

中图法分类号: TP391.7 文献标识码: A 文章编号: 1006-8961(2024)06-1765-20

论文引用格式: Jia M D, Yang J M, Meng W L, Guo J W, Zhang J G and Zhang X P. 2024. Survey on the fusion of point clouds and images for environmental object detection. Journal of Image and Graphics, 29(06):1765-1784(贾明达, 杨金明, 孟维亮, 郭建伟, 张吉光, 张晓鹏. 2024. 融合点云与图像的环境目标检测研究进展. 中国图象图形学报, 29(06):1765-1784)[DOI:10.11834/jig.240030]

融合点云与图像的环境目标检测研究进展

贾明达, 杨金明, 孟维亮*, 郭建伟, 张吉光, 张晓鹏

1. 中国科学院自动化研究所多模态人工智能系统全国重点实验室, 北京 100190;

2. 中国科学院大学人工智能学院, 北京 100049

摘要: 在数字仿真技术应用领域, 特别是在自动驾驶技术的发展中, 目标检测是至关重要的一个环节, 它涉及对周围环境中物体的感知, 为智能装备的决策和规划提供了关键信息。近年来, 随着传感器技术的进步, 图像和点云成为两种主要的感知数据源, 它们各自在基于深度学习技术的目标检测方法研究中具有独特的优势。为了更加全面地对现有基于点云和图像的目标检测方法进行研究, 本文对基于图像、点云及两者联合的3类目标检测算法进行系统的梳理和总结, 旨在探索如何将这两种数据源融合起来, 促进提高目标检测的准确性、稳定性和鲁棒性, 并对融合点云和图像的环境目标检测发展方向进行展望。

关键词: 点云; 自动驾驶; 多模态; 目标检测; 融合

Survey on the fusion of point clouds and images for environmental object detection

Jia Mingda, Yang Jinming, Meng Weiliang*, Guo Jianwei, Zhang Jiguang, Zhang Xiaopeng

1. State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences,

Beijing 100190, China; 2. School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China

Abstract: In the field of digital simulation technology applications, especially in the development of autonomous driving, object detection is a crucial component. It involves the perception of objects in the surrounding environment, which provides essential information for the decision-making process and planning of intelligent systems. Traditional object detection methods typically involve steps such as feature extraction, object classification, and position regression on images. However, these methods are limited by manually designed features and the performance of classifiers, which restrict their effectiveness in complex scenes and for objects with significant variations. The advent of deep learning technology has led to the widespread adoption of object detection methods based on deep neural networks. Notably, the convolutional neural network (CNN) has emerged as one of the most prominent approaches in this field. By leveraging multiple layers of convolution and pooling operations, CNNs are capable of automatically extracting meaningful feature representations from image data. In addition to image data, light detection and ranging (LiDAR) data play a crucial role in object detection tasks, particularly

收稿日期: 2024-01-11; 修回日期: 2024-02-11; 预印本日期: 2024-02-18

* 通信作者: 孟维亮 weiliang.meng@ia.ac.cn

基金项目: 北京市自然科学基金—丰台轨道交通前沿研究联合基金(L231013); 国家自然科学基金(U21A20515, 62376271, 62172416, 52175493, U22B2034, 62365014); 北京航空航天大学虚拟实现国家重点实验室开放课题(VRLAB2023B01)

Supported by: Beijing Natural Science Foundation (L231013); National Natural Science Foundation of China (U21A20515, 62376271, 62172416, 52175493, U22B2034, 62365014); the Open Project Program of State Key Laboratory of Virtual Reality Technology and Systems, Beihang University (VRLAB2023B01)

for 3D object detection. LiDAR data represent objects through a set of unordered and discrete points on their surfaces. Accurately detecting point cloud clusters representing objects and providing their pose estimation from these unordered points is a challenging task. LiDAR data, with their unique characteristics, offer high-precision obstacle detection and distance measurement, which contributes to the perception of surrounding roadways, vehicles, and pedestrian targets. In real-world autonomous driving and related environmental perception scenarios, using a single modality often presents numerous challenges. For instance, while image data can provide a wide variety of high-resolution visual information such as color, texture, and shape, it is susceptible to lighting conditions. In addition, models may struggle to handle occlusions caused by objects obstructing the view due to inherent limitations in camera perspectives. Fortunately, LiDAR exhibits exceptional performance in challenging lighting conditions and excels at accurately spatially locating objects in diverse and harsh weather scenarios. However, it possesses certain limitations. Specifically, the low resolution of LiDAR input data results in sparse point cloud when detecting distant targets. Extracting semantic information from LiDAR data is also more challenging than that from image data. Thus, an increasing number of researchers are emphasizing multimodal environmental object detection. A robust multimodal perception algorithm can offer richer feature information, enhanced adaptability to diverse environments, and improved detection accuracy. Such capabilities empower the perception system to deliver reliable results across various environmental conditions. Certainly, multimodal object detection algorithms also face certain limitations and pressing challenges that require immediate attention. One challenge is the difficulty in data annotation. Annotating point cloud and image data is relatively complex and time consuming, particularly for large-scale datasets. Moreover, accurately labeling point cloud data is challenging due to their sparsity and the presence of noisy points. Addressing these issues is crucial for further advancements in multimodal object detection. Moreover, the data structure and feature representation of point cloud and image data, as two distinct perception modalities, differ significantly. The current research focus lies in effectively integrating the information from the two modalities and extracting accurate and comprehensive features that can be utilized effectively. Furthermore, processing large-scale point cloud data are equally challenging. Point cloud data typically encompass a substantial number of 3D coordinates, which necessitates greater demands on computing resources and algorithmic efficiency compared with pure image data. This study aims to summarize and refine existing approaches to facilitate researchers in gaining a deeper and more efficient understanding of object detection algorithms that integrate images and point clouds. It classifies object detection algorithms based on multimodal fusion of point clouds, images, and combinations of both. Furthermore, we analyze the strengths and weaknesses of various methods while discussing potential solutions. Moreover, we provide a comprehensive review of the development of object detection algorithms that fuse point clouds and images, with considerations of aspects such as data collection, representation, and model design. Ultimately, we give a perspective on the future development direction of environmental target detection, and the goal is to enhance overall capabilities in autonomous systems.

Key words: point cloud; autonomous driving; multimodal; object detection; fusion

0 引言

目标检测是计算机视觉领域的一项重要任务,旨在从图像或视频中准确地识别和定位特定目标。它在许多应用中发挥着关键作用,如自动驾驶、视频监控和人脸识别等。目标检测算法的发展经历了多个阶段,逐渐从传统的基于手工特征的方法发展到基于深度学习的方法(Arora等,2007;Aldoma等,2012)。

传统的目标检测方法通常涉及针对图像进行特征提取、目标分类和位置回归等步骤。其中常用的特征包括 Harr 小波特征算子(Viola 和 Jones, 2001)、

HOG(histogram of gradients)(Dalal 和 Triggs, 2005)和 SIFT (scale-invariant feature transform) (Ng 和 Henikoff, 2003)。这些方法在一定程度上能够实现目标检测,但受限于手工设计的特征和分类器的性能,对于复杂场景和变化较大的目标具有局限性。随着深度学习技术的兴起,基于深度神经网络的目标检测方法逐渐成为主流。其中最具有代表性的是基于卷积神经网络(convolutional neural network, CNN)的方法。CNN 通过多层卷积和池化操作,能够自动学习图像中的特征表示。在目标检测中,基于 CNN 的方法通常可以按照两个主要的分类方式进行划分:单阶段(one-stage)方法和两阶段(two-stage)方

法。单阶段方法直接在图像上密集预测目标的类别和位置,通常速度较快,代表性的算法包括 YOLO (you only look once) 系列 (Bochkovskiy 等, 2020) 和 SSD (single shot multibox detector) 系列 (Liu 等, 2016) 等。两阶段方法则首先生成候选框,然后再对候选框进行分类和位置回归,代表性的算法包括基于选择性搜索的 R-CNN (region convolutional neural network) (Girshick 等, 2014)、Fast R-CNN 系列 (Girshick, 2015) 和基于候选框生成的 Faster R-CNN 系列 (Ren 等, 2017)。由于两阶段方法需要进行两个步骤,会影响模型最终的计算速度。随着硬件性能的不断提升,目前基于单阶段的环境目标检测方法已经成为主流计算方法。

与此同时,随着激光雷达 (light detection and ranging, LiDAR) 技术的不断发展和成本的下降, LiDAR 数据也逐渐引入到目标检测领域。在 LiDAR 数据中,任何物体都以其表面的无序离散点表示,从这些无序离散点中,准确检测出代表目标 (如行人、车辆) 的点云块并给出目标位姿是一个具有挑战性的任务。由于 LiDAR 数据的独特性,可以提供高精度的障碍物检测和距离测量,帮助车辆感知周围的道路、车辆和行人目标。通过与其他传感器相结合, LiDAR 可以实现多模态感知,从而显著提高目标检测的准确性和鲁棒性。

近年来,研究人员提出众多针对点云与图像的 3D 目标检测方法 (Arnold 等, 2019; Guo 等, 2021), 现有主流的 3D 目标检测方法按照输入数据模态的不同,可划分为以下 3 种:

1) 基于图像的 3D 目标检测。由于 2D 图像本身缺少深度信息,直接使用 2D 图像进行 3D 目标检测往往十分困难。主流的基于图像的 3D 目标检测方式可以分为基于纯图像的 3D 目标检测、基于深度估计的 3D 目标检测、基于先验引导的 3D 目标检测和基于多视图的 3D 目标检测。尽管这些方法在数据使用或特征表示层面有所不同,但它们都从二维图像中获取目标的三维特征,以进行对目标三维位置、姿态和形状等信息的推断。

2) 基于点云的 3D 目标检测。此类方法通常以 LiDAR 点云 (Yan 等, 2018) 作为输入,相比图像数据, LiDAR 点云提供了更为准确和稠密的三维数据,能够更精确地描述目标的几何属性。这类方法可以分为两个大类。一类方法直接使用原始点云数据进

行目标检测,对输入的三维点云数据直接进行数据预处理、特征提取和分类检测等操作;另一类方法则将原始点云先转化为其他的数据表示,如体素 (voxel) (Catmull, 1998)、鸟瞰图 (bird's eye view, BEV) (Chen 等, 2017) 和距离图像 (range view) (Fan 等, 2021) 等视图,然后进行基于图像的目标检测。

3) 融合点云和图像的 3D 目标检测 (Xu 等, 2019)。此类算法通常使用多模态融合的方法实现 3D 目标检测,结合图像和激光雷达两种数据源的特点,可以获得更加丰富的融合信息,具有更好的稳定性和鲁棒性 (Feng 等, 2021)。

目前已有研究者对基于图像和 LiDAR 的 3D 目标检测研究进行了总结整理。已有工作包括对 3D 点云方法中的分类、追踪和分割等各项任务进行研究 (Guo 等, 2021), 也包括面向多模态自动驾驶中的目标检测和语义分割方法 (Feng 等, 2021), 以及自动驾驶应用技术的发展现况 (Arnold 等, 2019) 和多模态目标检测等新型融合手段 (Wang 等, 2023)。这些综述大多针对 3D 目标检测的某一具体领域,并非注重不同方法之间的联系,且随着目标检测技术的迅速发展,部分综述的内容已经不能完全适用现有的研究进展。

为了研究人员可以对融合图像和点云的目标检测算法进行更好更快的了解,本文主要对已有目标检测算法进行总结提炼,按照基于点云、图像、图像和点云多模态融合的方法对目标检测算法进行分类,对各类方法进行优缺点分析及探讨可能的解决方案。同时从数据采集及表示、模型设计等不同角度对融合点云与图像的目标检测算法的发展进行更全面的综述,并对环境目标检测的未来进行展望。

本文的主要贡献如下: 1) 总结概括了当前环境目标检测中的常用数据采集方式、数据集、数据形式和发展现状; 2) 按照基于点云、图像、图像和点云多模态融合的方法对目标检测算法进行分类,对各类方法进行详细的介绍,对各类方法之间的联系做了分析及梳理; 3) 提出环境目标检测研究中的现存问题和未来展望。

1 环境目标检测研究现状

1.1 图像及点云数据的采集

图像数据使用通用相机即可进行采集,但在三

维位置信息的计算中,常受到精度方面的限制。激光雷达通过主动发射激光束,测量光线打到物体表面再反射回来所需要的时间,来计算激光雷达到目标点的距离,这个过程中会获取数百万计的数据点,构建出其正在测量的空间表面的复杂“地图”,称为“点云”。点云数据以离散无序的三维点的形式呈现,每个点都包含了其在空间中的准确坐标。这使得点云数据能够直接捕捉到物体的三维几何特征,而不受视角变化的影响。相比于图像数据,点云数据具有独特的优势,可以提供更加精确的深度信息,包括物体的位置、形状和大小,从而有效减轻二维图像中常见的遮挡问题(Cui等,2022)。

1.2 常用的3D目标检测数据集

在现有研究中,研究人员常常依赖于各种广泛使用的公开数据集来评估和验证他们的算法与自动驾驶系统。常用的3D目标检测数据集主要包括KITTI数据集、nuScenes数据集和Waymo数据集。

KITTI数据集(Geiger等,2012)是一个广泛应用于自动驾驶领域,由德国卡尔斯鲁厄理工学院KIT(Karlsruhe Institute of Technology)和美国芝加哥丰田技术研究院TTI-C(Toyota Technological Institute at Chicago)联合赞助的用于自动驾驶领域研究的数据集。KITTI数据集使用64线的Velodyne激光雷达完成点云数据采集,并使用通用相机进行对应图像的采集,其中包含车辆、行人和自行车等常见的道路上的物体,具有较高的标注精度和相对明确的场景范围,可用于多种自动驾驶场景的视觉算法评估,如光流、深度估计、2D/3D目标检测和目标跟踪等任务。

nuScenes(nuTonomy scenes)数据集(Caesar等,2020)是Motional(前身为nuTonomy)团队开发的用于自动驾驶的公共大规模数据集,由在美国波士顿和新加坡拍摄的不同驾驶场景组成,也是自动驾驶领域使用的另一个广泛的公开数据集之一,是目前最权威的自动驾驶纯视觉3D目标检测评测集。nuScenes数据集注重真实世界中的多样性和复杂性,涵盖不同地点、不同时间和不同天气情况的驾驶环境,每个环境的时长为20 s,其数据标注量是KITTI的7倍,且是首个包含全传感器套件的数据集。此外,KITTI数据集中大多是白天和天气条件较好的场景,而nuScenes数据集的场景包含夜晚及雨天,更具有挑战性。

Waymo数据集(Sun等,2020)是由谷歌旗下的

Waymo公司提供的大规模3D感知数据集。Waymo包含798个训练场景、202个验证场景和150个测试场景,每个场景的时长为20 s。Waymo数据集的标注频率是nuScenes的6倍,场景数量是nuScenes数据集的3倍,有数量更多且更密集的2D/3D标签。Waymo数据集对自动驾驶算法的鲁棒性和泛化能力提出了更高的要求,是目前为止最大、最多样化的数据集。

1.3 目标检测中常见的数据表示形式

在现有研究中,环境目标检测的数据主要表示形式可以划分为图像数据、原始点云、体素、前视图、鸟瞰图和柱体。

1.3.1 图像数据

常用的视觉传感器主要包括单目相机和深度相机。单目相机通常用于获得该视角下的二维图像,即RGB图像与灰度图像,可以提供丰富的纹理信息和色彩,且这类图像的处理速度相比点云数据要更快,占用的计算资源要更少。深度相机则用于获得图像中的每一个点到相机的距离,即每一个点的三维空间坐标(Xu等,2019)。图像数据一直以来都是环境目标检测研究中十分重要的一种输入数据。

1.3.2 原始点云

原始的三维点云可使用激光雷达、深度相机或双目相机获得,能够直接作为目标检测模型的输入数据,通过深度学习网络进行深层特征提取,从而完成目标检测或语义分割任务。典型的对原始点云数据直接处理的工作有PointNet(Charles等,2017)、PointNet++(Qi等,2017)等工作。原始点云数据在笛卡尔坐标系中使用 (x, y, z) 进行编码,其视觉显示效果如图1所示,一般附带每个点的反射强度等相关信息。直接使用原始点云进行目标检测的计算成本相对更高,设计模型时需要考虑点云数据本身的无序性和稀疏性。

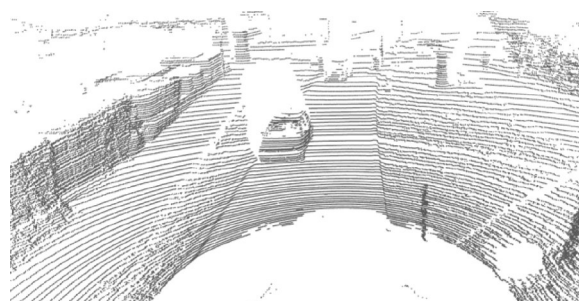


图1 来自激光雷达扫描的原始点云数据(Wang等,2023)

Fig. 1 Raw point cloud data (Wang et al., 2023)

1.3.3 体素

体素是一种有效表示原始点云的数据处理方式,点云的体素化是指将连续的三维点云数据划分为离散的三维网格空间的过程。每个体素网格代表一个固定大小的三维空间,如图2所示,类似图像中的像素(Zhou和Tuzel,2018)。一般这个过程需要选择合适的体素分辨率,用于确定每个体素网格的大小。更小的体素网格代表着更大的计算开销和更高的计算精度。对点云进行体素化是一种广泛使用于环境目标检测研究的数据预处理方式。

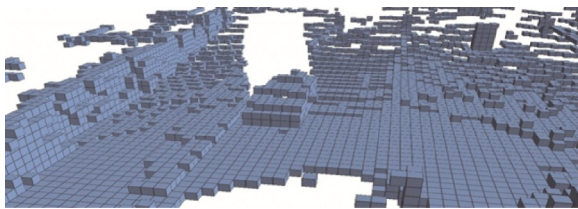


图2 体素数据(Wang等,2023)
Fig. 2 Voxel data(Wang et al.,2023)

1.3.4 前视图

前视图(front view,FV)是指从一个特定视角对场景或对象进行观察和捕捉,并将观察到的信息以二维图像形式的数据表示。前视图数据通常通过摄像头、相机或其他视觉传感器进行捕获,也可以通过将三维点云数据投影到二维平面上获取(Ku等,2018),如图3所示。通过3D点云数据获取前视图的过程可以选定不同的投影方式:正交投影和透视投影。正交投影将点云直接投影到平行于视平面的平面上,透视投影则模拟人眼观察物体时的透视视角完成投影。不同的前视图投影方式往往对应着不同的平移、缩放和旋转等坐标变换操作。

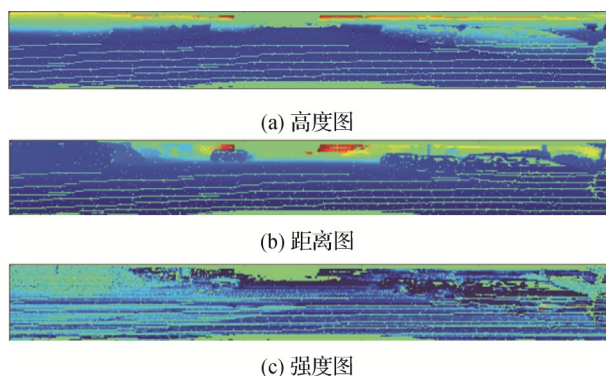


图3 前视图编码方式(Chen等,2017)
Fig. 3 Front view coding(Chen et al.,2017)
((a)height;(b)distance;(c)intensity)

1.3.5 鸟瞰图

鸟瞰图(BEV)与前视图类似,属于点云数据在某一视角(即俯视角)下的投影,鸟瞰图也可以使用多种编码方式对点云进行处理,如图4所示。

在鸟瞰图表示下,前后遮挡的物体在垂直视角无相互遮掩,且每个物体都保留长度和宽度信息。使用鸟瞰图转化而来的特征便于后续模型在垂直方向上的检测和定位(Chen等,2017)。

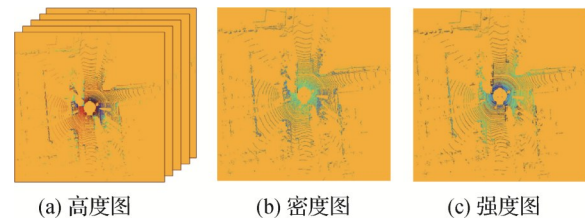


图4 鸟瞰图编码方式(Chen等,2017)
Fig. 4 Bird's eye view coding(Chen et al.,2017)
((a) height;(b) density;(c) intensity)

1.3.6 柱体

柱体表示(Vora等,2020)是一种将点云数据转为柱体空间的表示方法。在柱体表示中,点云中的点被分配到一个柱形空间,类似于体素的划分网格。每个柱体以一定步长对点云在 x 和 y 方向上进行区域划分,而不对 z 方向上进行划分。

1.4 国内外环境目标检测研究现状

近年来自动驾驶领域迎来快速发展,从自动驾驶的环境感知算法层面来看,目前主流技术路线可以分为两种:纯视觉技术路线和使用激光雷达的技术路线。纯视觉路线仅使用相机图像对周围环境进行目标检测感知,代表公司有特斯拉和百度 Apollo Lite等。使用激光雷达进行三维环境感知仍是目前大部分企业的首选,如华为、Uber、Waymo等都属于此类算法的研究团队。

就基于点云的环境目标检测算法而言,近年来国内已有许多单位对此开展研究工作,同时也有多家单位具备点云数据获取的扫描设备并取得了相应成果,如北京大学提出基于BEV特征空间的视觉与LiDAR高性能融合框架(Liu等,2023)、腾讯优图提出基于点云体素化的高效多阶段目标检测(Chen等,2019)以及基于点的点云稀疏到稠密多阶段目标检测模型(Yang等,2019)、商汤科技提出利用激光雷达线索进行非监督辅助的目标检测方法(Tian等,2021)等工作,都在一定程度上推动了基于图像和点

云的环境目标检测研究。

1.5 融合点云和图像多模态目标检测的现存问题

1)数据标注困难。点云和图像数据的标注工作相对复杂和耗时,尤其是对于大规模数据集。同时,由于点云数据的稀疏性和存在噪声点的影响,对点云数据进行精确的标注也是一项挑战。

2)跨模态融合。点云和图像数据是两种不同的感知模态,其数据结构和特征表示方式存在差异。如何有效地融合这两种模态的信息,提取出准确且完备的可用特征,是目前的一个研究热点。

3)大规模点云数据的处理。点云数据通常包含大量的三维坐标,相比纯图像数据而言,处理大规模点云数据对计算资源和算法效率有着更高的要求。如何高效地处理和利用大规模点云数据,也是亟待解决的问题之一。

2 基于图像的目标检测

基于图像的目标检测工作分为2D目标检测和3D目标检测两个大类。其中3D目标检测很多的研究是在2D目标检测算法的基础上进行的。

基于图像的2D目标检测可分为两种:基于候选区域的方法和基于回归的方法。基于候选区域的目标检测方法首先生成一组候选区域来进行目标的初步定位,再使用神经网络及各种附加条件进一步确认感兴趣区域(region of interest, ROI)中物体的位置和类别等信息(曹家乐等,2022)。这类工作的代表性方法有R-CNN(Girshick等,2014)、Fast R-CNN(Girshick,2015)、Faster R-CNN(Ren等,2017)和Mask R-CNN(He等,2017)等。基于回归的目标检测方法则将目标的边界定位问题转化成端到端的回归问题,舍弃了区域(region)的思想,直接获取目标的边界描述,如SSD(Liu等,2016)、YOLOv4(Bochkovskiy等,2020)、RetinaNet(Lin等,2017)、EfficientDet(Tan等,2020)等。

然而,单纯使用2D的目标检测模型并不能为自动驾驶场景提供有效的物体定位,因此很多研究人员试图从2D目标检测方法出发,将其中的检测模型扩展为3D目标检测。本文侧重介绍3D目标检测的相关方法,并将其分为基于纯图像的单目3D检测、基于深度估计的单目3D检测、基于先验引导的单目3D检测和基于多视角图像的BEV目标检测。

2.1 基于纯图像的单目3D检测

单目3D目标检测最直接的方案是通过神经网络直接从图像中获得物体的3D检测框参数,即与2D目标检测类似,同样可以分为基于候选区域的目标检测方法和基于回归的目标检测方法。此外,这些方法也可以按照是否使用锚点(anchor)和是否为端到端方式来分类。

基于锚点的目标检测方法有基于3D锚点、2D锚点、深度锚点等不同的锚点设置,旨在将物体的位置通过度量的方法确定,用真实值与给出的锚点预测值做偏差纠正,不断优化模型的目标检测效果,如YOLO系列和RCNN系列都是使用锚点进行目标检测的经典工作,目标检测任务中的锚点如图5所示。

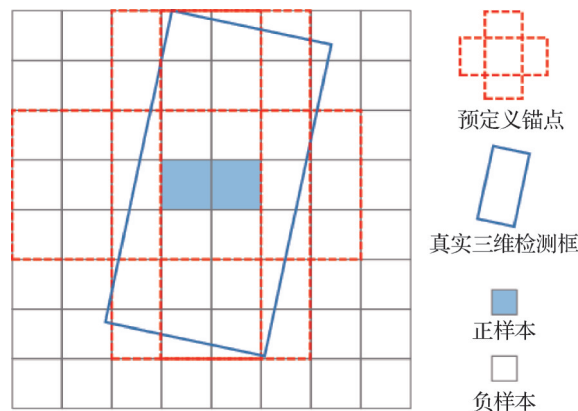


图5 目标检测任务中的锚点(Mao等,2022)

Fig. 5 Anchor of object detection(Mao et al.,2022)

无锚点的方法同样使用卷积神经网络对图像进行处理,区别在于无锚点方法直接对物体的相关参数进行预测,可看做网络自行学习锚点的位置与形状。无锚点预测算法有许多模型组件,包括分类头、中心点检测头、中心点偏置回归头、深度距离回归头和关键点检测头等。代表性的无锚点目标检测工作有:FCOS(fully convolutional one-stage)(Tian等,2019)、CornerNet(Law和Deng,2020)、CenterNet(Duan等,2019)、ExtremeNet(Zhou等,2019)等。

单阶段3D目标检测方法可以自然地借鉴2D目标检测网络架构的设计,以端到端的方式训练模型。其中基于锚点的单阶段3D目标检测方法有M3D-RPN(monocular 3D region proposal network)(Brazil和Liu,2019)、Kinematic3d(Brazil等,2020)、M3DSSD(monocular 3D single stage object detector)(Luo等,

2021)和FQNet(fitting quality network)(Liu等,2019)等。无锚点单阶段3D目标检测方法则有SMOKE(single-stage monocular 3D object detection via key-point estimation)(Liu等,2020)、Fcos3D(fully convolutional one-stage monocular 3D)(Wang等,2021d)、MonoEF(monocular extrinsic parameter free)(Zhou等,2021)、OFT(orthographic feature transform)(Roddick等,2019)和MoVi-3D(monocular virtual views 3D)(Simonelli等,2020)等。

多阶段3D目标检测方法通常需要一定的手工设计将传统的两阶段2D检测体系扩展到3D目标检测,即在第1阶段利用2D检测器从输入图像生成2D目标框,然后在第2阶段通过物体2D视角的ROI来预测3D视角下的各项参数,以形成完整的3D检测框,如Gs3d(guidance surface 3D)(Li等,2019)、MonoGRNet(monocular geometric reasoning network)(Qin等,2022)、MonoRCNN(monocular region convolutional neural networks)(Shi等,2021)、GUPNet(geometry uncertainty projection network)(Lu等,2021)。

总的来说,基于纯图像的单目3D目标检测方法通过改进的2D目标检测框架,从图像中直接回归3D检测框的参数,往往能直接受益于2D图像处理的相关研究,且其中大多数方法都可以进行端到端训练,不需要进行预训练或后处理。然而仅使用单目2D图像直接回归每个3D物体的深度是一项困难的任务。深度信息带来的误差容易成为阻碍3D目标检测任务的关键因素,该过程只有检测框级别的标注作为监督的真值,限制了模型的性能提升。

2.2 基于深度估计的单目3D检测

基于深度估计的单目3D检测是指利用深度学习从单幅图像中推断对应深度信息并估计物体位置和姿态的目标检测方法。深度估计是这类方法中重要的一环,许多工作采用预训练辅助深度估计网络的方法,以获得准确的单目检测结果。

具体来说,这类方法需要将单目2D图像送入预训练好的深度估计器以获得深度图像,如MonoDepth(monocular depth)(Godard等,2017)和DORN(deep ordinal regression network)(Fu等,2018),得到的深度图像可通过基于深度图像的方法和基于伪激光雷达的方法完成后续的3D目标识别和定位(晋帅等,2023)。

基于深度图像的方法将图像和深度映射与专门的神经网络融合,生成深度感知特征,可以提高检测性能,如MultiFusion(multilingual, multimodal diffusion)(Xu和Chen,2018)、MonoFENet(monocular feature enhancement networks)(Bao等,2020)、D4LCN(depth-guided dynamic-depthwise-dilated local convolutional network)(Ding等,2020a)和DDMP(depth-conditioned dynamic message propagation)(Wang等,2021c)。而基于伪激光雷达的方法将深度图像转换为伪激光雷达点云,并将伪激光雷达点云送入基于激光雷达的3D检测器,如Pseudo-LiDAR(Wang等,2019a)、AM3D(accurate monocular 3D)(Ma等,2019)和Deep Optics(Chang和Wetzstein,2019)。

2.3 基于先验引导的单目3D检测

研究人员也在不断探索图像中目标的形状和场景几何等先验知识,以建立3D目标检测框在世界坐标系下和像素坐标系下的对应关系,从而解决病态的单目3D目标检测。例如,通过引入预训练好的子网络来学习有关3D物体的先验知识,辅助2D目标检测获得3D目标检测的能力。其中可学习的先验知识包括:目标形状先验、类别先验、尺度先验、姿态先验等。代表性工作有Mono3d++(He和Soatto,2019)、ROI-10d(Manhardt等,2019)、MonoPSR(monocular proposals and shape reconstruction)(Ku等,2019)、Shift R-CNN(Naiden等,2019)、RTM3d(real-time monocular 3D)(Li等,2020)和Kinematic3d(Brazil等,2020)等。

2.4 基于多视角图像的BEV目标检测

使用鸟瞰图(BEV)表示进行3D目标检测的优势在于,可以有效减缓目标检测中的遮挡问题,并且保留了物体的高度和宽度信息。相比于在单目3D检测过程中使用的前视图,使用BEV进行物体的定位往往具有更小的垂直位置误差,更易获得准确的3D检测框。但是将点云转成BEV会丢失大量纵轴上的信息,对于行人、路标等物体来说,检测效果不佳。

相较点云直接转化成BEV特征进行感知,以多视角匹配的方式将多视角图像序列转化成BEV特征进行感知得到的语义信息更为丰富,同时能够得到较为准确的深度信息。常见的透视图到鸟瞰图方法可分为两大类。一类是基于集合投影将透视图转化为BEV的传统方法;另一类是以数据驱动的方式

隐式利用几何关系将多视角图像转化为 BEV 特征。近几年来基于数据驱动的透视图到鸟瞰图方法受到了较为广泛的关注,如 BEVFusion (Liu 等, 2023)、BEVFormer (Li 等, 2022)、BEVFormerV2 (Yang 等, 2022)和 M2BEV (Xie 等, 2022)等。

BEVFormer 提出一个采用纯视觉多视角图像的 3D 目标检测模型,通过 2D 卷积神经网络提取环视图像的多尺度特征,然后将其以模型学习的方式,通过时域和空域的注意力模块对图像特征编码生成 BEV 特征,最后接入特征解码器实现 3D 目标检测和地图分割任务。

BEVFormerV2 为克服 BEVFormer 检测器结构过于复杂导致编码器和解码器梯度流扭曲的问题,引入了透视 3D 检测头,即从透视视角产生的监督信号直接作用于骨干网络,指导骨干网络学习 2D 识别任务中丢失的 3D 信息,促进了模型的优化过程。同时模型也将上述透视检测头和 BEV 检测头合并为一个多阶段的 BEV 检测器,很大程度上提高了模型 BEV 识别的适应性和收敛速度。

随着 BEV 感知算法的不断发展,基于多视角图像的 BEV 目标检测已经成为目前主流自动驾驶系统的首选感知方案,在实时系统的高效计算、路径规划等任务上都有着亮眼的表现。

3 基于点云的目标检测

基于点云的目标检测算法是近年来计算机视觉领域的研究热点之一。早期研究阶段,研究人员主要探索基于传统的几何学和机器学习方法,依赖于手工设计的特征和规则对目标点云进行分割和分类等任务,具有很大的局限性。随着深度学习的兴起,在 PointNet 和 PointNet++ 等经典工作最早将点云数据应用于深度学习之后,研究人员开始将深度学习应用于点云目标检测中。目前基于点云的目标检测算法可以分为基于点的目标检测、基于体素的目标检测、基于 BEV 的目标检测和基于距离图像的目标检测。

3.1 基于点的目标检测

基于点的目标检测是指直接通过原始点云数据来预测 3D 目标的位置。这类方法通常包括以下几个模块:

1)数据预处理。包括移除离群点、点云滤波等

操作,以便消除点云中的噪声,从而提高点云的质量和位置的准确性。

2)特征学习器。通过深度学习的手段学习点云的特征表示,这些特征往往由局部特征和全局特征共同组成,以便更好地区分目标和背景。

3)目标检测器。即在特征提取模块之后给网络加上目标检测器。与 2D 目标检测器不同,点云目标检测器的输入特征往往更关注于点云的法线、周围点的分布和强度等属性,不同于直接从像素图像中提取特征。通常点云目标检测头会输出目标的类别、3D 边界框的坐标和目标的置信度等信息,与基于图像的 3D 目标检测器并无不同。

基于点的 3D 目标检测器最重要的模块为特征学习器部分,可分为点云采样模块和特征学习模块:

1)点云采样。点云采样决定输入点云数据的密度和分布,对后续特征学习和目标检测的性能有重要的影响。常见的传统点云采样方式有:均匀采样、全局采样、局部采样和自适应采样等。最近工作往往关注以深度学习模型来指导点云进行采样,通过学习点云的重要性权重和点云的分布来获得更优质的采样点云。点云采样中很经典的工作是 PointNet++ 为解决 PointNet 计算量过大的问题时使用的最远点采样算法,即从所有输入点云中采样了一定数量的点,希望这些点能够包含尽可能多的有用信息,以放大不同采样点之间的差异。

2)特征学习。基于点的目标检测器在对单个点进行特征提取时,可以采用单个点本身来描述该点的特征,但更有效的方式是使用该点的邻近点来获得局部特征。这种局部特征提取的操作需要对模型的性能和开销做出一定平衡。增加邻近点的数量可以使模型具有更强的表达能力,但同时会增大内存开销,且使用过多的邻近点来获取某一点对应的特征也会导致细粒度信息的丢失。减少邻近点的数量则可能导致局部特征描述不够充分,丢失相关的语义信息。

目前,基于原始点云数据的目标检测研究,大多是对上述采样模块和特征学习模块两个组件及目标检测分类头进行改进的结果,如 PointRCNN (Shi 等, 2019)、IPOD (intensive point-based object detector) (Yang 等, 2018c)、STD (sparse-to-dense 3D object detector) (Yang 等, 2019)、3DSSD (Yang 等, 2020)、Point-GNN (Shi 和 Rajkumar, 2020)、StarNet (Ngiam

等,2019)和Pointformer(Pan等,2021)。这些基于点的目标检测方法在小规模数据集上往往有着不错的效果,能表现出比较高的检测精度,但是在大规模自动驾驶数据集上难以达到实时且高效的检测效果。

3.2 基于体素的目标检测

基于体素(voxel)的目标检测算法将不规则的点云转化为紧凑形状的体素表示,然后通过3D卷积网络提取目标检测所需的点云特征。与基于点的目标检测算法相比,使用体素进行3D目标检测将更加高效。这种将点云体素化后再处理的方法已经受到了研究人员的广泛关注,代表性工作有Vote3D(Wang和Posner,2015)、Vote3Deep(Engelcke等,2017)、3D-FCN(Li,2017)、VoxelNet(Zhou和Tuzel,2018)、SECOND(sparse embedded convolutional detection)(Yan等,2018)、PointPillars(Lang等,2019)和CenterPoint(Yin等,2021)。

VoxelNet是基于体素进行目标检测的开山之作,是一种端到端的3D目标检测框架。VoxelNet将三维点云划分为一定数量的体素,经过点的随机采样以及归一化后,对每一个非空体素使用若干个体素特征编码器(voxel feature encoding, VFE)进行局部特征提取,得到体素级特征,然后经过3D中间卷积层完成进一步特征提取(增大感受野并学习几何空间表示),最后使用选框提议网络对物体进行分类检测与位置回归。相比VoxelNet,SECOND(sparse embedded convolutional detection)(Yan等,2018)最大的创新点在于将3D卷积替换为稀疏卷积,提高了Voxel模型的运行速度,并且减少了内存占用,同时提出一种新的朝向角度损失函数和新的数据增强方法,取得了不错的效果。

SECOND算法和VoxelNet都是将点云数据划分为一个个体素,构成规则的、密集分布的体素集。而PointPillars则采用了一种不同于上述两种思路的点云建模方法,提出将点云数据转化为一个个的Pillar柱体编码。首先利用柱体的方式将点云转化为稀疏的俯视角度伪图像,再使用2D神经网络进行特征学习和物体位置回归,大大提高了基于体素的目标检测网络的性能,为后续相关工作提供了重要思路和技术基础。

总的来说,将点云表示为体素的形式,并使其规则化,有利于后续网络进行高效的特征处理。目前大多数基于体素的目标检测算法改进的内容包括以

下几点:1)提高网络对体素空间的感知能力;2)改进点云的采样过程,以更好地保留物体的三维结构信息;3)与其他通用模型进行融合以及有效的结构迁移。然而体素化表示点云也存在一定缺陷,这种数据表示方式仍然难以避免因忽略有效信息而带来的量化误差,点云采样过程的不确定性也会影响检测模型结果的稳定性和一致性。

3.3 基于BEV的目标检测

对基于点云的BEV目标检测算法而言,在特征提取部分,主要有两种方式将点云数据转换为BEV表示。根据流水线顺序,可分为BEV前特征提取和BEV后特征提取方法,分别对应主干网络的输入来自3D表示和BEV表示。

BEV前特征提取算法先对原始点云进行体素化,然后对得到的体素网格表示做3D特征提取,最后转为BEV表示。如PV-RCNN(point-voxel RCNN)(Shi等,2020a)是BEV前特征提取方式中的经典工作。PV-RCNN提出一种体素到关键点的编码方法,首先将空间中的点云数据体素化,利用稀疏卷积网络进行多次的特征提取与下采样。而后PV-RCNN使用每一层的稀疏卷积网络特征图来提取多尺度特征,再将得到的特征投影到鸟瞰图上进行一阶段的目标检测,最后利用先前的多尺度特征对一阶段的目标检测结果进行优化以得到更精确的检测结果。类似基于BEV前特征提取的方法还有SA-SSD(structure aware single-stage 3D object detection)(He等,2020)、Voxel R-CNN(Deng等,2021)等。

BEV后特征提取方法则首先将原始点云数据压扁到BEV视角,再对得到的BEV视图进行2D特征提取,送到目标检测头进行物体定位。其中具有代表性且最早将点云数据转化为BEV表示的工作是MV3D(multi-view 3D)(Chen等,2017)。MV3D将点云数据离散化为BEV网格后,根据网格中点的高度、强度和密度特征,获取表示BEV网格的特征。不过由于单个BEV网格有许多点,这一处理过程中,网络信息损失较大。不少的BEV后特征提取方法都遵循了类似的模式,用BEV网格中的统计数据表示点云,例如最大高度和强度平均值。前文提到的PointPillars本质上也可归类为基于BEV后特征提取的目标检测方法,其他的相关工作还有PIXOR(PIXel-wise ORiented)(Yang等,2018b)、BirdNet(Beltrán等,2018)、RT3d(real-time 3D)(Zeng等,

2018)和YOLO3D(Ali等,2018)等。

3.4 基于距离图像的目标检测

距离图像(range image)是一种密集紧凑的2D表示,如图6所示,每个像素都包含3D深度信息。距离图像的优势在于信息稠密且无需进行栅格化,可以直接使用2D卷积网络进行特征提取及快速的领域查询,且距离图像的检测范围可以和传感器真实的检测范围一样远。

RangeDet(range view detection)(Fan等,2021)是基于距离图像进行3D目标检测的经典工作之一。RangeDet认为限制距离图像性能的因素主要有3个:距离图像中尺寸变化大、距离图像中的2D卷积丢失3D信息、距离图像中提取的特征十分稠密但是难以有效利用。对此,RangeDet采用距离图像金字塔以解决尺寸问题,使用Meta卷积从2D距离图像表示中获取3D几何信息,然后使用加权非最大抑制算法来充分提高检测器的性能,最终模型在Waymo(Sun等,2020)数据集上取得了不错的性能提升。

其他基于距离图像进行3D目标检测的方法还有LaserNet(Meyer等,2019)、LaserFlow(Meyer等,2021)、RangeRCNN(Liang等,2020)、RangeIoUDet(Liang等,2021)、RCD(range conditioned dilated convolution)(Bewley等,2020)和RSN(range sparse net)(Sun等,2021)等。

当然,基于距离图像进行3D目标检测也有其固有的局限性,这种检测方式往往很容易受到遮挡和尺度变化的影响,导致3D信息的损失,使得模型难以处理物体之间的遮挡。因此,从距离视图中提取特征,并从BEV视图进行目标检测有潜力成为基于距离图像进行3D目标检测最实用的解决方案。

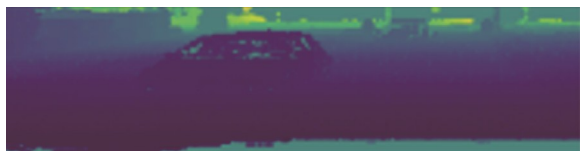


图6 距离图像(Wang等,2023)

Fig. 6 Range image(Wang et al., 2023)

4 融合点云和图像的目标检测

相机可以提供高分辨率的颜色图像,有助于模型提取丰富的语义特征。激光雷达则擅长3D定位,

提供精确的距离和空间信息,便于准确地捕捉到物体的几何结构。为了提升3D目标检测模型的性能,许多研究工作已经将相机和激光雷达的信息进行融合。通过融合两种传感器的信息,多模态目标检测模型可以综合利用每个模态各自的优势,提高目标检测的准确性和鲁棒性(刘祥等,2024)。

4.1 按照融合阶段分类

4.1.1 前期融合

前期融合又称做数据级融合,需要点云和图像数据在各自的分支都采用特征提取器,而后在特征层中进行一定的交互,将图像的知识整合到点云中,其结构如图7所示。经过整合的点云可以看做是经过图像增强的点云,模型将增强后的点云数据反馈给基于LiDAR的3D目标检测器,可以让模型获得更加丰富的语义信息。

前期融合方法通常会带来较高的推理延迟,融合步骤需要复杂的2D目标检测或语义分割网络,给模型带来较高的时间成本。

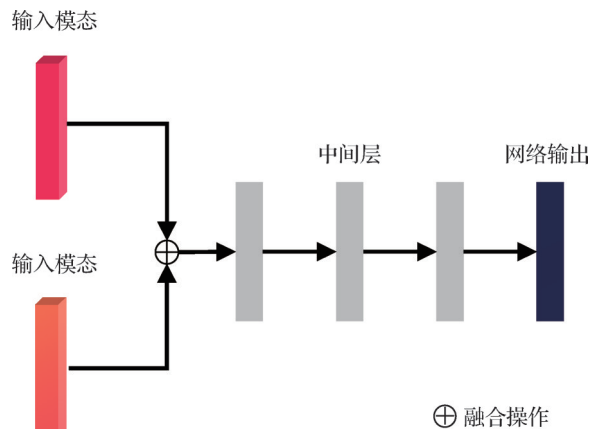


图7 前期融合(Wang等,2023)

Fig. 7 Early fusion(Wang et al., 2023)

4.1.2 深度融合

深度融合是指在基于LiDAR的3D目标检测器的中间阶段融合图像和激光雷达特征(例如骨干网络阶段、预选框生成阶段和RoI细化阶段),其网络结构如图8所示。

相比前期融合,深度融合方法可以对多模态表示进行更深层的融合,产生更高质量的3D检测框与深层特征,且降低了前期融合所需的计算量。然而图像数据和LiDAR数据特征本质上是异构的,深度融合过程要考虑在统一的特征空间下保留好图像数

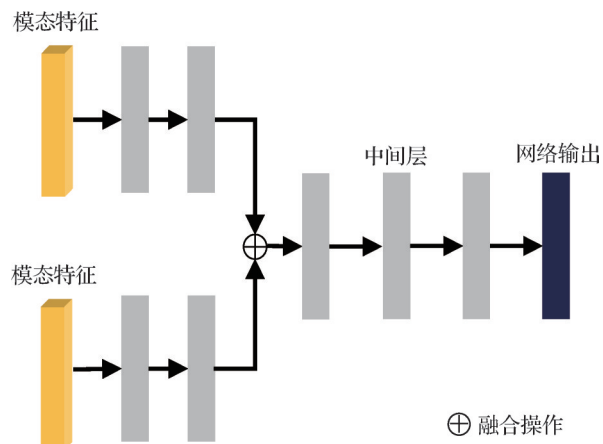


图8 深度融合(Wang等,2023)

Fig. 8 Deep fusion(Wang et al. ,2023)

据包含的语义信息和LiDAR数据对应的空间几何信息,以免损害模型目标检测任务的性能。

4.1.3 后期融合

后期融合也称决策融合,是指将来自不同传感器或模态的决策输出进行综合,以得出最终的目标检测结果,其网络结构如图9所示。与前期融合和深度融合相比,后期融合能更好地利用现有网络对每个模态的指导,容易得到每个模态的反馈。然而决策融合的缺点是:不能使用融合过程中丰富的深度融合特征,且不同模态的分支预测结果相差过大时,难以给出精确结果。

后期融合的选择需要考虑模态之间的相关性、不确定性以及任务的要求,合理的后期融合策略可以大大提高多模态环境目标检测的性能,消除不同模态之间的噪声和不一致性,使系统更加鲁棒和可靠。

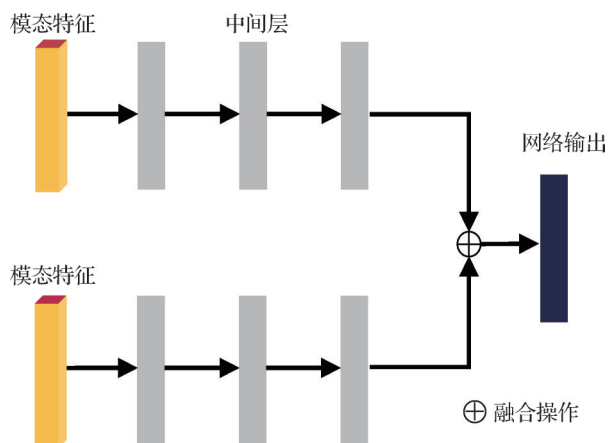


图9 后期融合(Wang等,2023)

Fig. 9 Late fusion(Wang et al. ,2023)

4.2 按照数据粒度分类

随着不同的融合方式及阶段的细分,越来越多的算法不能够单纯按照融合阶段进行分类(黄哲等,2023),例如MMF(multi-task multi-sensor fusion)(Liang等,2019)不仅有深度融合,也有后期的决策融合。同样,PointPainting(Vora等,2020)也不属于任何一种上述提到的融合策略,而是采用串行融合方式。尤其是前期融合和深度融合阶段,仅仅按照时间进行划分并不能较好地地区分不同多模态目标检测算法。

前期融合和深度融合阶段可以统称为特征级融合。特征级融合往往关注特征层之间的相互作用,关注不同特征层的数据表示,结合不同模态的特征实现多尺度和多层次融合。特征级融合可以按照特征层粒度的不同分为基于ROI粒度的特征融合、基于体素粒度的特征融合和基于点的特征融合。

基于ROI粒度的特征融合包括MV3D(Chen等,2017)、PointFusion(Xu等,2018)、AVOD(aggregate view object detection)(Ku等,2018)、F-Pointnet(Qi等,2018)、F-Convnet(Wang和Jia,2019b)、SCANet(spatial-channel attention network)(Lu等,2019)、RoarNet(region approximation refinement)(Shin等,2019)和SIFRNet(scale invariant and feature reweighting networks)(Zhao等,2019)。

基于体素粒度的特征融合包括ContFuse(Liang等,2018)、MVX-Net(multimodal VoXelNet)(Sindagi等,2019)、MMF(multi-task multi-sensor fusion)(Liang等,2019)和3D-CVF(3D cross-view spatial feature fusion)(Yoo等,2020)。

基于点的特征融合包括IPOD(Yang等,2018c)、PointPainting(Vora等,2020)、HorizonLiDAR3D(Ding等,2020b)、MVAf(multi-view adaptive fusion)(Wang等,2020)、VPF-Net(voxel-pixel fusion network)(Wang等,2021a)、PI-RCNN(pointcloud-image RCNN)(Xie等,2020)、FuseSeg(fuse segmentation)(Krispel等,2020)、CenterPointV2(Yin等,2021)、EPNet(enhancing point features network)(Huang等,2020)和PointAugmenting(Wang等,2021b)。

上述不同粒度的特征融合目标检测算法按照时间排序如图10所示。早期的特征融合算法的特征粒度普遍较为粗糙,随着多模态目标检测技术的逐渐成熟,越来越多的研究者将目光投向细粒度、多种类、高性能的多模态融合算法研究。

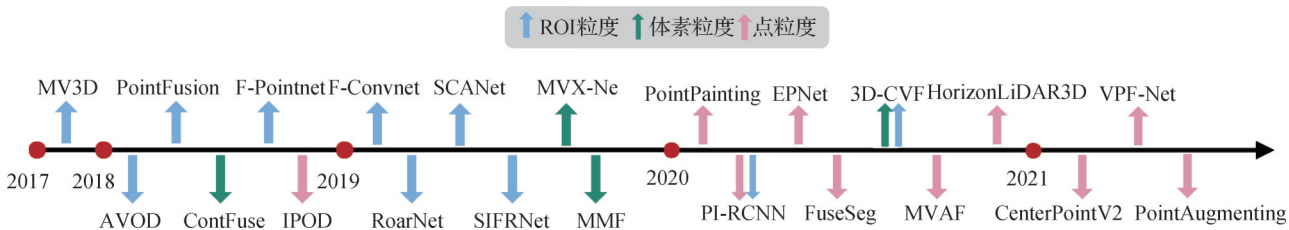


图 10 不同粒度特征融合的目标检测算法 (Wang 等, 2023)

Fig. 10 Object detection algorithm fused with features of different granularity (Wang et al., 2023)

4.3 常见 3D 目标检测数据集上的部分算法

KITTI 数据集对于 3D 目标检测任务将标签细分为机动车、行人和自行车等目标,通过比较预测边框和真实边框的重叠程度以及阈值的大小判定 3D 目标定位结果的正确性,通过置信度分数和阈值的比较确定目标识别结果的正确性,最终使用平均正确率来评估单类的目标检测模型结果。

对于汽车目标, KITTI 要求 3D 边界框重叠为 70%,而对于行人和骑自行车的人,重叠率为 50%即为正确。同时 KITTI 数据集根据边界框高度、遮挡和截断程度指标将标注目标划分为简单(easy)、中等(moderate)和困难(hard)3 个难度级别。目前在 KITTI 数据集上领先的部分 3D 目标检测算法如表 1—表 3 所示,展示了模型对应的平均正确率和用时。

表 1 KITTI 数据集上领先的 3D 目标检测算法 (车辆目标)

Table 1 Leading 3D object detection algorithms on KITTI dataset(car target)

模型	中等/%	简单/%	困难/%	时间/s
VirConv(Wu 等, 2023b)	87.20	92.48	82.45	0.09
UDeerPEP(Dong 等, 2023)	86.72	91.65	82.59	0.10
TED(Wu 等, 2023a)	85.28	91.61	80.68	0.10
LoGoNet(Li 等, 2023a)	85.06	91.80	80.74	0.10
SFD(Wu 等, 2022b)	84.76	91.73	77.92	0.10

注:加粗字体表示各列平均正确率最优结果。

nuScenes 数据集的评价指标相对 KITTI 数据集要更为丰富,包括平均精度均值(mean average percision, mAP)、平均平移误差(mean average translation error, mATE)、平均角度误差(mean average orientation error, mAOE)、平均属性误差(mean average attribute error, mAAE)和各项指标平均加权(nuScenes

表 2 KITTI 数据集上领先的 3D 目标检测算法 (骑行目标)

Table 2 Leading 3D object detection algorithms on KITTI dataset(cyclist target)

模型	中等/%	简单/%	困难/%	时间/s
TED(Wu 等, 2023a)	74.12	88.82	66.84	0.1
CasA++(Wu 等, 2022a)	73.79	87.76	66.84	0.1
CasA(Wu 等, 2022a)	73.47	87.91	66.17	0.1
LoGoNet(Li 等, 2023a)	71.70	84.47	64.67	0.1
BSAODet(Xiao 等, 2023)	70.48	83.17	62.46	0.4

注:加粗字体表示各列平均正确率最优结果。

表 3 KITTI 数据集上领先的 3D 目标检测算法 (行人目标)

Table 3 Leading 3D object detection algorithms on KITTI dataset(pedestrian target)

模型	中等/%	简单/%	困难/%	时间/s
CasA++(Wu 等, 2022a)	49.29	56.33	46.70	0.1
TED(Wu 等, 2023a)	49.21	55.85	46.52	0.1
VPFNet(Zhu 等, 2023)	48.36	54.65	44.98	0.2
LoGoNet(Li 等, 2023a)	47.43	53.07	45.22	0.1
CasA(Wu 等, 2022a)	47.09	54.04	44.56	0.1

注:加粗字体表示各列平均正确率最优结果。

detection score, NDS)等。

目前在 nuScenes 数据集上领先的部分 3D 目标检测算法如表 4 所示。表 4 展示了模型对应的输入模态、mAP、mATE 和 NDS。

Waymo 数据集也使用常见的平均精度指标(average precision, AP),但 Waymo 数据集计算 AP 时将其计算为 P-R(precession-recall)曲线下的面积,而非 KITTI 使用的插值方法。此外,Waymo 还引入了平均精度航向(mean average precision weighted by heading, mAPH)指标,用以计算航向信息的相关误差。关于检测 3D 目标的难度,Waymo 定义了 Level 1 和 Level 2 两个不同的级别。Level 2 中分配的目标

被认为是最难检测的目标,对应的 LiDAR 点数量更少,对基于 LiDAR 和融合图像与 LiDAR 的目标检测算法更具挑战性。

目前在 Waymo 数据集上领先的部分 3D 目标检测算法如表 5 所示。表 5 展示了模型对应的输入模

态、mAP、mATE 和 mAAE。

总体来说,在 KITTI 数据集上排名靠前的方法主要是基于纯点云的目标检测方法。与之不同,在最新数据集,如 nuScenes 和 Waymo 数据集上,排名靠前的方法则是以多模态融合方法为主。

表 4 nuScenes 数据集上领先的 3D 目标检测算法
Table 4 Leading 3D object detection algorithms on nuScenes dataset

模型	输入模态	mAP/%	mATE/m	NDS/%
EA-LSS(Hu 等,2023)	图像、激光雷达	76.6	0.234	77.6
BEVFusion++	图像、激光雷达	75.7	0.236	77.6
BEVFusion4D(Cai 等,2023)	图像、激光雷达	76.8	0.229	77.2
MMFusion(Ahmad 和 Del Bue,2023)	图像、激光雷达、毫米波雷达	75.0	0.220	77.1
MegFusion(Yan 等,2023)	图像、激光雷达	75.3	0.233	77.0

表 5 Waymo 数据集上领先的 3D 目标检测算法
Table 5 Leading 3D object detection algorithms on Waymo dataset

模型	输入模态	mAP/%	mAPH/%
DetZero(Ma 等,2023)	激光雷达	86.34	85.15
CTRL_FSD_TTA	激光雷达	84.10	82.52
LoGoNet_Ens v2(Li 等,2023a)	图像、激光雷达	83.36	81.96
HRI_ADLAB_HZ	激光雷达	82.57	81.31
WAVE3 LiPER Det v1	激光雷达	82.55	81.21

5 总结及未来展望

本文根据环境目标检测中基于图像和点云的相关技术,对基于深度学习的环境目标检测进行了阐述。首先介绍了 3D 目标检测的常用数据集和数据格式,其次按照不同的数据输入形式和模型内部的数据表示形式对 3D 目标检测算法进行划分。

尽管环境目标检测在自动驾驶及数字仿真技术领域有着很大的潜力,但目前仍然存在着一些问题和挑战:

1)数据获取困难。与 2D 图像数据集相比,获取和标注 3D 数据要更加困难,需要使用激光雷达或深度摄像头等特殊的传感器和设备,需要更加专业的标注人员进行复杂的信息标注,如物体的三维边界框、位姿和运动轨迹等,是一项相当耗时的任务。且对于多模态目标检测算法而言,可供使用的数据增强器类型相对较少,这点同样限制了模型的性能和

泛化性提升,如何在多模态数据集上进行有效的数据增强仍是一个需要探索的问题。

2)传感器故障和感知状态监控。感知系统尤其是自动驾驶系统,需要稳健地感知环境中的物体,同样要准确地检测出传感器自身发生的故障并采取措施。比如设计者可以采用多个相同传感器来增加信息的冗余度,通过比较每个传感器的操作来确定感知系统的稳定性。尤其对于多模态感知系统而言,以防某一模态传感器发生故障导致系统无法正常工作,需要使感知系统能够使用单一模态数据进行环境目标检测,保证在部分传感器故障的情况下稳定运行。

3)恶劣天气条件下的感知。在极端恶劣天气条件下的感知也是环境目标检测所面临的一个重要挑战。如对自动驾驶感知系统而言,在多雪多雾的条件下,无论是基于图像的系统还是基于 LiDAR 的系统都会存在感知困难。对于基于图像的系统而言,在多雪多雾的条件下,图像中的目标物体可能被雪

花、雾气或雨滴等遮挡和模糊化,从而导致目标检测的准确性下降。此外,由于光线的反射和折射,图像的对比度和清晰度也会受到影响,使得目标的边界和细节变得模糊不清。对于基于LiDAR的系统而言,激光束可能会被雪花、雾气或雨滴等散射介质所散射和吸收。这会导致激光雷达接收到的反射信号变弱或失真,从而使得检测目标的位置和形状难以准确地重建。基于图像和LiDAR的目标检测算法仍然需要进一步的研究和技术创新,以提高感知系统在恶劣天气条件下的鲁棒性和可靠性。

未来对环境目标检测研究的发展方向可包括以下几个方面:

1)时间序列模型引入。现实生活中人类驾驶员是依靠连续的视觉感知来得到周围的环境信息,然而目前大多数工作都是在单帧的基础上解决环境感知问题。随着目标检测过程的进行,感知系统可以不断地生成图像序列和点云序列,这些序列数据中包含着隐含的物体结构以及运动轨迹信息。以上述序列数据作为基础,以此增加时间线索和物体的位置约束,将会使得环境目标检测过程具有更好的鲁棒性和环境理解,以实现更智能的感知模型。

2)基于2D深度估计及多视角2D图像的目标检测方法。基于2D图像的环境目标检测在性能上与基于LiDAR的方法相比,仍然相差较大。提升2D图像进行深度估计的准确性,或采用多视角2D图像都可以作为使用2D图像进行3D目标检测的有效手段。这两种使用纯图像的环境目标检测方法,不仅可以直接作为3D目标检测的骨干,也可以为基于图像和LiDAR融合的多模态3D目标检测算法奠定基础。

3)多模态感知算法精度与速度的平衡。在自动驾驶和机器人导航领域中,环境目标检测过程需要在实时性要求较高的场景下进行。这是因为自动驾驶系统需要快速响应和适应变化的道路条件和交通情况,以确保车辆的安全性和稳定性。在多模态感知算法的设计过程中,往往加入新的模态与新的高级特征会导致算法在速度上损失,同时带来过大的开销。如何有效地筛选并简化多模态目标检测模型,以保证同时达到准确、实时和高效的要求,是多模态感知算法研究的未来发展趋势。

4)基于大模型的环境目标检测。即利用大型深度学习模型为点云和图像融合提供更强大的表征学

习能力。大模型的引入有望缓解点云和图像融合过程中的数据稀疏性问题,使得模型更好地捕捉图像和点云之间丰富的语义信息,更全面地理解环境中的物体,增强对局部遮挡和复杂背景的适应性。此外,大模型与环境目标检测算法的融合,将进一步促进自动驾驶系统的感知决策一体化进程,提升感知模型的可解释性、安全性和多模块可迭代性,为未来全栈自动驾驶、智能交通等研究领域的发展提供关键支持。

致谢 本文由中国图象图形学学会虚拟现实专委会组织撰写,该专委会链接为<https://www.csig.org.cn/16/201612/49314.html>。

参考文献(References)

- Ahmad J and Del Bue A. 2023. mmFUSION: multimodal fusion for 3D objects detection [EB/OL]. [2023-12-21].
<https://arxiv.org/pdf/2311.04058.pdf>
- Aldoma A, Marton Z C, Tombari F, Wohlkinger W, Potthast C, Zeisl B, Rusu R B, Gedikli S and Vincze M. 2012. Tutorial: point cloud library: three-dimensional object recognition and 6DOF pose estimation. *IEEE Robotics and Automation Magazine*, 19(3): 80-91 [DOI: 10.1109/MRA.2012.2206675]
- Ali W, Abdelkarim S, Zidan M, Zahran M and Sallab A E. 2018. YOLO3D: end-to-end real-time 3D oriented object bounding box detection from LiDAR point cloud//*Proceedings of 2018 European Conference on Computer Vision*. Munich, Germany: Springer: 716-728 [DOI: 10.1007/978-3-030-11015-4_54]
- Arnold E, Al-Jarrah O Y, Dianati M, Fallah S, Oxtoby D and Mouzakitis A. 2019. A survey on 3D object detection methods for autonomous driving applications. *IEEE Transactions on Intelligent Transportation Systems*, 20(10): 3782-3795 [DOI: 10.1109/ITITS.2019.2892405]
- Arora H, Loeff N, Forsyth D A and Ahuja N. 2007. Unsupervised segmentation of objects using efficient learning//*Proceedings of 2007 IEEE Conference on Computer Vision and Pattern Recognition*. Minneapolis, USA: IEEE: 1-7 [DOI: 10.1109/CVPR.2007.383011]
- Bao W T, Xu B and Chen Z Z. 2020. MonoFENet: monocular 3D object detection with feature enhancement networks. *IEEE Transactions on Image Processing*, 29: 2753-2765 [DOI: 10.1109/TIP.2019.2952201]
- Beltrán J, Guindel C, Moreno F M, Cruzado D, García F and De La Escalera A. 2018. BirdNet: a 3D object detection framework from LiDAR information//*Proceedings of the 21st International Conference on Intelligent Transportation Systems (ITSC)*. Maui, USA: IEEE: 3517-3523 [DOI: 10.1109/ITSC.2018.8569311]

- Bewley A, Sun P, Mensink T, Anguelov D and Sminchisescu C. 2020. Range conditioned dilated convolutions for scale invariant 3D object detection//Proceedings of the 4th Conference on Robot Learning. Cambridge, USA: PMLR: 627-641
- Bochkovskiy A, Wang C Y and Liao H Y M. 2020. YOLOv4: optimal speed and accuracy of object detection [EB/OL]. [2023-12-21]. <https://arxiv.org/pdf/2004.10934.pdf>
- Brazil G and Liu X M. 2019. M3D-RPN: monocular 3D region proposal network for object detection//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea (South): IEEE: 9286-9295 [DOI: 10.1109/ICCV.2019.00938]
- Brazil G, Pons-Moll G, Liu X M and Schiele B. 2020. Kinematic 3D object detection in monocular video//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 135-152 [DOI: 10.1007/978-3-030-58592-1_9]
- Caesar H, Bankiti V, Lang A H, Vora S, Liong V E, Xu Q, Krishnan A, Pan Y, Baldan G and Beijbom O. 2020. nuScenes: a multi-modal dataset for autonomous driving//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE: 11618-11628 [DOI: 10.1109/CVPR42600.2020.01164]
- Cai H X, Zhang Z Y, Zhou Z Y, Li Z Y, Ding W B and Zhao J H. 2023. BEVFusion4D: learning LiDAR-camera fusion under bird's-eye-view via cross-modality guidance and temporal aggregation [EB/OL]. [2023-12-21]. <https://arxiv.org/pdf/2303.17099.pdf>
- Cao J L, Li Y L, Sun H Q, Xie J, Huang K Q and Pang Y W. 2022. A survey on deep learning based visual object detection. *Journal of Image and Graphics*, 27(6): 1697-1722 (曹家乐, 李亚利, 孙汉卿, 谢今, 黄凯奇, 庞彦伟. 2022. 基于深度学习的视觉目标检测技术综述. *中国图象图形学报*, 27(6): 1697-1722) [DOI: 10.11834/jig.220069]
- Catmull E. 1998. Computer display of curved surfaces. *Seminal graphics: pioneering efforts that shaped the field*, 1: 35-41 [DOI: 10.1145/280811.280920]
- Chang J L and Wetzstein G. 2019. Deep optics for monocular depth estimation and 3D object detection//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea (South): IEEE: 10192-10201 [DOI: 10.1109/ICCV.2019.01029]
- Charles R Q, Su H, Kaichun M and Guibas L J. 2017. PointNet: deep learning on point sets for 3D classification and segmentation//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA: IEEE: 77-85 [DOI: 10.1109/CVPR.2017.16]
- Chen X Z, Ma H M, Wan J, Li B and Xia T. 2017. Multi-view 3D object detection network for autonomous driving//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA: IEEE: 6526-6534 [DOI: 10.1109/CVPR.2017.691]
- Chen Y, Liu S, Shen X and Jia J. 2019. Fast point R-CNN//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea (South): IEEE: 9774-9783 [DOI: 10.1109/ICCV.2019.00987]
- Cui Y D, Chen R, Chu W B, Chen L, Tian D X, Li Y and Cao D P. 2022. Deep learning for image and point cloud fusion in autonomous driving: a review. *IEEE Transactions on Intelligent Transportation Systems*, 23(2): 722-739 [DOI: 10.1109/TITS.2020.3023541]
- Dalal N and Triggs B. 2005. Histograms of oriented gradients for human detection//Proceedings of 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). San Diego, USA: IEEE: 886-893 [DOI: 10.1109/CVPR.2005.177]
- Deng J J, Shi S S, Li P W, Zhou W G, Zhang Y Y and Li H Q. 2021. Voxel R-CNN: towards high performance voxel-based 3D object detection//Proceedings of the 35th AAAI Conference on Artificial Intelligence. [s.l.]: AAAI: 1201-1209 [DOI: 10.1609/aaai.v35i2.16207]
- Ding M Y, Huo Y Q, Yi H W, Wang Z, Shi J P, Lu Z W and Luo P. 2020a. Learning depth-guided convolutions for monocular 3D object detection//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE: 11669-11678 [DOI: 10.1109/CVPR42600.2020.01169]
- Ding Z Z, Hu Y H, Ge R Z, Huang L, Chen S J, Wang Y and Liao J. 2020b. 1st place solution for Waymo open dataset challenge—3D detection and domain adaptation [EB/OL]. [2023-12-21]. <https://arxiv.org/pdf/2006.15505.pdf>
- Dong Z C, Ji H, Huang X F, Zhang W K, Zhan X and Chen J B. 2023. PeP: a Point enhanced Painting method for unified point cloud tasks [EB/OL]. [2023-12-21]. <https://arxiv.org/pdf/2310.07591.pdf>
- Duan K W, Bai S, Xie L X, Qi H G, Huang Q M and Tian Q. 2019. CenterNet: keypoint triplets for object detection//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea (South): IEEE: 6568-6577 [DOI: 10.1109/ICCV.2019.00667]
- Engelcke M, Rao D, Wang D Z, Tong C H and Posner I. 2017. Vote3Deep: fast object detection in 3D point clouds using efficient convolutional neural networks//Proceedings of 2017 IEEE International Conference on Robotics and Automation (ICRA). Singapore, Singapore: IEEE: 1355-1361 [DOI: 10.1109/ICRA.2017.7989161]
- Fan L, Xiong X, Wang F, Wang N Y and Zhang Z X. 2021. RangeDet: in defense of range view for LiDAR-based 3D object detection//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, Canada: IEEE: 2898-2907 [DOI: 10.1109/ICCV48922.2021.00291]
- Feng D, Haase-Schütz C, Rosenbaum L, Hertlein H, Gläser C, Timm

- F, Wiesbeck W and Dietmayer K. 2021. Deep multi-modal object detection and semantic segmentation for autonomous driving: datasets, methods, and challenges. *IEEE Transactions on Intelligent Transportation Systems*, 22(3): 1341-1360 [DOI: 10.1109/TITS.2020.2972974]
- Fu H, Gong M M, Wang C H, Batmanghelich K and Tao D C. 2018. Deep ordinal regression network for monocular depth estimation// *Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Salt Lake City, USA: IEEE: 2002-2011 [DOI: 10.1109/CVPR.2018.00214]
- Geiger A, Lenz P and Urtasun R. 2012. Are we ready for autonomous driving? The KITTI vision benchmark suite// *Proceedings of 2012 IEEE Conference on Computer Vision and Pattern Recognition*. Providence, USA: IEEE: 3354-3361 [DOI: 10.1109/CVPR.2012.6248074]
- Girshick R. 2015. Fast R-CNN// *Proceedings of 2015 IEEE International Conference on Computer Vision (ICCV)*. Santiago, Chile: IEEE: 1440-1448 [DOI: 10.1109/ICCV.2015.169]
- Girshick R, Donahue J, Darrell T and Malik J. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation// *Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition*. Columbus, USA: IEEE: 580-587 [DOI: 10.1109/CVPR.2014.81]
- Godard C, Aodha O M and Brostow G J. 2017. Unsupervised monocular depth estimation with left-right consistency// *Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Honolulu, USA: IEEE: 6602-6611 [DOI: 10.1109/CVPR.2017.699]
- Guo Y L, Wang H Y, Hu Q Y, Liu H, Liu L and Bennamoun M. 2021. Deep learning for 3D point clouds: a survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(12): 4338-4364 [DOI: 10.1109/TPAMI.2020.3005434]
- He C H, Zeng H, Huang J Q, Hua X S and Zhang L. 2020. Structure aware single-stage 3D object detection from point cloud// *Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Seattle, USA: IEEE: 11870-11879 [DOI: 10.1109/CVPR42600.2020.01189]
- He K M, Gkioxari G, Dollár P and Girshick R. 2017. Mask R-CNN// *Proceedings of 2017 IEEE International Conference on Computer Vision (ICCV)*. Venice, Italy: IEEE: 2980-2988 [DOI: 10.1109/ICCV.2017.322]
- He T and Soatto S. 2019. Mono3D++: monocular 3D vehicle detection with two-scale 3D hypotheses and task priors// *Proceedings of the 33rd AAAI Conference on Artificial Intelligence*. Honolulu, USA: AAAI: 8409-8416 [DOI: 10.1609/aaai.v33i01.33018409]
- Hu H T, Wang F Y, Su J W, Hu L F, Feng T P, Zhang Z K and Zhang W Z. 2023. EA-BEV: edge-aware bird's-eye-view projector for 3D object detection [EB/OL]. [2023-12-21]. <https://arxiv.org/pdf/2303.17895.pdf>
- Huang T T, Liu Z, Chen X W and Bai X. 2020. EPNet: enhancing point features with image semantics for 3d object detection// *Proceedings of the 16th European Conference on Computer Vision*. Glasgow, UK: Springer: 35-52 [DOI: 10.1007/978-3-030-58555-6_3]
- Huang Z, Wang Y C and Li D Y. 2023. A survey of 3D object detection algorithms. *Chinese Journal of Intelligent Science and Technology*, 5(1): 7-31 (黄哲, 王永才, 李德英. 2023. 3D目标检测方法研究综述. *智能科学与技术学报*, 5(1): 7-31) [DOI: 10.11959/j.issn.2096-6652.202312]
- Jin S, Li X P, Yang F and Zhang W G. 2023. 3D object detection in road scenes by pseudo-LiDAR point cloud augmentation. *Journal of Image and Graphics*, 28(11): 3520-3535 (晋帅, 李焯鹏, 杨凤, 张为公. 2023. 伪激光点云增强的道路场景三维目标检测. *中国图象图形学报*, 28(11): 3520-3535) [DOI: 10.11834/jig.220986]
- Krispel G, Opitz M, Waltner G, Possegger H and Bischof H. 2020. FuseSeg: LiDAR point cloud segmentation fusing multi-modal data// *Proceedings of 2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*. Snowmass, USA: IEEE: 1863-1872 [DOI: 10.1109/WACV45572.2020.9093584]
- Ku J, Mozifian M, Lee J, Harakeh A and Waslander S L. 2018. Joint 3D proposal generation and object detection from view aggregation// *Proceedings of 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. Madrid, Spain: IEEE: 1-8 [DOI: 10.1109/IROS.2018.8594049]
- Ku J, Pon A D and Waslander S L. 2019. Monocular 3D object detection leveraging accurate proposals and shape reconstruction// *Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Long Beach, USA: IEEE: 11859-11868 [DOI: 10.1109/CVPR.2019.01214]
- Lang A H, Vora S, Caesar H, Zhou L B, Yang J and Beijbom O. 2019. PointPillars: fast encoders for object detection from point clouds// *Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Long Beach, USA: IEEE: 12689-12697 [DOI: 10.1109/CVPR.2019.01298]
- Law H and Deng J. 2020. CornerNet: detecting objects as paired keypoints. *International Journal of Computer Vision*, 128(3): 642-656 [DOI: 10.1007/s11263-019-01204-1]
- Li B. 2017. 3D fully convolutional network for vehicle detection in point cloud// *Proceedings of 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. Vancouver, Canada: IEEE: 1513-1518 [DOI: 10.1109/IROS.2017.8205955]
- Li B Y, Ouyang W L, Sheng L, Zeng X Y and Wang X G. 2019. GS3D: an efficient 3D object detection framework for autonomous driving// *Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Long Beach, USA: IEEE: 1019-1028 [DOI: 10.1109/CVPR.2019.00111]
- Li P X, Zhao H C, Liu P F and Cao F D. 2020. RTM3D: real-time mon-

- ocular 3d detection from object keypoints for autonomous driving//
Proceedings of the 16th European Conference on Computer Vision.
Glasgow, UK: Springer: 644-660 [DOI: 10.1007/978-3-030-
58580-8_38]
- Li X, Ma T, Hou Y N, Shi B T, Yang Y C, Liu Y Q, Wu X J, Chen
Q, Li Y K, Qiao Y and He L. 2023a. LoGoNet: towards accurate
3D object detection with local-to-global cross- modal fusion//Pro-
ceedings of 2023 IEEE/CVF Conference on Computer Vision and
Pattern Recognition (CVPR). Vancouver, Canada: IEEE: 17524-
17534 [DOI: 10.1109/CVPR52729.2023.01681]
- Li Z Q, Wang W H, Li H Y, Xie E Z, Sima C H, Lu T, Qiao Y and
Dai J F. 2022. BEVFormer: learning bird's-eye-view representa-
tion from multi-camera images via spatiotemporal Transformers//
Proceedings of the 17th European Conference on Computer Vision.
Tel Aviv, Israel: Springer: 1-18 [DOI: 10.1007/978-3-031-20077-
9_1]
- Liang M, Yang B, Chen Y, Hu R and Urtasun R. 2019. Multi-task
multi-sensor fusion for 3D object detection//Proceedings of 2019
IEEE/CVF Conference on Computer Vision and Pattern Recogni-
tion (CVPR). Long Beach, USA: IEEE: 7337-7345 [DOI: 10.
1109/CVPR.2019.00752]
- Liang M, Yang B, Wang S L and Urtasun R. 2018. Deep continuous
fusion for multi-sensor 3D object detection//Proceedings of the 15th
European Conference on Computer Vision. Munich, Germany:
Springer: 663-678 [DOI: 10.1007/978-3-030-01270-0_39]
- Liang Z D, Zhang M, Zhang Z H, Zhao X and Pu S L. 2020. RangeR-
CNN: towards fast and accurate 3D object detection with range
image representation [EB/OL]. [2023-12-21].
<https://arxiv.org/pdf/2009.00206.pdf>
- Liang Z D, Zhang Z H, Zhang M, Zhao X and Pu S L. 2021. RangeIoU-
Det: range image based real-time 3D object detector optimized by
intersection over union//Proceedings of 2021 IEEE/CVF Confer-
ence on Computer Vision and Pattern Recognition (CVPR). Nash-
ville, USA: IEEE: 7136-7145 [DOI: 10.1109/CVPR46437.2021.
00706]
- Lin T Y, Goyal P, Girshick R, He K M and Dollár P. 2017. Focal loss
for dense object detection//Proceedings of 2017 IEEE International
Conference on Computer Vision (ICCV). Venice, Italy: IEEE:
2999-3007 [DOI: 10.1109/ICCV.2017.324]
- Liu L J, Lu J W, Xu C J, Tian Q and Zhou J. 2019. Deep fitting degree
scoring network for monocular 3D object detection//Proceedings of
2019 IEEE/CVF Conference on Computer Vision and Pattern Rec-
ognition (CVPR). Long Beach, USA: IEEE: 1057-1066 [DOI:
10.1109/CVPR.2019.00115]
- Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu C Y and Berg A
C. 2016. SSD: single shot MultiBox detector//Proceedings of the
14th European Conference on Computer Vision. Amsterdam, the
Netherlands: Springer: 21-37 [DOI: 10.1007/978-3-319-46448-
0_2]
- Liu X, Li H, Cheng Y Z, Kong X Z and Chen S M. 2024. 3D multi-
object tracking based on image and point cloud multi-information
perception association. *Journal of Image and Graphics*, 29 (1) :
163-178 (刘祥, 李辉, 程远志, 孔祥振, 陈双敏. 2024. 图像与
点云多重信息感知关联的三维多目标跟踪. *中国图象图形学
报*, 29(1): 163-178) [DOI: 10.11834/jig.221003]
- Liu Z C, Wu Z Z and Tóth R. 2020. SMOKE: single-stage monocular
3D object detection via keypoint estimation//Proceedings of 2020
IEEE/CVF Conference on Computer Vision and Pattern Recogni-
tion Workshops (CVPRW). Seattle, USA: IEEE: 4289-4298
[DOI: 10.1109/CVPRW50498.2020.00506]
- Liu Z J, Tang H T, Amini A, Yang X Y, Mao H Z, Rus D L and Han
S. 2023. BEVFusion: multi-task multi-sensor fusion with unified
bird's-eye view representation//Proceedings of 2023 IEEE Interna-
tional Conference on Robotics and Automation (ICRA). London,
UK: IEEE: 2774-2781 [DOI: 10.1109/ICRA48891.2023.
10160968]
- Lu H H, Chen X S, Zhang G Y, Zhou Q H, Ma Y B and Zhao Y. 2019.
Scanet: spatial-channel attention network for 3D object detection//
Proceedings of ICASSP 2019—2019 IEEE International Confer-
ence on Acoustics, Speech and Signal Processing (ICASSP).
Brighton, UK: IEEE: 1992-1996 [DOI: 10.1109/ICASSP.2019.
8682746]
- Lu Y, Ma X Z, Yang L, Zhang T Z, Liu Y T, Chu Q, Yan J J and Ouy-
ang W L. 2021. Geometry uncertainty projection network for mon-
ocular 3D object detection//Proceedings of 2021 IEEE/CVF Interna-
tional Conference on Computer Vision (ICCV). Montreal, Canada:
IEEE: 3091-3101 [DOI: 10.1109/ICCV48922.2021.00310]
- Luo S J, Dai H, Shao L and Ding Y. 2021. M3DSSD: monocular 3D
single stage object detector//Proceedings of 2021 IEEE/CVF Confer-
ence on Computer Vision and Pattern Recognition (CVPR). Nash-
ville, USA: IEEE: 6141-6150 [DOI: 10.1109/CVPR46437.2021.
00608]
- Ma T, Yang X M, Zhou H B, Li X, Shi B T, Liu J J, Yang Y C, Liu Z
Z, He L, Qiao Y, Li Y K and Li H S. 2023. DetZero: rethinking
offboard 3D object detection with long-term sequential point
clouds//Proceedings of 2023 IEEE/CVF International Conference
on Computer Vision (ICCV). Paris, France: IEEE: 6713-6724
[DOI: 10.1109/ICCV51070.2023.00620]
- Ma X Z, Wang Z H, Li H J, Zhang P B, Ouyang W L and Fan X. 2019.
Accurate monocular 3D object detection via color-embedded 3D
reconstruction for autonomous driving//Proceedings of 2019 IEEE/
CVF International Conference on Computer Vision (ICCV). Seoul,
Korea (South) : IEEE: 6850-6859 [DOI: 10.1109/ICCV.2019.
00695]
- Manhardt F, Kehl W and Gaidon A. 2019. ROI-10D: monocular lifting
of 2D detection to 6D pose and metric shape//Proceedings of 2019
IEEE/CVF Conference on Computer Vision and Pattern Recogni-
tion (CVPR). Long Beach, USA: IEEE: 2064-2073 [DOI: 10.

- 1109/CVPR.2019.00217]
- Mao J G, Shi S S, Wang X G and Li H S. 2022. 3D object detection for autonomous driving: a comprehensive survey [EB/OL]. [2023-12-21]. <https://arxiv.org/pdf/2206.09474.pdf>
- Meyer G P, Charland J, Pandey S, Laddha A, Gautam S, Vallespi-Gonzalez C and Wellington C K. 2021. LaserFlow: efficient and probabilistic object detection and motion forecasting. *IEEE Robotics and Automation Letters*, 6(2): 526-533 [DOI: 10.1109/LRA.2020.3047793]
- Meyer G P, Laddha A, Kee E, Vallespi-Gonzalez C and Wellington C K. 2019. LaserNet: an efficient probabilistic 3D object detector for autonomous driving//*Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Long Beach, USA: IEEE: 12669-12678 [DOI: 10.1109/CVPR.2019.01296]
- Naiden A, Paunescu V, Kim G, Jeon B M and Leordeanu M. 2019. Shift R-CNN: deep monocular 3d object detection with closed-form geometric constraints//*Proceedings of 2019 IEEE international conference on image processing (ICIP)*. Taipei, China: IEEE: 61-65 [DOI: 10.1109/ICIP.2019.8803397]
- Ng P C and Henikoff S. 2003. SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Research*, 31(13): 3812-3814 [DOI: 10.1093/nar/gkg509]
- Ngiam J, Caine B, Han W, Yang B, Chai Y N, Sun P, Zhou Y, Yi X, Alsharif O, Nguyen P, Chen Z F, Shlens J and Vasudevan V. 2019. StarNet: targeted computation for object detection in point clouds [EB/OL]. [2023-12-21]. <https://arxiv.org/pdf/1908.11069.pdf>
- Pan X R, Xia Z F, Song S J, Li L E and Huang G. 2021. 3D object detection with pointformer//*Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Nashville, USA: IEEE: 7459-7468 [DOI: 10.1109/CVPR46437.2021.00738]
- Qi C R, Liu W, Wu C X, Su H and Guibas L J. 2018. Frustum PointNets for 3D object detection from RGB-D data//*Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Salt Lake City, USA: IEEE: 918-927 [DOI: 10.1109/CVPR.2018.00102]
- Qi C R, Yi L, Su H and Guibas L J. 2017. Pointnet++: deep hierarchical feature learning on point sets in a metric space//*Proceedings of the 31st International Conference on Neural Information Processing Systems*. Long Beach, USA: Curran Associates Inc.: 5105-5114
- Qin Z Y, Wang J L and Lu Y. 2022. MonoGRNet: a general framework for monocular 3D object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9): 5170-5184 [DOI: 10.1109/TPAMI.2021.3074363]
- Ren S Q, He K M, Girshick R and Sun J. 2017. Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6): 1137-1149 [DOI: 10.1109/TPAMI.2016.2577031]
- Roddick T, Kendall A and Cipolla R. 2019. Orthographic feature transform for monocular 3D object detection//*Proceedings of the 30th British Machine Vision Conference 2019*. Cardiff, UK: BMVC: #285
- Shi S S, Guo C X, Jiang L, Wang Z, Shi J P, Wang X G and Li H S. 2020a. PV-RCNN: point-voxel feature set abstraction for 3D object detection//*Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Seattle, USA: IEEE: 10526-10535 [DOI: 10.1109/CVPR42600.2020.01054]
- Shi S S, Wang X G and Li H S. 2019. PointRCNN: 3D object proposal generation and detection from point cloud//*Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Long Beach, USA: IEEE: 770-779 [DOI: 10.1109/CVPR.2019.00086]
- Shi W J and Rajkumar R. 2020. Point-GNN: graph neural network for 3D object detection in a point cloud//*Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Seattle, USA: IEEE: 1708-1716 [DOI: 10.1109/CVPR42600.2020.00178]
- Shi X P, Ye Q, Chen X Z, Chen C R, Chen Z X and Kim T K. 2021. Geometry-based distance decomposition for monocular 3D object detection//*Proceedings of 2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. Montreal, Canada: IEEE: 15152-15161 [DOI: 10.1109/ICCV48922.2021.01489]
- Shin K, Kwon Y P and Tomizuka M. 2019. RoarNet: a robust 3D object detection based on RegiOn approximation refinement//*Proceedings of 2019 IEEE Intelligent Vehicles Symposium (IV)*. Paris, France: IEEE: 2510-2515 [DOI: 10.1109/IVS.2019.8813895]
- Simonelli A, Buló S R, Porzi L, Ricci E and Kotschieder P. 2020. Towards generalization across depth for monocular 3D object detection//*Proceedings of the 16th European Conference on Computer Vision*. Glasgow, UK: Springer: 767-782 [DOI: 10.1007/978-3-030-58542-6_46]
- Sindagi V A, Zhou Y and Tuzel O. 2019. MVX-Net: multimodal Voxel-Net for 3D object detection//*Proceedings of 2019 International Conference on Robotics and Automation (ICRA)*. Montreal, Canada: IEEE: 7276-7282 [DOI: 10.1109/ICRA.2019.8794195]
- Sun P, Kretschmar H, Dotiwala X, Chouard A, Patnaik V, Tsui P, Guo J, Zhou Y, Chai Y N, Caine B, Vasudevan V, Han W, Ngiam J, Zhao H, Timofeev A, Ettinger S, Krivokon M, Gao A, Joshi A, Zhang Y, Shlens J, Chen Z F and Anguelov D. 2020. Scalability in perception for autonomous driving: Waymo open dataset//*Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Seattle, USA: IEEE: 2443-2451 [DOI: 10.1109/CVPR42600.2020.00252]
- Sun P, Wang W Y, Chai Y N, Elsayed G, Bewley A, Zhang X, Sminchisescu C and Anguelov D. 2021. RSN: range sparse net for efficient, accurate LiDAR 3D object detection//*Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*

- tion (CVPR). Nashville, USA: IEEE: 5721-5730 [DOI: 10.1109/CVPR46437.2021.00567]
- Tan M X, Pang R M and Le Q V. 2020. EfficientDet: scalable and efficient object detection//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE: 10778-10787 [DOI: 10.1109/CVPR42600.2020.01079]
- Tian H, Chen Y, Dai J, Zhang Z and Zhu X. 2021. Unsupervised object detection with LiDAR clues//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE: 5958-5968 [DOI: 10.1109/CVPR46437.2021.00590]
- Tian Z, Shen C H, Chen H and He T. 2019. FCOS: fully convolutional one-stage object detection//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea (South): IEEE: 9626-9635 [DOI: 10.1109/ICCV.2019.00972]
- Viola P and Jones M. 2001. Rapid object detection using a boosted cascade of simple features//Proceedings of 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001. Kauai, USA: IEEE: I-I [DOI: 10.1109/CVPR.2001.990517]
- Vora S, Lang A H, Helou B and Beijbom O. 2020. PointPainting: sequential fusion for 3D object detection//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE: 4603-4611 [DOI: 10.1109/CVPR42600.2020.00466]
- Wang C H, Chen H W and Fu L C. 2021a. VPFNet: voxel-pixel fusion network for multi-class 3d object detection [EB/OL]. [2023-12-21]. <https://arxiv.org/pdf/2111.00966.pdf>
- Wang C W, Ma C, Zhu M and Yang X K. 2021b. PointAugmenting: cross-modal augmentation for 3D object detection//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE: 11789-11798 [DOI: 10.1109/CVPR46437.2021.01162]
- Wang D Z and Posner I. 2015. Voting for voting in online point cloud object detection//Robotics: Science and Systems [DOI: 10.15607/RSS.2015.XI.035]
- Wang G J, Tian B, Zhang Y C, Chen L, Gao D P and Wu J. 2020. Multi-view adaptive fusion network for 3D object detection [EB/OL]. [2023-12-21]. <https://arxiv.org/pdf/2011.00652.pdf>
- Wang L, Du L, Ye X Q, Fu Y W, Guo G D, Xue X Y, Feng J F and Zhang L. 2021c. Depth-conditioned dynamic message propagation for monocular 3D object detection//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE: 454-463 [DOI: 10.1109/CVPR46437.2021.00052]
- Wang L, Zhang X Y, Song Z Y, Bi J F, Zhang G X, Wei H Y, Tang L Y, Yang L, Li J, Jia C Y and Zhao L J. 2023. Multi-modal 3D object detection in autonomous driving: a survey and taxonomy. IEEE Transactions on Intelligent Vehicles, 8 (7) : 3781-3798 [DOI: 10.1109/TIV.2023.3264658]
- Wang T, Zhu X E, Pang J M and Lin D H. 2021d. FCOS3D: fully convolutional one-stage monocular 3D object detection. Proceedings of 2021 IEEE/CVF International Conference on Computer Vision Workshops. Montreal, Canada: IEEE: 913-922 [DOI: 10.1109/ICCVW54120.2021.00107]
- Wang Y, Chao W L, Garg D, Hariharan B, Campbell M and Weinberger K Q. 2019a. Pseudo-LiDAR from visual depth estimation: bridging the gap in 3D object detection for autonomous driving//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA: IEEE: 8437-8445 [DOI: 10.1109/CVPR.2019.00864]
- Wang Z X and Jia K. 2019. Frustum ConvNet: sliding frustums to aggregate local point-wise features for Amodal 3D object detection//Proceedings of 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Macau, China: IEEE: 1742-1749 [DOI: 10.1109/IROS40897.2019.8968513]
- Wu H, Deng J H, Wen C L, Li X, Wang C and Li J. 2022a. CasA: a cascade attention network for 3-D object detection from LiDAR point cloudS. IEEE Transactions on Geoscience and Remote Sensing, 60: #5704511 [DOI: 10.1109/TGRS.2022.3203163]
- Wu H, Wen C L, Li W, Li X, Yang R G and Wang C. 2023a. Transformation-equivariant 3D object detection for autonomous driving//Proceedings of the 37th AAAI Conference on Artificial Intelligence. Washington, USA: AAAI: 2795-2802 [DOI: 10.1609/aaai.v37i3.25380]
- Wu H, Wen C L, Shi S S, Li X and Wang C. 2023b. Virtual sparse convolution for multimodal 3D object detection//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE: 21653-21662 [DOI: 10.1109/CVPR52729.2023.02074]
- Wu X P, Peng L, Yang H H, Xie L, Huang C X, Deng C Q, Liu H F and Cai D. 2022b. Sparse fuse dense: towards high quality 3D detection with depth completion//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE: 5408-5417 [DOI: 10.1109/CVPR52688.2022.00534]
- Xiao W P, Peng Y, Liu C, Gao J T, Wu Y Q and Li X M. 2023. Balanced sample assignment and objective for single-model multi-class 3D object detection. IEEE Transactions on Circuits and Systems for Video Technology, 33 (9) : 5036-5048 [DOI: 10.1109/TCSVT.2023.3248656]
- Xie E Z, Yu Z D, Zhou D Q, Philion J, Anandkumar A, Fidler S, Luo P and Alvarez J M. 2022. M2BEV: multi-camera joint 3D detection and segmentation with unified birds-eye view representation [EB/OL]. [2023-12-21]. <https://arxiv.org/pdf/2204.05088.pdf>
- Xie L, Xiang C, Yu Z X, Xu G D, Yang Z, Cai D and He X F. 2020. PI-RCNN: an efficient multi-sensor 3D object detector with point-based attentive cont-conv fusion module//Proceedings of 2020 AAAI Conference on Artificial Intelligence. New York, USA:

- AAAI; 12460-12467 [DOI: 10.1609/aaai.v34i07.6933]
- Xu B and Chen Z Z. 2018. Multi-level fusion based 3D object detection from monocular images//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE; 2345-2353 [DOI: 10.1109/CVPR.2018.00249]
- Xu D F, Anguelov D and Jain A. 2018. PointFusion: deep sensor fusion for 3D bounding box estimation//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE; 244-253 [DOI: 10.1109/CVPR.2018.00033]
- Xu H T, Xu J M and Xu W W. 2019. Survey of 3D modeling using depth cameras. *Virtual Reality and Intelligent Hardware*, 1(5): 483-499 [DOI: 10.1016/j.vrih.2019.09.003]
- Yan J J, Liu Y F, Sun J J, Jia F, Li S L, Wang T C and Zhang X Y. 2023. Cross modal Transformer: towards fast and robust 3D object detection [EB/OL]. [2023-12-21].
<https://arxiv.org/pdf/2301.01283.pdf>
- Yan Y, Mao Y X and Li B. 2018. SECOND: sparsely embedded convolutional detection. *Sensors (Basel)*, 18(10): #3337 [DOI: 10.3390/s18103337]
- Yang B, Luo W J and Urtasun R. 2018b. PIXOR: real-time 3D object detection from point clouds//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE; 7652-7660 [DOI: 10.1109/CVPR.2018.00798]
- Yang C Y, Chen Y T, Tian H, Tao C X, Zhu X Z, Zhang Z X, Huang G, Li H Y, Qiao Y, Lu L W, Zhou J and Dai J F. 2022. BEVFormer v2: adapting modern image backbones to bird's-eye-view recognition via perspective supervision//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE; 17830-17839 [DOI: 10.1109/CVPR52729.2023.01710]
- Yang Z T, Sun Y N, Liu S and Jia J Y. 2020. 3DSSD: point-based 3D single stage object detector//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE; 11037-11045 [DOI: 10.1109/CVPR42600.2020.01105]
- Yang Z T, Sun Y N, Liu S, Shen X Y and Jia J Y. 2018c. IPOD: intensive point-based object detector for point cloud [EB/OL]. [2018-12-13]. <https://arxiv.org/pdf/1812.05276.pdf>
- Yang Z T, Sun Y N, Liu S, Shen X Y and Jia J Y. 2019. STD: sparse-to-dense 3D object detector for point cloud//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea (South): IEEE; 1951-1960 [DOI: 10.1109/ICCV.2019.00204]
- Yin T W, Zhou X Y and Krähenbühl P. 2021. Center-based 3D object detection and tracking//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE; 11779-11788 [DOI: 10.1109/CVPR46437.2021.01161]
- Yoo J H, Kim Y, Kim J and Choi J W. 2020. 3D-CVF: generating joint camera and LiDAR features using cross-view spatial feature fusion for 3D object detection//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer; 720-736 [DOI: 10.1007/978-3-030-58583-9_43]
- Zeng Y M, Hu Y, Liu S C, Ye J, Han Y H, Li X W and Sun N H. 2018. RT3D: real-time 3-D vehicle detection in LiDAR point cloud for autonomous driving. *IEEE Robotics and Automation Letters*, 3(4): 3434-3440 [DOI: 10.1109/LRA.2018.2852843]
- Zhao X, Liu Z, Hu R L and Huang K Q. 2019. 3D object detection using scale invariant and feature reweighting networks//Proceedings of the 33rd AAAI Conference on Artificial Intelligence. Honolulu, USA: AAAI; 9267-9274 [DOI: 10.1609/aaai.v33i01.33019267]
- Zhou X Y, Zhuo J C and Krähenbühl P. 2019. Bottom-up object detection by grouping extreme and center points//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA: IEEE; 850-859 [DOI: 10.1109/CVPR.2019.00094]
- Zhou Y and Tuzel O. 2018. VoxelNet: end-to-end learning for point cloud based 3D object detection//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE; 4490-4499 [DOI: 10.1109/CVPR.2018.00472]
- Zhou Y S, He Y, Zhu H Z, Wang C, Li H Y and Jiang Q H. 2021. Monocular 3D object detection: an extrinsic parameter free approach//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE; 7552-7562 [DOI: 10.1109/CVPR46437.2021.00747]
- Zhu H Q, Deng J J, Zhang Y, Ji J M, Mao Q Y, Li H Q and Zhang Y Y. 2023. VPFNet: improving 3D object detection with virtual point based LiDAR and stereo data fusion. *IEEE Transactions on Multimedia*, 25: 5291-5304 [DOI: 10.1109/TMM.2022.3189778]

作者简介

贾明达,男,硕士研究生,主要研究方向为计算机视觉和三维目标检测。E-mail: jiamingda2024@ia.ac.cn

孟维亮,通信作者,男,副研究员,主要研究方向为计算机图形图像、计算机视觉、点云处理、深度学习、人工智能和无人驾驶环境感知。E-mail: weiliang.meng@ia.ac.cn

杨金明,男,硕士研究生,主要研究方向为场景定位和环境感知。E-mail: yangjinming23@mails.ucas.ac.cn

郭建伟,男,副研究员,主要研究方向为计算机图形学、3D视觉、几何深度学习和虚拟现实。

E-mail: jianwei.guo@nlpr.ia.ac.cn

张吉光,男,副研究员,主要研究方向为计算机视觉和三维场景建模。E-mail: jiguang.zhang@ia.ac.cn

张晓鹏,男,研究员,主要研究方向为计算机图形学、计算机视觉、数字几何处理与模式识别。

E-mail: xiaopeng.zhang@ia.ac.cn