLO算法,使用单阶段YOLO网络在鸟瞰图上

进行检测,并达到50 FPS的运行速度. Meyer等[89]提出了LaserNet算法,将点云转换为二维距离图,利用全卷积网络来预测检测结果. 在距离图中,数据是密集的,能够学习到更加丰富的特征,但存在尺度变化和遮挡问题. 为了克服这个问题,Bewley等[90]提出了RCD算法,根据目标距离远近来动态调整卷积核大小,从而有效地处理尺度变化问题,再进行3D候选框优化来解决遮挡问题. 这类方法在生成二维视图时可能会忽略压缩轴上的大量信息. 因此,对于行人、路标、立交桥下的物体等来说,这可能不是一个可行的选择,因为从这样的角度来看,这些列举的案例可能只是在一定高度采样后的几个点,这显然不利于网络特征提取.

2) 基于体素的方法.

将点云数据转换为体素是处理不规则点云数据的常用方法,如图10(b)所示. 基于体素的方法通过3D卷积神经网络有效地提取点特征进行3D检测,在计算上虽然是有效的,但代价是由于离散化过程中的信息丢失而降低了细粒度定位精度.

Li[91]提出了3DFCN算法,将点云离散成具有长度、高度、宽度和通道尺寸的4D张量的体素表示,并将基于2D全卷积网络的检测技术扩展到3D空间域中进行3D目标检测. Engelcke等[92]提出了Vote3Deep算法,利用中心点对称的投票机制为每个非空体素生成一组投票,并通过累积投票的方式获得卷积结果,以处理输入点云的稀疏问题. 由于三维空间中存在大量的体素,整个检测和定位过程非常耗时. 另外,传统的3D卷积不能很好地学习不同尺度的局部特征. Zhou等[93]提出了代表性算法VoxelNet,将点云按长×宽×高的比例划分为一定数量的体素,并把点云空间中的点划分到位置所对应的体素中,对每一个非空体素使用若干个体素特征编码层进行局部特征提取. Yan等[32]提出了SECOND算法,通过稀疏卷积来代替3D卷积操作,从而达到减少计算量和提高训练速度的效果. Deng等[28]提出的Voxel r-cnn是在3D特征上聚合3D结构信息,使用粗粒度的体素完成高精度检测. Lang等[33]提出了PointPillars算法,将点云按柱状(长×宽)划分,减少了需要处理的体素数量,提升了检测速度. 考虑到不同的体素比例划分会影响检测的精度和计算时间,Ye等[94]提出了HVNet算法,通过改变多组体素划分的参数大小,对不同尺度的信息做特征提取,再进行多尺度的特征映射和融合. Liu等[95]提出了TANet算法,使用堆叠的三元注意力模块分别处理每个体素,增强目标的关键信息,同时抑制不稳定的点以获得更具判别性的特征表示,从而改善现有算法对行人等小目标的检测率低以及算法稳定性不高的问题. Ye等[96]认为形状先验在3D目标检测中应得到进一步增强,提出了SARPNET算法,设计3D形状注意模块用于学习物体的3D先验形状. Mao等[97]提出了一种基于transformer的3D主干网络VoTr,可作为标准稀疏卷积层的替代方案. Hu等[98]提出的PDV通过体素点质心有效地定位来自3D稀疏卷积主干的体素特征.

一些体素方法也利用到Anchor-free的思想,例如,Yin等[99]提出的CenterPoint算法,在第1阶段预测3D框的中心点,并回归其大小、方向和速度;第2阶段利用中心点特征回归检测框的得分并进行优化. Chen等[100]提出的HotSpotNet算法,对场景进行体素化处理并把体素中每一个点的特征集合起来,根据原始点到特征图的映射关系和标注框来决定生成hotspot表示,这些hotspot被用来直接训练预测物体的定位和方向.

3) 基于空间点的方法.

点云作为一种不规则的数据,如图10(c)所示,将其转换成规则的3D体素或者二维视图会造成数据不必要地冗长,并掩盖了原本的数据自然不变性. 出于这个原因,近几年提出了一系列点云处理网络,如PointNet[101]、PointNet++[102]等,直接从原始点中获取空间几何特征,再根据提取的特征对感兴趣的目标进行分类和定位.

Shi等[29]提出的PointRcnn算法,第1阶段利用PointNet++提取点云特征,将整个场景的点云分割为前景点和背景点,并从前景点中生成少量3D候选框;第2阶段将每个候选框池化的点转换为规范坐标,从而更好地学习局部空间特征来优化3D框. Yang等[103]提出的STD算法,使用基于点的球形锚框来生成更精确的候选框,由于球形锚框包括任何角度,可以削减锚框的生成数量,从而大大减少了计算量. 考虑到PointNet++中所使用的基于欧氏距离的最远点采样方法会产生大量的背景点,从而影响检测的效率,Yang等[24]提出了3D-SSD算法,采用一种基于特征距离的最远点采样法,通过结合语义信息排除大量背景点. 为了避免完全使用基于特征距离的采样而造成一定的冗余,他们选择结合基于欧氏距离和基于特征距离最远点采样的方法,并去除了PointNet++中非常耗时的FP模块和优化模块,使计算损耗大大减少. 现有的采样策略通常以距离作为标准,选择较远的点来尽可能覆盖整个场景,但是,这样会导

致keypoints包含过多的背景点,Chen等[104]提出的SASA通过引入点级的语义信息,避免了语义增强模块选择较多的背景点. Xu等[105]提出的BtcDet发现通过对遮挡缺失的补全可以提升性能,通过预测RoI的形状占有率,将其整合到点云特征中再进行目标检测. Xu等[106]提出的SPG首先在预测的前景点区域生成语义点集,然后将语义点集与原始点云相结合得到增强点云,最后再使用一个点云检测器得到检测结果.

在三维空间中,点之间共享的上下文特征仅限于点之间的关系,形状仍然缺乏结构性. 图卷积(GCN)[107-109]如图11所示(其中:conv为卷积操作,pool为池化操作,label表示分类标签,每种颜色表示一个子图,在池化后形成新的节点),可通过沿边缘聚集特征迭代更新顶点特征,通常不需要重复采样和分组顶点便能学习点的形状特征. 因此,Zarzar等[110]提出了PointRGCN算法,首次利用图卷积作为点云的特征提取网络,实现了3D目标检测. Shi等[111]提出了Point-GNN算法,利用自动配准的图神经网络更好地学习每个顶点的特征.

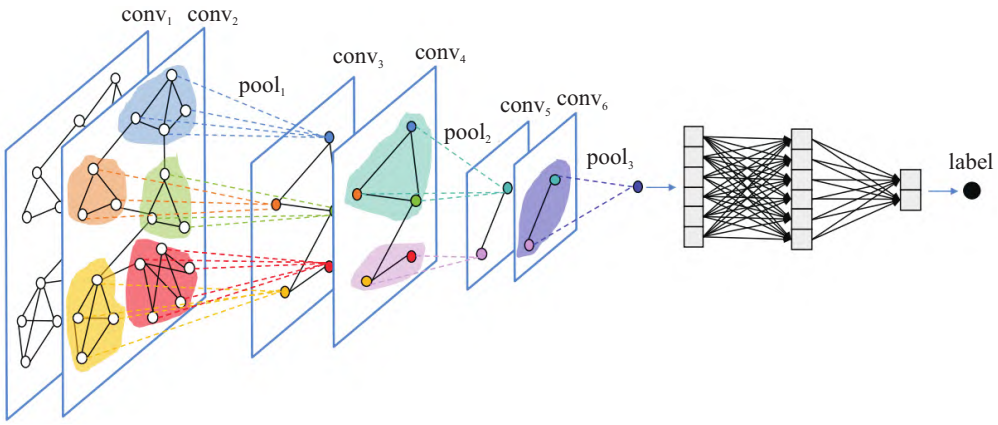


图11　图卷积神经网络实现分类任务示意图[107]

目前,由于3D目标检测数据集制作耗时费力,许多学者开始研究减少依赖于真实3D检测框的方法. Ren等[112]提出的WyPR方法是一个用于点云识别的弱监督框架,只需要场景级的类别标签作为监督,结合标准的多实例学习目标,WyPR可以检测和分割点云数据中的物体,而无需在训练时获得任何空间标签. Zhao等[113]提出的基于自封闭半监督点云的三维对象检测框架SESS,不需要大量通常难以获得的强标签,设计了一个特定的基于点的数据扰动方案和3种一致性损失,使网络能够生成更准确的检测. Qin等[114]提出了VS3D,利用一种跨模态知识蒸馏策略,通过利用标准化的点云密度来生成目标提案.

由于点云的稀疏性,单帧检测可能会受到一些限制,例如遮挡、远距离和不均匀采样等情况,从而导致性能下降. 然而,点云视频中包含了丰富的前景物体时空信息,于是,Yin等[115]提出了一种处理点云序列的视频检测算法PMPNet. 网络在PointPillars[33]的基础上添加了图卷积,使得每个节点的感受野扩大,以此设计了空间特征提取模块用于提取独立的每一帧点云特征,然后将空间特征送入时空特征融合模块,得到连续帧之间更加丰富的特征信息用于实现3D检测.

4) 基于点与体素相结合的方法.

通常,基于体素的方法计算效率更高,但会丢失局部信息,导致降低定位精度. 基于原始点云的方法计算成本较高,但更容易通过点抽象获得更大的感受野,从而得到更加精确的定位.

Shi等[116]结合基于原始点云和基于体素特征学习的优点,提出了PV-RCNN算法,通过体素划分操作获取高效编码的多尺度特征并生成高质量3D候选框,再利用原始点云可变的感受野来保留更精确的位置信息. 考虑到PV-RCNN取得成功的部分原因是通过随机采样关键点来捕获多尺度特征,改进候选框和定位信息,然而,随机抽样对潜在的模糊场景并不有效,于是,Bhattacharyya等[117]在PV-RCNN的基础上提出了Deformable PV-RCNN,通过收集不均匀分布的上下文信息来提升对于远距离目标的检测性能. He等[25]发现点云的部分空间信息会随着网络中分辨率逐步缩小的特征图丢失,为了处理这个问题,他们提出了SA-SSD算法,在主网络中加入辅助网络,将各个分辨率的特征层上下不为0的特征点通过体素到点反栅格化映射到三维空间,再采用反距离加权法插值来获取每个点的特征,将不同分辨率下的点

特征进行融合. Zheng 等[118] 提出的 CIA-SSD 算法校准单步目标检测中分类和定位两个任务. Zheng 等[26] 提出的 SE-SSD 算法利用具有公式化约束的软目标和硬目标来联合优化模型,而不在推理中引入额外的计算. Noh 等[119] 设计了一个双流编码器 HVPR 来分别提取体素和点的特征,对于每个体素特征,根据其相似性聚合点特征,并得到体素-点混合表示. Chen 等[120] 提出了 Fast Point R-CNN 算法,第 1 阶段,将点云进行体素化,通过卷积操作生成少量 3D 候选框,并采用注意力机制结合坐标信息和对应的每个点卷积后的特征,以达到保留上下文信息和准确坐标位置的效果;第 2 阶段,在上述得到的混合特征上进行优化.

综上所述,基于二维视图的方法可以提高检测效率,但会造成数据不必要地冗长并丢失 3D 空间信息. 基于体素的方法很容易适应具有显著精度和延迟的高效硬件实现,但体素化过程对体素的长度、宽度和高度等参数的选择更为敏感,而且考虑到超出体素占用的许多点将被丢弃,不可避免地会遭受量化损失. 基于空间点的方法更合理地保留了点云的原始几何信息,但其提取点云特征过程比体素化更耗时. 采用图卷积可以捕捉到点之间的相关性和更为复杂的空间局部结构,即“顶点-边缘”信息,使得 3D 形状将更容易被感知,但计算消耗量更大. 基于点与体素相结合的方法可以得到更精确的定位,但同时也会增加额外的计算消耗.

2.3 基于图像与点云融合的 3D 目标检测

基于图像的处理方法只提供纹理信息,不提供深度信息;而基于点云的处理方法提供深度信息,但缺乏纹理信息. 纹理信息对于目标检测和分类很重要,而深度信息对于准确估计 3D 位置和目标大小至关重要. 此外,随着与传感器的距离增加,点云密度按比例减少. 由于纹理和深度模式在 3D 目标检测中是必不可少的,将两者互补信息相融合,理论上可以获得更好的检测效果. 本文根据对图像和点云数据不同的处理流程,将 3D 目标检测分为顺序融合的方法和并行融合的方法,如图 12 所示.

1) 顺序融合.

这类方法以顺序的方式提取图像和点云特征,并且后阶段的特征提取严重依赖于前阶段. 将前阶段提取到的特征作用到后阶段有两种方式,分别为视锥和投影,如图 13 所示.

① 利用视锥融合的方法. Qi 等[78] 提出了 F-Point Net 算法,首次利用视锥的方法实现了 3D 目标检测. 首先通过优秀的 2D 目标检测算法生成候选区域,

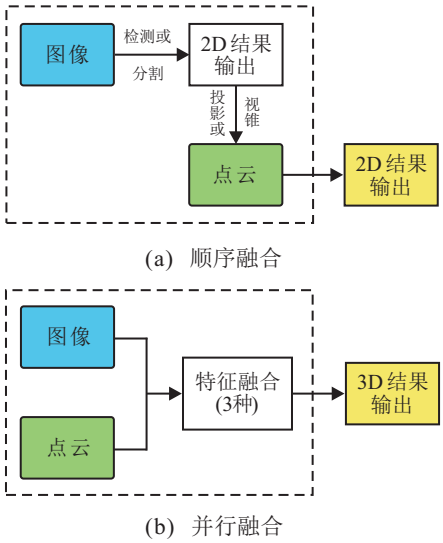


图 12 图像和点云两种融合策略

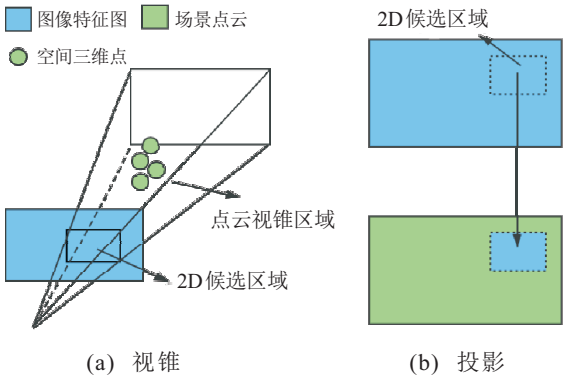


图 13 顺序融合中两种特征融合方式

结合候选区域的深度信息提取 3D 视锥;然后对其用 PointNet 进行 3D 实例分割;最后利用 T-Net 网络对坐标进行变换,使得视锥体的中心轴与图像平面正交并预测最终的 3D 包围框. Wang 等[121] 提出了 F-ConvNet 算法,将视锥体进行分块,并对每个视锥体使用 PointNet 做特征提取;然后排列成一个 2D 的特征图,送入到全卷积网络中实现分类和 3D 框回归.

② 介绍利用投影融合的方法. Yang 等[122] 提出了 IPOD 算法,利用 2D 语义分割网络对图像数据做语义分割,再将预测到的 mask 投影到点云上过滤背景点,利用 PointNet++ 网络对每个前景点预测分类得分和 3D 框. Vora 等[123] 提出的 PointPainting 算法与其类似,将融合了语义分割信息的点云传递到已有的点云检测器 (PointRCNN 或 PointPillars) 进行训练回归. Yin 等[124] 提出的 MVP 算法则是利用图像实例分割结果,对激光点云做了稠密化. Sindagi 等[125] 提出的 MVX-Net 算法,先利用 2D 卷积网络提取图像特征,再对点云进行体素化,并分别尝试将图像特征融入到体素中的点上和融入到经过 VFE 层编码过后的体素特征中来实现最终的 3D 检测.

2) 并行融合.

此类方法对图像和点云的数据进行并行特征提取,其中特征融合的方法可分为感兴趣区域融合、特征图后融合和连续特征融合,如图14所示.

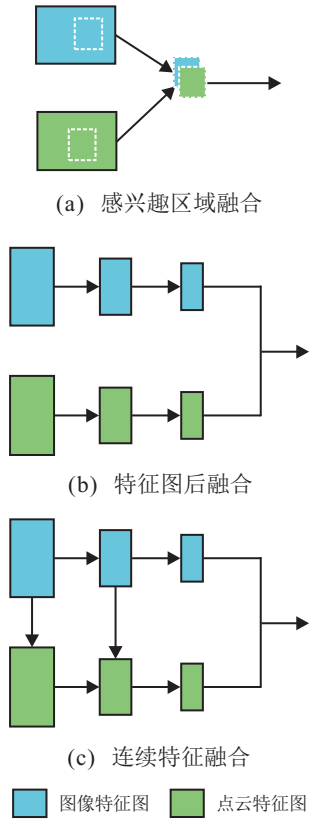


图 14　并行融合中3种特征融合方式

① 感兴趣区域特征融合的方法. Chen等[30]提出了MV3D算法,先将3D点云投影到前视图和鸟瞰图,其中鸟瞰图被编码成高度、强度和密度信息,并利用前视图对鸟瞰图提供了互补的信息. 输入的每个视图进行特征提取,结合基于区域的特征融合实现分类和3D框回归. 由于算法实现过程中特征图进行下采样导致小目标的空间信息丢失,从而无法准确检测出小目标,对此,Ku等[31]提出了AVOD来解决这个问题,通过使用FPN网络进行特征提取和1×1卷积操作进行降维处理,从而在一定程度上解决了小目标空间信息丢失问题. 另外,网络在数据输入中舍弃了MV3D中鸟瞰图的强度信息和前视图,减少了一个分支的预处理量和后续融合的计算量. 根据实验表明,RGB图像和鸟瞰图足够很好地诠释了3D空间中的信息. 由于KITTI数据集中骑行者和行人类别中物体的大小和长宽比不同,目前的目标检测器在提出合适的锚大小方面的性能较差,导致最终的目标检测性能较低. Raffiee等[126]提出的CAP-AVOD扩展了AVOD框架,添加了特定的锚框,通过解决类中对象大小和长宽比变化较大的问题来改进锚定策略. Zhu等[127]提出的CM3D,第1阶段主要目的是通过稀疏点云的特征融合产生3D候选框,第2阶段主要融合2D和3D候选框区域的密集特征;还提出了一种数据增强方法来增强点云密度,防止遥远的目标因过少的点而导致识别遗漏.

与在二维视图中提取候选框不同,Xie等[128]提出的PI-RCNN算法,直接对原始点云进行特征提取并生成候选3D框,图像部分进行语义分割,并提出PACF模块将图像的语义特征与点云特征相融合,得到的融合特征信息输入到检测网络中以实现分类和边框回归. 稀疏点云由于缺乏几何和语义信息造成检测性能下降,为了提高远处和被遮挡的点云的检测质量,Wu等[129]提出了一种新型多模态融合框架SFD,其利用深度补全将2D图片转换成3D伪点云,从而统一了图像和雷达点云的数据表达方式.

② 特征图后融合的方法. Xu等[130]提出了Point Fusion算法并给出了两个版本:第1个版本通过融合图像特征和点云的全局特征直接回归3D框的8个顶点位置;第2个最终采用的版本,额外融入了点云的局部特征用于预测每个角点相对于中心点的偏移量. Yoo等[131]提出的3D-CVF算法,第1阶段,为了覆盖更广阔的视野,对6个视角下所采集到的图像分别进行特征提取,并采用能够校正空间偏移的插值投影(自校准投影)将6个图像特征映射转化为平滑且密集的鸟瞰图特征. 点云通道,先对原始点云进行体素化,利用Voxelnet进行编码,再通过3D稀疏卷积提取点云特征. 之后利用注意力机制来权衡不同模态特征的重要性,融合图像和点云特征图. 第2阶段,利用图像和点云融合特征图生成3D候选区域,再应用感兴趣区域池化进行优化. Li等[132]提出的DeepFusion方法,为了实现激光雷达点与图像像素的精确几何对齐,使用了反转与几何相关的增强,并利用交叉注意力动态捕获图像与激光雷达之间的相关性融合特征.

③ 连续特征融合的方法. Liang等[133]提出了ContFuse算法,分别在图像和点云生成的鸟瞰图上进行特征提取,并利用PCCN[134]将不同尺度下的图像特征图投影到其同尺度下的鸟瞰特征图上,用以实现最终的3D检测. Liang等[135]在此基础上进一步提出了MMF算法,添加了两个额外任务:地面估计和深度补全. 地面估计任务是使用UNet模型估计点云相对于地面的高度值,作为额外通道附加到鸟瞰图中再输入到特征提取网络中;深度补全任务是根据激光雷达相机外参把点云投影到成像平面上,得到稀疏深度图并作为额外通道附加到RGB图像中,再一起输

入到特征提取网络中. Piergiovanni等[136]提出的4D-Net算法能同步利用3D点云和RGB传感信息,通过在不同特征表示和抽象层次上进行新的动态连接学习以及观察几何约束纳入4D信息,还能够更好地使用运动线索和密集的图像信息更成功地检测远处的物体. Huang等[137]提出了Epnet算法,用逐点的方式使用语义图像特征来提高点的特征,不需要任何图像标注,并且使用一致性强制损失促进定位和分类置信度的一致性.

综上所述,顺序融合的方式,即当前阶段取决于前一个阶段的方式,对于内存要求高,模型架构通常无法进行端到端的训练. 不同阶段之间更紧密的耦合是所有基于顺序融合的方法都存在的典型特征之一,即前一阶段的性能不佳可能会恶化其余阶段. 此外,信息丰富的中间特征在很大程度上被抛弃了,这对于目标检测来说似乎是至关重要的. 基于并行融合的方法将特征集成在一种多模态表示中,只需要一个学习阶段,但是,传感器之间的视图错位问题不容易解决. 目前,基于图像与点云融合的方法并没有优于纯点云的方法,主要是因为原始数据存在噪声、未充分利用信息以及多模态传感器的错位. 因此,想要实现相当好的性能还需要进一步研究融合策略,使其能够兼顾各数据模态的优点,并在有效保留原始信息的同时实现深层次融合.

3 评 估

常见的3D目标检测评估标准包括:交并比IoU (intersection-over-union)、查准率P、查全率R、平均精度AP、多类别平均精度mAP (mean average precision).

1) 交并比(IoU).

交并比表示产生的候选框与真实框之间的重叠率. 交并比的值越高,检测的精度越准确. 可通过下式计算:

$$\mathrm{IOU}=\frac{\mathrm{BBox}_{\mathrm{pred}}\bigcap\mathrm{BBox}_{\mathrm{gt}}}{\mathrm{BBox}_{\mathrm{pred}}\bigcup\mathrm{BBox}_{\mathrm{gt}}}. \tag{9}$$

其中:$\mathrm{BBox}_{\mathrm{pred}}$为检测算法预测的候选框区域,$\mathrm{BBox}_{\mathrm{gt}}$为真实标记的候选框区域. 图15给出了IoU的概念示意.

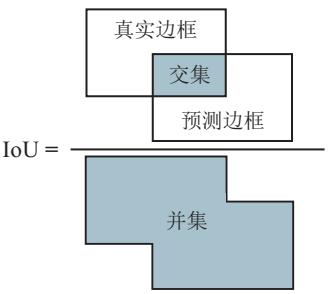


图15 IoU概念图示

2) 查准率P和查全率R.

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

$$P=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}, \tag{10}$$

$$R=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}. \tag{11}$$

其中:TP为判断检测结果中被正确识别的正样本的数量,TN为被正确识别的负样本的数量,FP为被错误识别为正样本的负样本数量,FN为被错误识别为负样本的正样本的数量,这些变量可以通过设定IoU的阈值t来调整. 根据计算所得到的查准率和查全率值可绘制precision-recall (P-R)曲线,平均精度AP即是P-R曲线下方的面积. 其面积值越大,代表分类检测的效果越好.

3) 多类别平均精度(mAP).

mAP即为多类别AP的平均值. 为了减少计算量,PASCAL VOC数据集[138]采用插值方式计算AP值,在[0,1]上以步长为0.1等间距取查全率上的查准率值. 可通过下式计算:

$$\mathrm{AP}=\frac{1}{11}\sum_{r\in 0,0.1,\ldots,1}P_{\mathrm{interp}}(r). \tag{12}$$

其中:r表示查准率值,$P_{\mathrm{interp}}(r)$表示插值函数.

3.1 KITTI数据集评估方法

目前,大多数3D目标检测工作都是使用KITTI数据集进行训练和评估,这为算法性能比较提供了基准. KITTI的评估指标包括2D检测平均准确率、3D检测平均准确率以及平均角度相似性AOS. 采用2D目标检测的AP计算方式,将世界坐标系下3D检测框投影到图像坐标系下,通过IoU计算AP值,且不同的目标类别设置的IOU阈值不同. 对于车辆来说,如果3D检测框与真实标记的3D框之间重叠率超过0.7,则认为预测框是准确的. 对于行人和骑车者的类别IoU的阈值则设置为0.5. 但是,检测得到的位于世界坐标系下不同位置、不同大小的目标3D框,被投影到2D图像上可能会得到相同的检测框. 因此,图像坐标系下的坐标并不能直接表示3D检测框的准确性. 为了解决这个问题,Chen等[30]在MV3D中引入了$\mathrm{AP}_{\mathrm{Ioc}}$指标,将3D检测投影到鸟瞰图上. 与AP计算过程中不同之处在于,其IoU值为世界坐标系下3D检测框值与真实值之间的重叠率. 针对3D目标检测任务,KITII数据集定义了AOS (average orientation similarity)指标,用来评价目标航向角预测的结果.

平均方向相似性AOS被定义为2D检测器的平

均准确率与方位角方向的平均余弦距离相似度的乘积,可通过下式计算:

$$\mathrm{AOS}=\frac{1}{11}\sum_{r\in\{0,0.1,\ldots,1\}}\max_{\tilde{r}:\tilde{r}\geqslant r}s(\tilde{r}). \tag{13}$$

其中:r为2D目标检测的查全率,$s(\tilde{r})$为方向相似性. $s(r)$可通过下式计算:

$$s(r)=\frac{1}{|D(r)|}\sum_{i\in D(r)}\frac{1+\cos\Delta_\theta^{(i)}}{2}\delta_i. \tag{14}$$

其中:$D(r)$表示在查全率r下所有预测为正样本的集合,$\Delta_\theta^{(i)}$表示检出目标i的预测角度与真实值的差. 为了防止多个检出匹配到同一个真实值,如果检出目标i已经匹配到真实值(IoU至少为50%),则设置$\delta_i=1$;否则$\delta_i=0$.

3.2　NuScenes数据集评估方法

NuScenes数据集的检测任务要求检测10类带有3D检测框的目标,包括汽车、卡车、公交车、行人、自行车等. 并提出平均精度指标(average precision metric, AP)、真正类指标(true positive metrics, TP)和nuScenes检测得分(nuScenes detection score, NDS)三种新的评估指标.

1) 平均精度指标(AP).

数据集使用平均精度(AP)进行评估,但是,定义匹配的阈值是2D中心距地平面的距离,而不是交并比. 这样做的目的:其一是为了使检测与物体的大小和方向分离开;其二是因为像行人和自行车这样占地面积小的对象,如果检测到一个小的平移误差,则会给出交并比为0的结果,这使得那些倾向于具有大定位误差的纯视觉方法的性能变得难以比较. 然后,计算出在查全率和查准率超过10%的情况下,精确查全率曲线下的归一化面积. 召回率或准确率低于10%的点将被删除,以最大限度地减少低准确率和召回率区域常见的噪声影响. 如果在该区域没有达到阈值点,则设置该类别的AP为零,并将匹配阈值$D=0.5,1,2,4\,\mathrm{m}$和类别C的集合进行平均. 可通过下式计算:

$$\mathrm{mAP}=\frac{1}{|C||D|}\sum_{c\in C}\sum_{d\in D}\mathrm{AP}_{c,d}. \tag{15}$$

其中:C为目标检测类别的集合,D为匹配的阈值,AP表示平均精度.

2) 真正类指标(TP).

除AP之外,还为每个与地面实况框匹配的预测框设置一组TP度量指标. 所有TP指标都是在匹配期间使用$d=2\,\mathrm{m}$中心距离计算得出的,并且都设计为正标量. 匹配和评分在每个类别中独立进行,每个指标是所有达到10%以上的召回水平时的累积平均值的均值. 如果特定类别的召回率达不到10%,则该类别的所有TP错误均设置为1.

这些TP错误定义如下:平均平移误差ATE表示二维欧几里得中心距离;平均尺度误差ASE表示调整方向和平移后的3D交并比;平均方向误差AOE是预测值与真值之间的最小偏航角;平均速度误差AVE是以2D (m/s)为单位的速度差的L_2范数的绝对速度误差;平均属性误差AAE定义为1减去属性分类精度$(1-\mathrm{acc})$. 对于每个TP指标,计算所有类别的平均TP指标(mTP),可通过下式计算:

$$\mathrm{mTP}=\frac{1}{|C|}\sum_{c\in C}\mathrm{TP}_c. \tag{16}$$

其中:C为目标检测类别的集合,TP表示真正类指标.

3) nuScenes检测得分(NDS).

带有IoU阈值的mAP可能是当前最常用的目标检测指标. 但该指标不能捕捉检测任务的所有方面,如垂向和属性估计. 因此,nuScenes将不同的错误类型合并为一个标量分数:nuScenes检测得分NDS,可通过下式计算:

$$\mathrm{NDS}=\frac{1}{10}\Big[5\mathrm{mAP}+\sum_{\mathrm{mTP}\in\mathrm{TP}}(1-\min(1,\mathrm{mTP}))\Big]. \tag{17}$$

其中:mAP是平均精度均值,TP是5个平均真实正指标的集合. 因此,NDS分数的一半是基于检测性能,另一半则根据检测框的位置、大小、方向、属性和垂向来量化检测的质量. 由于mAVE、mAOE和mATE可以大于1,在式(17)中将每个指标限定在0～1之间.

4　结　论

本综述针对自动驾驶任务中的3D目标检测方法进行了总结与分析. 首先,介绍了自动驾驶感知周围环境所需相关传感器,作为感知算法的3D目标检测算法所需的数据集、基础模型、算法的输出表示以及训练模型所需要的损失函数;其次,系统地梳理了基于不同传感器数据的3D目标检测算法的研究进展,将其分类为基于单目/立体图像、点云以及图像与点云融合的方法;最后,介绍了自动驾驶相关数据集及其3D检测评估指标,并在KITTI数据集和NuScenes数据集上进行了大量的实验结果对比. 3D目标检测作为自动驾驶中的关键任务,近几年来得到了飞速的发展. 但是,3D目标检测算法受复杂环境和不同传感器自身局限性的影响仍然存在许多的难题和挑战. 结合本文对该领域算法研究现状的回顾及实验结果对比,未来可能的研究趋势包含以下几个方

面.

1) 输入数据处理.

目前,点云输入数据的处理形式主要分为两类:2D视图形式(鸟瞰图BEV、距离图RV和前视图FV)、体素或柱状形式. 例如,C-YOLO[88]将原始点云数据压缩成2D鸟瞰图,PointPillars[33]将其划分为柱状以减少大量计算元素,从而提高网络的处理效率. 最近,FCOS-LiDAR[139]与使用鸟瞰图的主流方法不同,利用从激光雷达的距离视图检测物体,仅使用标准2D卷积的基于RV的3D检测器就可以实现与最先进的基于BEV的探测器相当的性能,而且检测器速度更快、更简单. 更重要的是:几乎以前所有基于RV的检测器都只关注单帧点云,单帧点云通常在3D空间中稀疏且分辨率较低;而该方法使用新的距离视图投影机制将多帧点云融合到单个距离视图,这是一个全新的研究思路.

针对图像数据,基于图像的3D目标检测的性能严重依赖于估计对象的精准距离的能力,近几年提出的利用估计所得的深度信息生成伪激光点云的方法,例如Pseudo-lidar[73]、Pseudo-lidar++[79]等,将3D目标检测和预训练的深度估计配对,提高了一定的性能,但研究还在初步阶段,深度和检测的方法仍然是完全独立的. 为了克服这一点,DD3D[140]将3D目标检测和深度估计两个任务一起训练,验证了深度估计和目标检测联合的潜力,或许会成为未来研究的一个新方向. 因此,为了满足实时应用所需的高效性和准确性,需要对输入的图像或点云数据的处理形式进行更充分地探索.

2) 特征提取策略.

如何捕捉到有利于检测的特征是至关重要的. 针对图像数据,MixConv[141]探索了多组不同的卷积核以提高模型的准确性和效率. 深度分离卷积核[142]能够加快卷积操作的速度,帮助满足实时需求和在低计算设备中进行模型集成. 除此之外,将注意力机制融入主干网络[143]或者将已知特征与自学习特征相结合[144]也是一种发展趋势. 针对点云数据,Pointnet系列是实现对原始点云的分类和分割的经典之作. 为了更好地处理点云数据,近年来不断涌现出一些新的方法,例如PointCNN[145]根据输入点学习一种X变换,然后将其用于同时加权与点关联的输入特征以及将它们重新排列成潜在隐含的规范顺序,之后再在元素上应用求积和求和运算. PointASNL[146]通过减轻离群点的偏差效应,进一步捕获采样点的区域和相关性,以提高点云处理任务中的鲁棒性和优越性.

此外,使用图卷积网络来处理原始点云的PointASNL[147]也是一种可行的方法. 图卷积引入可学习的卷积参数,可对顶点与边对应关系的拓扑图进行优化,有利于提取到更具结构性的点的形状特征,提高3D检测的效果.

3) 多传感器融合问题.

图像与点云融合需要考虑对两个不同测量空间的数据进行投影对齐时所产生的量化误差,跨模态异构特征的语义对齐是融合问题的难点. 目前,已有一些工作[128,133]利用双线性插值来提高图像与点云融合检测网络的性能. 另一种趋势则是将图像分割特征[123]直接附加到点云上,从而构建丰富的逐点输入数据以获得更好的检测效果.

此外,可以考虑其他形式的数据来进一步提高算法的准确性和鲁棒性. 例如,与激光雷达相比,雷达在一些极端天气条件下更稳定,而且具有更长的感应距离,可以提高远处物体的精度. Major等[148]首次提出了一种基于雷达的深度神经网络目标检测方法. CenterFusion[149]是一种利用雷达和摄像机数据进行3D目标检测的融合方法,专注于将雷达检测与图像中获得的初步检测结果相关联,然后生成雷达特征图,并将其与图像特征结合使用,准确估计物体的三维包围盒. 尽管低成本雷达在自动驾驶领域很受欢迎,但很少有研究关注于将雷达数据与其他传感器融合. 主要原因是包含自动驾驶雷达数据的数据集并不多[150],而且雷达数据更稀疏,无法有效提取物体的几何信息. 尽管利用聚合多个雷达扫描增加扫描点的密度,但同时也会给系统带来延迟. 总之,理想的检测算法应该整合各种数据,以覆盖复杂和极端的条件.

4) 数据分布不均衡.

泛化能力对自动驾驶汽车的安全性起着重要的作用,但是,3D目标检测数据集存在数据分布不均衡(长尾分布)的问题,如图16所示. 这个问题限制了模型在实际应用中的实用性,训练后的模型容易偏向训

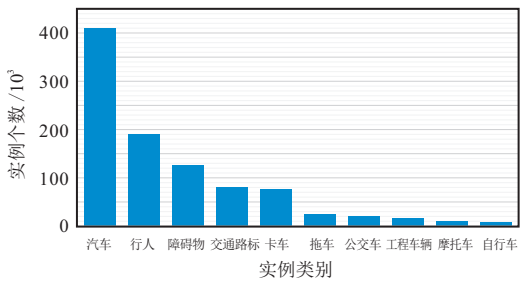


图16 NuScenes数据集不平衡分类

练数据量大的头部类,导致模型在数据量有限的尾类上表现不佳. 利用数据增强手段可以有效解决这个问题,提高检测性能和防止过拟合.

图像领域中,常用到的数据增强方法有随机翻转、缩放和旋转等. 在三维激光点云中,常用到剪切、旋转和加遮挡等方法,例如Yan等[32]将罕见类(如公共汽车)的目标复制粘贴到激光雷达场景中,从而生成新的训练数据. Zhu等[151]提出了类分组概念,让形状相似的类别划分为一个组,让该组中样本数量较多的类去提升样本数量较少的类的精度,并且每个组之间的总数接近,有助于减缓网络学习时数量较多类别主导整个训练的问题. 单独在图像或点云数据上进行数据增强操作较为容易,但对于同时利用两类数据的方法,则需要构建数据元素(如点或像素)之间的映射. 当在每个数据模态中执行数据增强操作时,需要有效地考虑数据之间的映射关系来获取更好的融合特征. Zhang等[152]提出了多模态剪贴的方法,通过切割点云和地面真实对象的图像斑块,并以一致的方式将它们粘贴到不同的场景中,避免了多模态数据之间的错位. 对点云进行平移、旋转、翻转等数据增强操作时按顺序记录相应的参数,在特征融合的阶段,把需要投影到图像的点按相反顺序转换到原始点云所在坐标系下进行点到像素的映射投影.

5) 对标注数据依赖性高.

创建3D目标检测数据集是一项极其昂贵且耗时的操作,通常涉及不同传感器数据处理技术以及大量劳动力之间的协同作用. 注释的精度要求很高,即使进行许多质量检查,也不可避免地受到错误的影响. 而且,目前大多数的3D目标检测方法都是完全监督的,即需要训练3D边框注释. 也有小部分在尝试使用半监督学习或自监督学习进行3D目标检测[106,112-114,153-156]. 在这些尝试中,值得强调的是Qin等[114]提出的VS3D方法,引入了一个无监督的三维提案模块,通过利用标准化的点云密度来生成目标提案,此外还提出了一种跨模态知识蒸馏策略,使在未标记的点云上训练三维目标检测器成为可能. 通过实施一些先进的自监督、无监督机制,使模型减少对真实3D边框注释的需求是非常有价值的.

使用视频数据也可能放松完全监督的要求. 目前,大多数工作都是从单帧的角度来解决3D目标检测问题的,只有最近的一项工作[157]开始考虑时间上的约束. 在深度估计领域已经有使用视频序列和运动信息实现自监督的深度估计方法[158],如果相似的监督也用于恢复对象的形状和外观,则可以用于3D目标检测任务. 因此,可以尝试通过引入时间数据在时空空间中建立新的约束实现新的突破.

过去20年间见证了自动驾驶技术的巨大进步,随着可用硬件设备计算能力的提高以及传感和计算技术成本的降低,自动驾驶汽车的相关技术趋于成熟. 无论是使用图像还是点云以及两种数据融合的方式进行3D目标检测,都为自动驾驶领域带来了大量的可能性,研究兼备准确性和实时性的3D目标检测算法一定能促进自动驾驶领域的进一步发展.

参考文献(References)

[1] Taeihagh A, Lim H S M. Governing autonomous vehicles: Emerging responses for safety, liability, privacy, cybersecurity, and industry risks[J]. Transport Reviews, 2019, 39(1): 103-128.

[2] Badue C, Guidolini R, Carneiro R V, et al. Self-driving cars: A survey[J]. Expert Systems with Applications, 2021, 165: 113816.

[3] Buehler M, Iagnemma K, Singh S. The DARPA urban challenge[M]. Berlin, Heidelberg: Springer, 2009: 1-622.

[4] Ozguner U, Stiller C, Redmill K. Systems for safety and autonomous behavior in cars: The DARPA grand challenge experience[J]. Proceedings of the IEEE, 2007, 95(2): 397-412.

[5] Shadrin S S, Ivanova A A. Analytical review of standard Sae J3016 ńtaxonomy and definitions for terms related to driving automation systems for on-road motor vehicleszż with latest updates[J]. Automobile Doroga Infrastruktura, 2019, 3(21): 1-9.

[6] 詹德凯. 自动驾驶汽车环境感知系统传感器技术现状及发展趋势[J]. 辽宁省交通高等专科学校学报, 2021, 23(3): 21-26.
(Zhan D K. The situation and trends of sensors for environmental perception system of autonomous vehicles[J]. Journal of Liaoning Provincial College of Communications, 2021, 23(3): 21-26.)

[7] Chen S H, Liu B A, Feng C, et al. 3D point cloud processing and learning for autonomous driving[J/OL]. 2020, arXiv: 2003.00601.

[8] Arnold E, Al-Jarrah O Y, Dianati M, et al. A survey on 3D object detection methods for autonomous driving applications[J]. IEEE Transactions on Intelligent Transportation Systems, 2019, 20(10): 3782-3795.

[9] Shen X K. A survey of Object Classification and Detection based on 2D/3D data[J/OL]. 2019, arXiv: 1905.12683.

[10] Feng D, Haase-Schütz C, Rosenbaum L, et al. Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges[J]. IEEE Transactions on Intelligent Transportation Systems, 2021, 22(3):

1341-1360.

[11] Zou Z, Shi Z, Guo Y, et al. Object detection in 20 years: A survey[J/OL]. 2019, arXiv: 1905.05055.

[12] Ahmad W A, Wessel J, Ng H J, et al. IoT-ready millimeter-wave radar sensors[C]. 2020 IEEE Global Conference on Artificial Intelligence and Internet of Things. Dubai, 2020: 1-5.

[13] Zou Z X, Shi Z W, Guo Y H, et al. Object detection in 20 years: A survey[J/OL]. 2019, arXiv: 1905.05055.

[14] 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. Providence, 2012: 3354-3361.

[15] 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). Seattle, 2020: 11618-11628.

[16] 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). Seattle, 2020: 2443-2451.

[17] Huang X Y, Wang P, Cheng X J, et al. The ApolloScape open dataset for autonomous driving and its application[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 42(10): 2702-2719.

[18] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]. 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, 2014: 580-587.

[19] Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection[C]. 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, 2016: 779-788.

[20] Redmon J, Farhadi A. YOLO9000: Better, faster, stronger[C]. 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, 2017: 6517-6525.

[21] Redmon J, Farhadi A. YOLOv3: An incremental improvement[J/OL]. 2018, arXiv: 1804.02767.

[22] Gkioxari G, Girshick R, Malik J. Contextual action recognition with R CNN[C]. 2015 IEEE International Conference on Computer Vision. Santiago, 2015: 1080-1088.

[23] 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.

[24] 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). Seattle, 2020: 11037-11045.

[25] 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). Seattle, 2020: 11870-11879.

[26] Zheng W, Tang W L, Jiang L, et al. SE-SSD: Self-ensembling single-stage object detector from point cloud[C]. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, 2021: 14489-14498.

[27] Kundu A, Li Y, Rehg J M. 3D-RCNN: Instance-level 3D object reconstruction via render-and-compare[C]. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, 2018: 3559-3568.

[28] 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.

[29] Shi S S, Wang X G, Li H S. PointRCNN: 3D object proposal generation and detection from point cloud[C]. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, 2019: 770-779.

[30] 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. Honolulu, 2017: 6526-6534.

[31] 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). Madrid, 2018: 1-8.

[32] Yan Y, Mao Y, Li B. SECOND: Sparsely embedded convolutional detection[J]. Sensors: Basel, 2018, 18(10): E3337.

[33] Lang A H, Vora S, Caesar H, et al. PointPillars: Fast encoders for object detection from point clouds[C]. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, 2019: 12689-12697.

[34] Chen X Z, Kundu K, Zhu Y K, et al. 3D object proposals using stereo imagery for accurate object class detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(5): 1259-1272.

[35] Chen X Z, Kundu K, Zhang Z Y, et al. Monocular 3D object detection for autonomous driving[C]. 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, 2016: 2147-2156.

[36] He T, Soatto S. Mono3D++: Monocular 3D vehicle detection with two-scale 3D hypotheses and task priors[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33(1): 8409-8416.

[37] Xiang Y, Choi W, Lin Y Q, et al. Data-driven 3D voxel patterns for object category recognition[C]. 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, 2015: 1903-1911.

[38] Chabot F, Chaouch M, Rabarisoa J, et al. Deep MANTA: A coarse-to-fine many-task network for joint 2D and 3D vehicle analysis from monocular image[C]. 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, 2017: 1827-1836.

[39] Sun J M, Chen L H, Xie Y M, et al. Disp R-CNN: Stereo 3D object detection via shape prior guided instance disparity estimation[C]. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, 2020: 10545-10554.

[40] Liu Z D, Zhou D F, Lu F X, et al. AutoShape: real-time shape-aware monocular 3D object detection[C]. 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, 2021: 15621-15630.

[41] Mousavian A, Anguelov D, Flynn J, et al. 3D bounding box estimation using deep learning and geometry[C]. 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, 2017: 5632-5640.

[42] 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). Seoul, 2019: 9286-9295.

[43] Simonelli A, Bulò S R, Porzi L, et al. Disentangling monocular 3D object detection[C]. 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, 2019: 1991-1999.

[44] Luo S J, Dai H, Shao L, et al. M3DSSD: Monocular 3D single stage object detector[C]. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, 2021: 6141-6150.

[45] Li P L, Chen X Z, Shen S J. Stereo R-CNN based 3D object detection for autonomous driving[C]. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach,2019: 7636-7644.

[46] He K M, Gkioxari G, Dollár P, et al. Mask R-CNN[C]. 2017 IEEE International Conference on Computer Vision. Venice, 2017: 2980-2988.

[47] Qin Z Y, Wang J L, Lu Y. Triangulation learning network: From monocular to stereo 3D object detection[C]. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, 2019: 7607-7615.

[48] Li P X, Su S, Zhao H C. RTS3D: Real-time stereo 3D detection from 4D feature-consistency embedding space for autonomous driving[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 35(3): 1930-1939.

[49] Guo X Y, Shi S S, Wang X G, et al. LIGA-stereo: Learning LiDAR geometry aware representations for stereo-based 3D detector[C]. 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, 2021: 3133-3143.

[50] Kong T, Sun F C, Liu H P, et al. FoveaBox: Beyound anchor-based object detection[J]. IEEE Transactions on Image Processing, 2020, 29: 7389-7398.

[51] Law H, Deng J. CornerNet: Detecting objects as paired keypoints[J]. International Journal of Computer Vision, 2020, 128(3): 642-656.

[52] Tian Z, Shen C H, Chen H, et al. FCOS: Fully convolutional one-stage object detection[C]. 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, 2019: 9626-9635.

[53] Zhou X Y, Wang D Q, Krähenbühl P. Objects as points[J/OL]. 2019, arXiv: 1904.07850.

[54] Li P X, Zhao H C, Liu P F, et al. RTM3D: Real-time monocular 3D detection from object keypoints for autonomous driving[C]. Computer Vision — ECCV 2020. Cham: Springer International Publishing, 2020: 644-660.

[55] Liu Z C, Wu Z Z, Tóth R. SMOKE: Single-stage monocular 3D object detection via keypoint estimation[C]. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Seattle, 2020: 4289-4298.

[56] 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). Nashville, 2021: 4719-4728.

[57] 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.

[58] Lu Y, Ma X Z, Yang L, et al. Geometry uncertainty projection network for monocular 3D object detection[C]. 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, 2021: 3091-3101.

[59] Wang L, Zhang L, Zhu Y, et al. Progressive coordinate transforms for monocular 3D object detection[J/OL]. 2021, arXiv: 2108.05793.

[60] Chen Y J, Tai L, Sun K, et al. MonoPair: Monocular 3D object detection using pairwise spatial relationships[C]. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, 2020: 12090-12099.

[61] Chen H S, Huang Y Y, Tian W, et al. MonoRUn: Monocular 3D object detection by reconstruction and uncertainty propagation[C]. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, 2021: 10374-10383.

[62] Zhao C Q, Sun Q Y, Zhang C Z, et al. Monocular depth estimation based on deep learning: An overview[J]. Science China Technological Sciences, 2020, 63(9): 1612-1627.

[63] 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. Salt Lake City, 2018: 2345-2353.

[64] Qin Z Y, Wang J L, Lu Y. MonoGRNet: A geometric reasoning network for monocular 3D object

localization[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33(1): 8851-8858.

[65] Ku J, Pon A D, Waslander S L. Monocular 3D object detection leveraging accurate proposals and shape reconstruction[C]. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, 2019: 11859-11868.

[66] 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). Nashville, 2021: 3288-3297.

[67] Peng W L, Pan H, Liu H, et al. IDA-3D: Instance-depth-aware 3D object detection from stereo vision for autonomous driving[C]. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, 2020: 13012-13021.

[68] Chen Y L, Liu S, Shen X Y, et al. DSGN: Deep stereo geometry network for 3D object detection[C]. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, 2020: 12533-12542.

[69] Ding M Y, Huo Y Q, Yi H W, et al. Learning depth-guided convolutions for monocular 3D object detection[C]. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, 2020: 11669-11678.

[70] 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). Nashville, 2021: 8551-8560.

[71] Lian Q, Li P L, Chen X Z. MonoJSG: Joint semantic and geometric cost volume for monocular 3D object detection[C]. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, 2022: 1060-1069.

[72] Payen de la Garanderie G, Atapour Abarghouei A, Breckon T P. Eliminating the blind spot: Adapting 3D object detection and monocular depth estimation to 360 panoramic imagery[C]. Computer Vision — ECCV 2018. Munich, 2018: 789-807.

[73] 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). Long Beach, 2019: 8437-8445.

[74] Fu H, Gong M M, Wang C H, et al. Deep ordinal regression network for monocular depth estimation[C]. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, 2018: 2002-2011.

[75] Chang J R, Chen Y S. Pyramid stereo matching network[C]. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, 2018: 5410-5418.

[76] Mayer N, Ilg E, Häusser P, et al. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation[C]. 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, 2016: 4040-4048.

[77] Yamaguchi K, McAllester D, Urtasun R. Efficient joint segmentation, occlusion labeling, stereo and flow estimation[C]. Computer Vision — ECCV 2014. Zurich 2014: 756-771.

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

[79] You Y, Wang Y, Chao W L, et al. Pseudo-lidar++: Accurate depth for 3D object detection in autonomous driving[J/OL]. 2019, arXiv: 1906.06310.

[80] Ma X Z, Wang Z H, Li H J, et al. Accurate monocular 3D object detection via color-embedded 3D reconstruction for autonomous driving[C]. 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, 2019: 6850-6859.

[81] Pon A D, Ku J, Li C Y, et al. Object-centric stereo matching for 3D object detection[C]. 2020 IEEE International Conference on Robotics and Automation. Paris, 2020: 8383-8389.

[82] Li C Y, Ku J, Waslander S L. Confidence guided stereo 3D object detection with split depth estimation[C]. 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Las Vegas, 2020: 5776-5783.

[83] Li B, Zhang T L, Xia T. Vehicle detection from 3D lidar using fully convolutional network[J/OL]. 2016, arXiv: 1608.07916.

[84] Yang B, Luo W J, Urtasun R. PIXOR: Real-time 3D object detection from point clouds[C]. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, 2018: 7652-7660.

[85] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2015: 3431-3440.

[86] Yang B, Liang M, Urtasun R. Hdnet: Exploiting hd maps for 3D object detection[C]. Conference on Robot Learning. Zurich, 2018: 146-155.

[87] Beltrán J, Guindel C, Moreno F M, et al. BirdNet: A 3D object detection framework from LiDAR information[C]. The 21st International Conference on Intelligent Transportation Systems (ITSC). Maui, 2018: 3517-3523.

[88] Simon M, Amende K, Kraus A, et al. Complexer-YOLO: Real-time 3D object detection and tracking on semantic point clouds[C]. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Long Beach, 2019: 1190-1199.

[89] Meyer G P, Laddha A, Kee E, et al. LaserNet: An efficient probabilistic 3D object detector for autonomous driving[C]. 2019 IEEE/CVF Conference on Computer

Vision and Pattern Recognition (CVPR). Long Beach, 2019: 12669-12678.

[90] Bewley A, Sun P, Mensink T, et al. Range conditioned dilated convolutions for scale invariant 3D object detection[J/OL]. 2020, arXiv: 2005.09927.

[91] Li B. 3D fully convolutional network for vehicle detection in point cloud[C]. 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Vancouver, 2017: 1513-1518.

[92] Engelcke M, Rao D, Wang D Z, et al. Vote3Deep: Fast object detection in 3D point clouds using efficient convolutional neural networks[C]. 2017 IEEE International Conference on Robotics and Automation. Singapore, 2017: 1355-1361.

[93] 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. Salt Lake City, 2018: 4490-4499.

[94] Ye M S, Xu S J, Cao T Y. HVNet: Hybrid voxel network for LiDAR based 3D object detection[C]. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, 2020: 1628-1637.

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

[96] Ye Y Y, Chen H J, Zhang C, et al. SARPNET: Shape attention regional proposal network for liDAR-based 3D object detection[J]. Neurocomputing, 2020, 379: 53-63.

[97] Mao J G, Xue Y J, Niu M Z, et al. Voxel transformer for 3D object detection[C]. 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, 2021: 3144-3153.

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

[99] Yin T W, Zhou X Y, Krähenbühl P. Center-based 3D object detection and tracking[C]. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, 2021: 11779-11788.

[100] Chen Q, Sun L, Wang Z X, et al. Object as hotspots: An anchor-free 3D object detection approach via firing of hotspots[C]. Computer Vision — ECCV 2020. Cham: Springer International Publishing, 2020: 68-84.

[101] 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. Honolulu, 2017: 77-85.

[102] Qi C R, Yi L, Su H, et al.Pointnet++: Deep hierarchical feature learning on point sets in a metric space[J/OL]. 2017, arXiv: 1706.02413.

[103] 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). Seoul, 2019: 1951-1960.

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

[105] Xu Q G, Zhong Y Q, Neumann U. Behind the curtain: Learning occluded shapes for 3D object detection[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2022, 36(3): 2893-2901.

[106] Xu Q G, Zhou Y, Wang W Y, et al. SPG: Unsupervised domain adaptation for 3D object detection via semantic point generation[C]. 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, 2021: 15426-15436.

[107] 徐冰冰, 岑科廷, 黄俊杰, 等. 图卷积神经网络综述[J]. 计算机学报, 2020, 43(5): 755-780.
(Xu B B, Cen K T, Huang J J, et al. A survey on graph convolutional neural network[J]. Chinese Journal of Computers, 2020, 43(5): 755-780.)

[108] Wu Z H, Pan S R, Chen F W, et al. A comprehensive survey on graph neural networks[J]. IEEE Transactions on Neural Networks and Learning Systems, 2021, 32(1): 4-24.

[109] 孔玮, 刘云, 李辉, 等. 基于图卷积网络的行为识别方法综述[J]. 控制与决策, 2021, 36(7): 1537-1546.
(Kong W, Liu Y, Li H, et al. A survey of action recognition methods based on graph convolutional network[J]. Control and Decision, 2021, 36(7): 1537-1546.)

[110] Zarzar J, Giancola S, Ghanem B. PointRGCN: Graph convolution networks for 3D vehicles detection refinement[J/OL]. 2019, arXiv: 1911.12236.

[111] Shi W J, Rajkumar R. Point-GNN: Graph neural network for 3D object detection in a point cloud[C]. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, 2020: 1708-1716.

[112] Ren Z Z, Misra I, Schwing A G, et al. 3D spatial recognition without spatially labeled 3D[C]. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, 2021: 13199-13208.

[113] Zhao N, Chua T S, Lee G H. SESS: Self-ensembling semi-supervised 3D object detection[C]. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, 2020: 11076-11084.

[114] Qin Z Y, Wang J L, Lu Y. Weakly supervised 3D object detection from point clouds[C]. Proceedings of the 28th ACM International Conference on Multimedia. Seattle, 2020: 4144-4152.

[115] Yin J B, Shen J B, Guan C Y, et al. LiDAR-based online 3D video object detection with graph-based message passing and spatiotemporal transformer attention[C]. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, 2020:

11492-11501.

[116] Shi S S, Guo C X, Jiang L, et al. PV-RCNN1: Point-voxel feature set abstraction for 3D object detection[C]. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, 2020: 10526-10535.

[117] Bhattacharyya P, Czarnecki K. Deformable PV-RCNN: Improving 3D object detection with learned deformations[J/OL]. 2020, arXiv: 2008.08766.

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

[119] Noh J, Lee S, Ham B. HVPR: Hybrid voxel-point representation for single-stage 3D object detection[C]. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, 2021: 14600-14609.

[120] Chen Y L, Liu S, Shen X Y, et al. Fast point R-CNN[C]. 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, 2019: 9774-9783.

[121] Wang Z X, Jia K. Frustum ConvNet: Sliding Frustums to aggregate local point-wise features for amodal 3D object detection[C]. 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Macao, 2019: 1742-1749.

[122] Yang Z T, Sun Y N, Liu S, et al. IPOD: Intensive point-based object detector for point cloud[J/OL]. 2018, arXiv: 1812.05276.

[123] Vora S, Lang A H, Helou B, et al. PointPainting: Sequential fusion for 3D object detection[C]. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, 2020: 4603-4611.

[124] Yin T W, Zhou X Y, Krähenbühl P. Multimodal virtual point 3D detection[J/OL]. 2021, arXiv: 2111.06881.

[125] Sindagi V A, Zhou Y, Tuzel O. MVX-net: Multimodal VoxelNet for 3D object detection[C]. 2019 International Conference on Robotics and Automation (ICRA). Montreal, 2019: 7276-7282.

[126] Raffiee A H, Irshad H. Class-specific anchoring proposal for 3D object recognition in LIDAR and RGB images[J/OL]. 2019, arXiv: 1907.09081.

[127] Zhu M, Ma C, Ji P, et al. Cross-modality 3D object detection[C]. 2021 IEEE Winter Conference on Applications of Computer Vision. Waikoloa, 2021: 3771-3780.

[128] 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.

[129] Wu X P, Peng L, Yang H H, et al. Sparse fuse dense: Towards high quality 3D detection with depth completion[C]. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, 2022: 5408-5417.

[130] Xu D F, Anguelov D, Jain A. PointFusion: Deep sensor fusion for 3D bounding box estimation[C]. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, 2018: 244-253.

[131] Yoo J H, Kim Y, Kim J, et al. 3D-CVF: Generating joint camera and LiDAR features using cross-view spatial feature fusion for 3D object detection[C].Computer Vision — ECCV 2020. Cham: Springer International Publishing, 2020: 720-736.

[132] Li Y, Yu A W, Meng T, et al. DeepFusion: Lidar-camera deep fusion for multi-modal 3D object detection[J/OL]. 2022, arXiv: 2203.08195,

[133] Liang M, Yang B, Wang S L, et al. Deep continuous fusion for multi-sensor 3D object detection[C]. Computer Vision — ECCV 2018. Cham: Springer International Publishing, 2018: 663-678.

[134] Wang S L, Suo S, Ma W C, et al. Deep parametric continuous convolutional neural networks[C]. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, 2018: 2589-2597.

[135] Liang M, Yang B, Chen Y, et al. Multi-task multi-sensor fusion for 3D object detection[C]. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, 2019: 7337-7345.

[136] Piergiovanni A, Casser V, Ryoo M S, et al. 4D-net for learned multi-modal alignment[C]. 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, 2021: 15415-15425.

[137] Huang T T, Liu Z, Chen X W, et al. EPNet: Enhancing point features with image semantics for 3D object detection[C]. Computer Vision — ECCV 2020. Cham: Springer International Publishing, 2020: 35-52.

[138] Everingham M, Gool L, Williams C K I, et al. The pascal visual object classes (VOC) challenge[J]. International Journal of Computer Vision, 2010, 88(2): 303-338.

[139] Tian Z, Chu X, Wang X, et al. Fully convolutional one-stage 3D object detection on LiDAR range images[J/OL]. 2022, arXiv: 2205.13764.

[140] Park D, Ambru R, Guizilini V, et al. Is pseudo-lidar needed for monocular 3D object detection?[C]. 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, 2021: 3122-3132.

[141] Qiao S Y, Chen L C, Yuille A. DetectoRS: Detecting

objects with recursive feature pyramid and switchable atrous convolution[C]. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, 2021: 10208-10219.

[142] Lim J S, Astrid M, Yoon H J, et al. Small object detection using context and attention[C]. 2021 International Conference on Artificial Intelligence in Information and Communication (ICAIIC). Jeju Island, 2021: 181-186.

[143] Paigwar A, Erkent O, Wolf C, et al. Attentional PointNet for 3D-object detection in point clouds[C]. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Long Beach, 2019: 1297-1306.

[144] Gao F, Wang C M. Hybrid strategy for traffic light detection by combining classical and self-learning detectors[J]. IET Intelligent Transport Systems, 2020, 14(7): 735-741.

[145] Li Y, Bu R, Sun M, et al. Pointcnn: Convolution on x-transformed points[J]. Advances in Neural Information Processing Systems, 2018, 31: 820-830.

[146] Yan X, Zheng C D, Li Z, et al. PointASNL: Robust point clouds processing using nonlocal neural networks with adaptive sampling[C]. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, 2020: 5588-5597.

[147] He Q D, Wang Z N, Zeng H, et al. SVGA-net: Sparse voxel-graph attention network for 3D object detection from point clouds[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2022, 36(1): 870-878.

[148] 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). Seoul, 2019: 924-932.

[149] 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. Waikoloa, 2021: 1526-1535.

[150] Sheeny M, de Pellegrin E, Mukherjee S, et al. RADIATE: A radar dataset for automotive perception in bad weather[C]. 2021 IEEE International Conference on Robotics and Automation. Xi'an, 2021: 1-7.

[151] Zhu B J, Jiang Z K, Zhou X X, et al. Class-balanced grouping and sampling for point cloud 3D object detection[J/OL]. 2019, arXiv: 1908.09492.

[152] Zhang W, Wang Z, Change Loy C. Multi-modality cut and paste for 3D object detection[J/OL]. 2020, arXiv: 2012.12741.

[153] Meng Q H, Wang W G, Zhou T F, et al. Weakly supervised 3D object detection from lidar point cloud[C]. Computer Vision — ECCV 2020. Cham: Springer International Publishing, 2020: 515-531.

[154] Li P X, Zhao H C. Monocular 3D detection with geometric constraint embedding and semi-supervised training[J]. IEEE Robotics and Automation Letters, 2021, 6(3): 5565-5572.

[155] Beker D, Kato H, Morariu M A, et al. Monocular differentiable rendering for self-supervised 3D object detection[C]. Computer Vision — ECCV 2020. Cham: Springer International Publishing, 2020: 514-529.

[156] Peng L, Yan S, Wu B, et al. WeakM3D: Towards weakly supervised monocular 3D object detection[J/OL]. 2022, arXiv: 2203.08332.

[157] Brazil G, Pons-Moll G, Liu X M, et al. Kinematic 3D object detection in Monocular video[C]. Computer Vision — ECCV 2020. Cham: Springer International Publishing, 2020: 135-152.

[158] Guizilini V, Ambru R, Pillai S, et al. 3D packing for self-supervised monocular depth estimation[C]. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, 2020: 2482-2491.

作者简介

任柯燕 (1982–), 女, 副教授, 博士, 从事计算机视觉、机器学习理论、方法和关键技术等研究, E-mail: keyanren@bjut.edu.cn;

谷美颖 (1998–), 女, 硕士生, 从事图像合成、目标检测等研究, E-mail: g1224598163@163.com;

袁正谦 (1998–), 男, 硕士生, 从事6D目标检测、机器人抓取等研究, E-mail: 824131408@qq.com;

袁帅 (1998–), 男, 硕士生, 从事轨迹预测、计算机视觉等研究, E-mail: yua_shuai@126.com.

(责任编辑: 李君玲)