中图法分类号:TP391.41　　**文献标识码**:A　　**文章编号**:1006-8961(2023)06-1709-32

论文引用格式:Li X Y, Ye Z H, Wei S K, Chen Z, Chen X T, Tian Y H, Dang J W, Fu S J and Zhao Y. 2023. 3D object detection for autonomous driving from image:a survey——benchmarks,constraints and error analysis. Journal of Image and Graphics, 28(06):1709-1740(李熙莹, 叶芝桧, 韦世奎, 陈泽, 陈小彤, 田永鸿, 党建武, 付树军, 赵耀. 2023. 基于图像的自动驾驶3D目标检测综述——基准、制约因素和误差分析. 中国图象图形学报, 28(06):1709-1740)[DOI:10.11834/jig.230036]

基于图像的自动驾驶3D目标检测综述
——基准、制约因素和误差分析

李熙莹[1,2,3],叶芝桧[1,2,3],韦世奎[4*],陈泽[1,2,3],陈小彤[4],田永鸿[5],
党建武[6],付树军[7],赵耀[4]

1. 中山大学智能工程学院,深圳　518107; 2. 中山大学·深圳,深圳　518107; 3. 广东省智能交通系统(ITS)重点实验室,深圳　518107; 4. 北京交通大学信息科学研究所,北京　100044; 5. 北京大学信息科学技术学院,北京　100871; 6. 兰州交通大学电子与信息工程学院,兰州　730070; 7. 山东大学数学学院,济南　250100

摘　要:从高分辨率图像中获取周边目标的精准3D位置和尺寸信息是实现自动驾驶控制和行为决策的基础,因此基于图像的3D目标检测是自动驾驶领域中的研究热点。已有学者对该领域方法论及成果进行了比较详细的综述,但对于导致现有方法检测精度不尽如意的制约因素未能进行深入系统的分析。考虑自动驾驶领域在工程应用方面的要求高,且现有方法以数据驱动类型为主,本文从常用数据集和评价基准、数据影响、方法论的制约因素和误差等角度,对学术界和产业界在3D目标检测方面的研究成果及行业应用进行较为系统的阐述。首先,从学术界探索成果以及自动驾驶行业的应用角度进行概要介绍。然后,从数据采集设备、数据精度和标注信息3方面详细分析总结了KITTI等4个通用数据集,并对这些数据集提出的主要评价指标进行对比分析。接着,从数据和方法论方面分析制约算法性能的主要因素及由此造成的误差影响。在数据方面,制约因素主要是数据精度、样本差异、标注数据量和标注规范;在方法论方面,制约因素主要包括先验几何关系、深度预测误差和数据模态等。最后,对国内外研究现状进行总结,并在数据集、评价指标和目标深度预测等方面提出了未来需要重点关注的研究方向。

关键词:3D目标检测;基准;制约因素;误差分析;自动驾驶;图像处理;计算机视觉

3D object detection for autonomous driving from image:a survey
——benchmarks,constraints and error analysis

Li Xiying[1,2,3], Ye Zhihui[1,2,3], Wei Shikui[4*], Chen Ze[1,2,3], Chen Xiaotong[4], Tian Yonghong[5], Dang Jianwu[6], Fu Shujun[7], Zhao Yao[4]

1. School of Intelligent Systems Engineering, Sun Yat-sen University, Shenzhen　518107, China; 2. Shenzhen Campus of Sun Yat-sen University, Shenzhen　518107, China; 3. Guangdong Province Key Laboratory of Intelligent Transportation System (ITS), Shenzhen　518107, China; 4. Institute of Information Science, Beijing Jiaotong University, Beijing　100044, China; 5. School of Electronics Engineering and Computer Science, Peking University, Beijing　100871, China; 6. School of Electronic and Information Engineering, Lanzhou Jiaotong University, Lanzhou　730070, China; 7. School of Mathematics, Shandong University, Jinan　250100, China

收稿日期:2023-01-16;**修回日期**:2023-03-07;**预印本日期**:2023-03-14

*通信作者:韦世奎　shkwei@bjtu.edu.cn

基金项目:国家自然科学基金项目(U21B2090,61972022);深圳市科技计划项目(JSGG20210802153412036)

Supported by: National Natural Science Foundation of China (U21B2090, 61972022); Science and Technology Program of Shenzhen (JSGG20210802153412036)

Abstract: Autonomous driving-oriented accurate perception and measurement of the three-dimensional (3D) spatial position and scale can be as the basis for realizing the control ability and decision-making level. Sensing technology-driven autonomous vehicles are equipped with high-resolution camera, light detection and ranging(LiDAR), radar, global positioning system(GPS)/inertial measurement unit(IMU) and other related sensors. Current LiDAR or multi-modal data-based 3D object detection algorithms are challenged for its deployment and application because of the shortcomings of LiDAR sensors like high price, limited sensing range, and sparse point clouds data. In contrast, such high-resolution cameras are commonly-used and featured by its lower price, and it can obtain high-resolution spatial information, richer shape, and appearance details as well. The emerging image-based 3D object detection is focused on further. At present, constraints of detection accuracy of the existing methods are still to be analyzed thoroughly and systematically. We summary the research results and industrial applications in relevance to such 1) perspectives of commonly used datasets and evaluation criteria, 2) data impact, 3) methodological constraints and prediction errors. First, a brief introduction is linked to perspective of academic domain and application of autonomous driving industry. We briefly review latest growths of Baidu Apollo, Google Waymo, Tesla and other related autonomous driving companies, and the thread of 3D object detection methods for autonomous driving. Then, we analyzed and summarized four popular datasets like KITTI, nuScenes, Waymo open dataset, and DAIR-V2X dataset from three aspects of: 1) data acquisition/sensors, data accuracy and data label information; 2) key evaluation standards proposed by these data sets, and 3) pros/cons and applicability of these evaluation standards. Third, main constraints of the image-based 3D object detection algorithm and the errors are derived from two sides of: data and methodology. Such main data constraints are originated from their data accuracy, sample difference, data volume, and data annotation. The data accuracy is mainly limited by equipment performance. The sample difference is mainly restricted by such image processing problems in related to object distance difference, angle difference, occlusion, and truncation. Data volume is affected by variety of 3D data types and high difficulty of labeling. The volume of 3D object detection data set is much smaller in comparison with the 2D object detection data set. Data annotation is mainly focused on 3D bounding box labeling, the labeling details, and quality of the dataset, especially for image annotation used in image-based 3D object detection. For non-rigid objects like pedestrians, the annotation error is larger, and there are some optimal for improving the labeling method. The general framework of image-based 3D object detection can be classified as one-stage methods and two-stage methods, and the limitations consists of 1) the prior geometric relationship, 2) depth prediction accuracy, and 3) data modality. The prior geometric relationship is focused on 2D-3D geometric constraints for 2D images-projected 3D objects and objects-between position relationships. The image-based 3D object detection methods face such problems as: prior 2D-3D geometric constraints and occluded and truncated objects. The prediction of depth information from 2D images is an ill conditioned problem, and dimension collapse will cause depth prediction error-relevant loss of depth information in the image. On the one hand, the depth prediction is often not accurate due to the influence of projection relationship. On the other hand, the performance of continuous depth prediction is often poor at the depth mutation of the image (such as edge of objects). When the prediction depth is discretized, there is a problem that the classification of depth is relatively rough, and the accuracy classification cannot be arbitrarily divided. The limitation of single image-based data modality is mainly reflected via large error of depth prediction. The detection performance of the algorithm can be optimized by 1) simulating the stereo signal and LiDAR point clouds, or 2) using stereo image as the aided input, or 3) leveraging point clouds data with accurate 3D information as supervision signal. In addition, video data can be adopted to improve the detection accuracy to a certain extent. Forth, current research situation is summarized and compared from academic and industrial domain. Finally, some future research directions are predicted in terms of such factors of datasets, evaluation indicators, and depth prediction.

Key words: 3D object detection; benchmark; constraint; error analysis; autonomous driving; image processing; computer vision

0 引 言

随着高分辨率相机、激光雷达和毫米波雷达等传感器的发展，自动驾驶汽车对环境的感知能力越来越强，高度自动驾驶甚至完全自动驾驶越来越成为可能。作为环境感知的核心技术之一，基于图像的2D目标检测技术（Redmon等，2016；Girshick等，2014；Girshick，2015；Ren等，2017；Liu等，2016；Zhou等，2019）发展相对成熟，但其缺少对目标物体在3维世界中位置、姿态和尺寸信息的准确估计，限制了自动驾驶汽车的感知精度和安全性。为此，3D目标检测逐渐成为工业界和学术界的研究热点，相关的数据集和评价基准（Geiger等，2012；Caesar等，2020；Sun等，2020）也不断建立和完善。3D目标检测是一种通过利用高分辨率相机、立体相机、激光雷达和毫米波雷达等传感器的一种或多种数据来预测目标3维属性信息的技术。通常，3维属性信息包括目标类别（c）、目标3维坐标（x,y,z）、3维尺寸（长l、宽w、高h）和姿态（俯仰角、滚动角、偏航角），可利用目标的最小外接立方体的信息来表示。由于汽车位于地平面且无翻转，只需要考虑汽车在地平面的偏航角（θ）。因此，自动驾驶的3D目标检测本质上是关于c,x,y,z,l,w,h,θ这8个参数的回归优化问题。当投影到鸟瞰图视角下进行3D目标检测时，该问题可以进一步简化为回归目标物体的类别和鸟瞰图视角下的3维信息（c,x,y,w,l,θ）（Novák，2017）。

在工业界，3D目标检测可分为两类技术路线，即基于多传感器融合的技术路线和基于纯视觉的技术路线。以百度、谷歌、小马智行、华为和滴滴等厂商为代表，采用基于多传感器融合的技术路线。百度自动驾驶平台Apollo（Baidu，2022）在其Apollo1.5版本中增加了激光雷达设备，构成了一个综合GPS/IMU（global positioning system/inertial measurement unit）、高分辨率相机、激光雷达和毫米波雷达的环境感知系统，使其自动驾驶车辆对周围环境有了更高的感知能力。Apollo2.0版本对Apollo1.5进行升级，使车辆能够在简单的城市道路上自主驾驶。更进一步，Apollo3.5版本已经能够实现车辆的360°感知。目前Apollo已经升级到Apollo7.0版本，通过搭载多传感器能够实现对复杂城市道路环境的感知与车辆行为决策。谷歌旗下Waymo采用以激光雷达为主的多传感器融合方案，感知系统包括激光雷达、毫米波雷达、高分辨率相机和补充传感器，能够实现全时段360°监控（池娟，2021）。小马智行（Pony.ai，2022）的第3代自动驾驶方案Pony Alpha在感知层采用了多传感器融合的方案，感知系统包括高分辨率相机、激光雷达和毫米波雷达等传感器，并在其自研的域控制器上实现了对相机、雷达的精准耦合，实现了全天候感知车身周围360°环境的能力。华为ADS（autonomous driving solution）高阶自动驾驶全栈解决方案同样采用多传感器融合方案，感知系统包括自研的激光雷达、毫米波雷达和高分辨率相机等传感器，不同的数据间采用后融合算法获得障碍物的分布信息（WorldAuto，2022）。滴滴出行的双子星自动驾驶系统同样采用多传感器融合的方案，感知系统包括远、中、近距的激光雷达、相机、雷达和红外相机等传感器，前向视角可实现12层传感器冗余覆盖叠加（孟醒，2021）。

以特斯拉为代表的厂商，采用纯视觉为主导的技术路线。特斯拉的完全自动驾驶（full self-drive，FSD）摒弃了昂贵的激光雷达等非视觉传感器，环绕车身构建了以可见光相机为主的感知系统，并通过多头神经网络HydraNet构建出真实世界的3维向量空间（Talpes等，2020）。安途AutoX在其第一代无人驾驶方案Gen1中采用纯视觉感知方案，虽然在后续几代版本中加入了激光雷达、毫米波雷达等传感器，但其始终坚持以相机为主导的多传感器感知方案（郭文佳，2021）。百度也推出了纯视觉城市道路自动驾驶闭环解决方案Apollo Lite，能够实现360°道路环境感知，为合作伙伴提供轻传感器、轻算力需求的轻量化产品解决方案，并基于该技术打造了智能领航辅助驾驶产品ANP（Apollo navigation pilot）（陈念航，2020）。丰田汽车旗下子公司Woven Planet在其辅助驾驶和更高级别的自动驾驶项目中，统一采用纯视觉方案开发自动驾驶（李琳，2022）。

在学术界，学者们提出了很多通用的3D目标检测理论和框架（Brazil和Liu，2019；Chen等，2020b；Wang等，2021b；Luo等，2021；Liu等，2020；Weng和Kitani，2019；You等，2020）。如图1所示，有学者从输入数据来源和方法论两个维度对目前的3D目标检测进行了层次化分类（Mao等，2022）。

目前，多线束激光雷达和雷达系统价格高昂，限制了其在车辆上的大规模部署与应用。与激光雷达

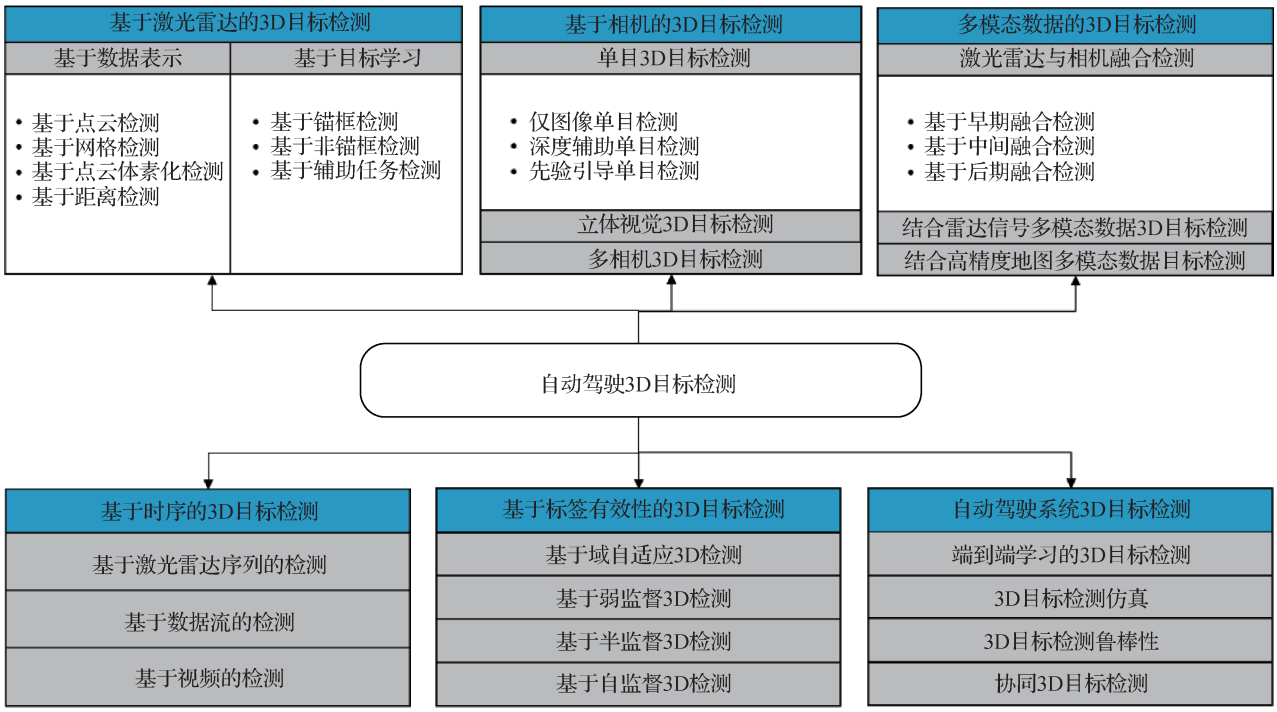


图1　用于自动驾驶3D目标检测的层次结构分类法(Mao等,2022)

Fig. 1　A hierarchical taxonomy for 3D object detection in autonomous driving (Mao et al., 2022)

和雷达系统相比,高分辨率相机、立体相机等视觉传感器具备经济实惠和便于车载的优势。因此,基于视觉图像来感知物体的3维信息逐渐成为研究的热点。由于物体通过相机投影到图像平面时会造成深度信息的丢失,所以从2维图像中获取3维信息是一个病态的问题,这导致基于图像的3D目标检测算法与基于激光雷达点云或多模态数据融合算法相比仍存在较大的性能差距。即便如此,许多学者和科技公司仍在此领域进行了深入研究,并取得了激励人心的结果。本文从数据集及评价基准、制约因素和误差分析3个方面来分析和总结基于图像的自动驾驶领域3D目标检测的已有成果,并对未来的发展方向进行展望,以期帮助从业者了解该领域的发展水平和未来发展方向。

本文主要贡献如下:1)对目前基于图像的自动驾驶领域3D目标检测算法常用的数据集和评价基准进行分析和总结,并指出需要完善之处。虽然已有多个公开数据集,并建立了诸多具有一定通用性的检测标准,但评价标准仍有改进空间。2)对制约基于图像的3D目标检测算法性能的主要因素进行总结,并对其误差进行分析。尽管现在已经有很多通用的理论和检测框架,但是当算法落地到应用时依然存在诸如深度预测不准确、目标类别之间的差异等问题。3)从多角度对基于图像的3D目标检测的发展方向进行展望。

1　数据集及评价基准

数据集及评价基准对于设计和评测基于图像的3D目标检测算法至关重要。本节从数据采集设备、通用数据集构成和算法评价指标等方面进行分析和总结。一般来说,基于图像的3D目标检测算法在训练和测试评价阶段只需要包含目标类别、属性信息以及目标外接立方体等标注信息的图像数据。然而,为了更准确地从图像中检测3D目标,一些算法在训练阶段引入2D标注信息、激光点云或深度图作为监督信号进行优化学习。因此,一些多传感器融合的数据集也可以用来训练和测试基于图像的3D目标检测算法。事实上,主流的数据集大多包含多个传感器的数据。

1.1　数据采集设备

由于图像采集设备的精度对感知能力影响较大,所以自动驾驶系统对车载传感器有较为严苛的要求。以车载相机为例,自动驾驶系统对相机数量、分辨率、视距范围、视场角、自身尺寸和复杂工况下的稳定性等有严格的规范。依据车规级约束,车载

相机需要在高低温、湿热、强微光和振动等复杂工况下保持工作稳定。

特斯拉的Autopilot系统从HW2.0开始，设计搭载了8个相机、12个远程超声波传感器和1个前置毫米波雷达，并在后续升级版本保持这一设计。在该系统中，后置相机采用了豪威科技OV10635型80万像素CMOS相机，其余相机则采用安森美半导体公司的120万像素CMOS相机，整个系统最远有效视距为250 m、单个相机的最大视场角120°。Autopilot HW2.0处理器搭载了Nvidia PG418 MXM模块，包含一个GP106 GPU和4 GB的GDDR5内存，可实现L4级别的自动驾驶（程增木，2022）。理想ONE使用了全世界第一个量产车载800万像素相机，同时搭载4颗200万像素环视相机、12颗超声波雷达、4颗毫米波雷达以及1颗前向毫米波雷达。其中，前置相机的水平视场达到120°，视距达200 m，并可以持续跟踪目标。安途（AutoX，2021）第5代系统采用了28个车规级800万像素相机，总像素达到2.2亿像素/帧。高分辨率的相机带来了对系统算力和算法的更高要求（刘岸泽，2022）。安途第5代系统采纳英特尔32核双CPU架构，主频率3.4 GHz。在车规级GPU方面，XCU域控制器算力更是达到2 200 TOPS（Tera operations per second）。表1列出了一些自动驾驶数据采集平台常用设备及性能。

表1 一些数据采集设备信息

Table 1 Information on data collection equipment

设备	型号	主要性能参数	应用厂商	备注
相机	Aptina AR0132/ AR0136A	1/3″, 1.2 MP (MegaPixels,百万像素), 1 280×960像素	Tesla	最远有效视距250 m，最大视场角120°
相机	豪威科技OV10635	0.8 MP	Tesla	最远有效视距约50 m，视场角140°
相机	大陆集团MFC535	8 MP	蔚来、理想、极氪等	视场角为125°
相机	OnSemi	8 MP	AutoX	
相机	Point Grey Flea 2 (FL2-14S3M-C)	1.4 MP	KITTI数据集	灰度相机
相机	Point Grey Flea 2 (FL2-14S3C-C)	1.4 MP	KITTI数据集	彩色相机
相机	Basler acA1600-60gc	1/1.8″, 1 600×1 200像素	nuScenes数据集	
激光雷达	禾赛AT128	128线，分辨率1 200×128像素	理想	测距范围≤200 m，视场角120°×25.4°
激光雷达	Velodyne HDL32E	32线旋转式激光雷达，20 Hz采样频率	nuScenes数据集	探测范围≤70 m，测距精度2 cm，采样率1.4 M pts/s
毫米波雷达	ARS-4B	60 GHz	Tesla	
毫米波雷达	博世LLR4	76～77 GHz	小鹏	最大探测范围≤250 m，最大探测目标24个
毫米波雷达	Continental ARS 408-21	77 GHz	nuScenes数据集	探测范围≤250 m，速度精度±0.1 km/h

1.2 通用数据集

1.2.1 KITTI 3D数据集

KITTI（Karlsruhe Institute of Technolgoy and Toyota Technological Institute at Chicago）数据集（Geiger等，2012）是自动驾驶领域最受欢迎的数据集之一，由装载了高分辨率的彩色相机和灰度相机、激光雷达等传感器的采集系统收集。根据标注信息不同，分为2D目标检测数据集和3D目标检测数据集（即KITTI 3D）等（Liao等，2023）。KITTI 3D数据集是最受欢迎的3D目标检测数据集之一，具体的数据精度如表2所示，其中，KITTI官网图像标称分辨率为1 382×512像素，实际下载图像分辨率不等，以

1 242 × 375像素最多。采集设备参数如表3所示。KITTI数据集的收集场景包括中等城市、农村地区和高速公路，其目标检测数据集包含7 481幅训练图像和7 518幅测试图像以及对应的点云数据，共有80 256个标记物体。KITTI数据集的目标类别包括小汽车、面包车、卡车、行人、坐着的人、骑行者和有轨电车共7类，常用于3D目标检测的类别为小汽车(car)、行人(pedestrian)和骑行者(cyclist)3类。KITTI数据集根据边界框(bounding box)高度、遮挡和截断程度指标将标注目标划分为简单、中等和困难3个难度级别。KITTI 3D数据集的标注信息如表4所示。图2给出了KITTI 3D数据集的采集设备坐标系定义及标注信息中目标的观测角Alpha(红色)和航向角Rotation_y(蓝色)的定义示意图。KITTI数据集的测试集不提供具体标注信息，算法性能测试需要上传到官方网页进行。为了在本地进行模型性能测试，有学者将KITTI 3D数据集的训练集划分为训练集和验证集，常用的划分标准是将数据集的7 481幅训练集图像划分为3 712幅训练图像和3 769幅验证图像(Chen等，2015)，或者划分为3 682幅训练图像和3 799幅验证图像(Xiang等，2017)。

表2　KITTI 3D数据集的数据精度

Table 2　Data accuracy of KITTI 3D dataset

采集设备	数据模态	数据精度
彩色相机(left)	图像	1 382 × 512像素
彩色相机(right)	图像	1 382 × 512像素
激光雷达	点云	1.3 Mpts/s

表3　KITTI数据集采集设备信息

Table 3　Sensors used to setup KITTI dataset

设备名称	数量	型号	参数
GPS/IMU 惯性导航系统	1	OXTS RT 3003	6轴，采集频率100 Hz；L1/L2 RTK；分辨率0.02 m/0.1°
灰度相机	2	Point Grey Flea 2 (FL2-14S3M-C)	分辨率1.4 MP；1/2″Sony ICX267CCD；全景镜头
彩色相机	2	Point Grey Flea2 (FL2-14S3C-C)	分辨率1.4 MP；1/2″Sony ICX267CCD；全景镜头
变焦镜头	4	Edmund Optics NT59-917	焦距4 ~ 8 mm；视场角最大为90°；感兴趣区域竖直视场角最大为35°
激光雷达	1	Velodyne HDL-64E	旋转3D激光扫描；扫描频率10 Hz；发射线束64线；角度分辨率0.09°；测距精度2 cm；采样率1.3 Mpts/s；水平视场角360°；竖直视场角26.8°；测距范围 ≤ 100 m

表4　KITTI 3D数据集标注信息

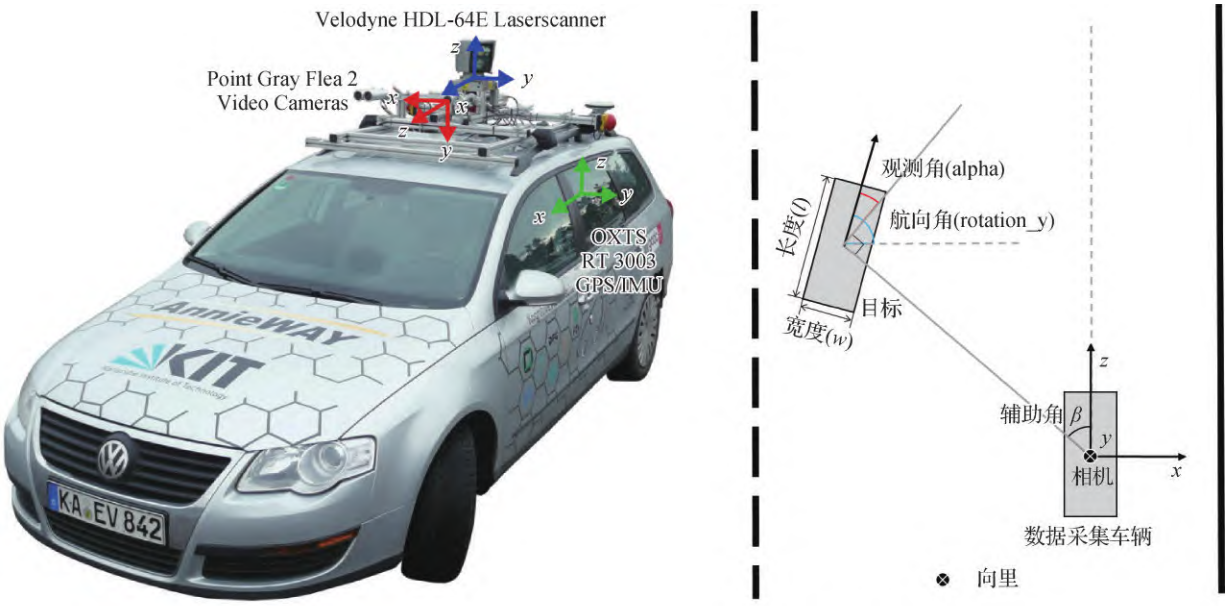
Table 4　Annotation information in KITTI 3D dataset

字段	字段长度	单位	含义
Type	1	-	目标类别
Truncated	1	-	目标截断程度：0(非截断) ~ 1(截断)之间的浮点数
Occluded	1	-	目标遮挡程度：整数0，1，2，3(0：完全可见，1：部分遮挡，2：大部分遮挡，3：未知)
Alpha	1	弧度	相机坐标系下，目标的观测角：[-π，π]
Bbox	4	像素	图像中目标的2D边界框位置，包括左上顶点和右下顶点的像素坐标
Dimension	3	m	目标的3D尺寸：高、宽、长
Location	3	m	相机坐标系下，目标的3D边界框中心坐标(x，y，z)
Rotation_y	1	弧度	相机坐标系下，目标的航向角：[-π，π]

注："-"表示没有单位。

1.2.2　nuScenes数据集

nuScenes数据集(nuTonomy scenes)(Caesar等，2020)发表于2019年，也是自动驾驶领域比较常用的数据集之一。数据集采集系统包含1个激光雷达、5个雷达、6个彩色相机和1个惯性导航系统，提供360°的扫描结果，主要的数据精度如表5所示，各采集设备参数如表6所示。nuScenes数据集包含1 000个驾驶场景，每个场景的采集时间为



(a) KITTI数据集的设备坐标系(Geiger等, 2022)　　(b) 标注信息中的观测角和航向角定义

图2　KITTI数据集坐标系、目标的观测角和航向角定义

Fig. 2　Definitions of coordinate system, alpha and rotation_y of KITTI dataset

((a) equipment coordinate system of KITTI dataset (Geiger et al., 2022); (b) definitions of alpha and rotation_y)

表5　nuScenes数据集数据精度

Table 5　Data accuracy of nuScenes dataset

采集设备	数据模态	数据精度
相机	图像	1 600 × 900像素
激光雷达	点云	1.4 Mpts/s

20 s，包含各种驾驶行为、交通状况和意外行为。数据集收集地点为美国波士顿的海港(Seaport)和南波士顿(South Boston)地区、新加坡的纬壹科技城(One North)、女王镇(Queenstown)和荷兰村(Holland Village)地区。这两个城市交通密集，驾驶环境复杂，车辆类型、植被、道路标志和交通规则(分属左右手交通)都各不相同。数据集收集场景涵盖不同的地点、不同的时间(白天和黑夜)和不同的天气条件(晴天、雨天和多云)，对研究算法的通用性很有帮助。数据集包含约140万帧图像、39万帧激光雷达扫描帧、140万帧雷达扫描帧，标注了4万个关键帧中的140万个目标边界框，标注数据量约为KITTI数据集的7倍。nuScenes数据集是以2 Hz采样率对图像、激光雷达和雷达数据抽取关键帧进行标注，采用了层次分类法，最底层共定义了小汽车(vehicle. car)、成年人(human. pedestrian. adult)等23个物体类。所有物体都标注有一个语义类别、属性(如可见性、活动和姿势)和3D边界框$(x, y, z, w, l, h, \theta)$。具体标注信息如表7所示。

表6　nuScenes数据集采集设备信息

Table 6　Sensors used to setup nuScenes dataset

设备名称	数量	参数
彩色相机	6	颜色通道RGB；采样频率12 Hz; 传感器1/1.8″ CMOS；分辨率1 600 × 900像素；自动曝光，JPEG压缩
激光雷达	1	旋转扫描，32线；采样频率20 Hz; 水平视场角360°; 竖直视场角−30° ~ 10°; 测距范围 ≤ 70 m; 测距精度±2 cm; 采样率1.4 Mpts/s
雷达	5	测距范围 ≤ 250 m, 77 GHz，FMCW，采样频率13 Hz；测速精度±0.1 km/h
GPS/IMU惯性导航	1	GPS，IMU，AHRS；偏航角精度0.2°；滚动/俯仰角精度0.1°；20 mm RTK定位；更新频率1 000 Hz

表7 nuScenes数据集3D标注信息

Table 7 3D annotation information in nuScenes dataset

字段	字段长度	单位	含义
Visibility	1	–	目标可见程度,分为0~40%,40%~60%,60%~80%,80%~100%,用1~4表示
Category_name	–	–	目标类别名称
Attribute	–	–	目标属性
Translation	3	m	全局坐标系下,目标的3D边界框中心:(x,y,z)
Size	3	m	目标的3D边界框尺寸:宽、长、高
Rotation	4	–	3D边界框的旋转矩阵,用四元组表示
Num_lidar_pts	1	个	目标的激光雷达点云数目
Num_radar_pts	1	个	目标的雷达点云数目

注:"–"表示没有统一规定或者单位。

1.2.3 Waymo Open Dataset数据集

Waymo Open Dataset(Sun等,2020)发布于2019年,由装载有5个激光雷达和5个高分辨率针孔相机的数据采集车采集得到,其中激光雷达和针孔相机均进行了同步和标定处理,采集数据精度在表8中给出,各采集设备参数如表9所示。数据集的采集地点为美国凤凰城(Phoenix)、山景城(Mountain View)和旧金山(San Francisco)3个城市,采集时间段包括白天、夜晚和黄昏。最初发布的数据集(Sun等,2020)包含1 150个场景,每个场景的采集时间为20 s,划分为训练集场景798个,验证集场景202个,测试集场景150个,后续一直有更新。数据集包含约1 200万个有激光雷达数据的3D边界框(对应11.3万个激光雷达跟踪ID)以及约1 200万个图像2D边界框(对应25.4万个图像跟踪ID)。标注类别包括小汽车、行人、交通标志和骑行者。对于激光雷达数据,每个目标有唯一的跟踪ID,采用7个自由度的3D边界框$(c_x,c_y,c_z,l,w,h,\theta)$标注,其中$c_x,c_y,c_z$代表3D边界框的中心坐标,$l,w,h$代表目标的尺寸长宽高,$\theta$表示目标的航向角。除了激光雷达标签,在所有图像中分别对小汽车、行人和骑行者进行目标的2D边界框(x,y,l,w)标注,分别代表2D边界框的中心坐标以及长宽,这与3D边界框在图像的2D投影保持一致。根据目标对应的激光雷达点云数量,数据集将目标划分为LEVEL_1和LEVEL_2两个级别。LEVEL_1级别的目标对应的点云数大于5个,LEVEL_2级别的目标对应的点云数小于等于5个。

表8 Waymo Open Dataset数据集数据精度

Table 8 Data accuracy of Waymo Open Dataset

采集设备	数据模态	数据精度
相机(front, front-left, front-right)	图像	1 920×1 280像素
相机(side-left, side-right)	图像	1 920×1 040像素
激光雷达	点云	1.77 Mpts/s

表9 Waymo Open Dataset数据集采集设备信息表

Table 9 Sensors used to setup Waymo Open Dataset

设备名称	数量	参数
激光雷达(top)	1	垂直视场角[−17.6°, +2.4°];测距范围≤75 m;回波模式双回波
激光雷达(front, right, side-lift, side-right)	4	垂直视场角[−90°, +30°];测距范围≤20 m;回波模式双回波
高分辨率针孔相机(front)	1	分辨率1 920×1 280像素;水平视场角±25.2°
高分辨率针孔相机(flont-left, front-right)	2	分辨率1 920×1 280像素;水平视场角±25.2°
高分辨率针孔相机(side-left, side-right)	2	分辨率1 920×1 040像素;水平视场角±25.2°

1.2.4 DAIR-V2X数据集

DAIR-V2X数据集(Yu等,2022)是目前全球首个用于车路协同自动驾驶研究的大规模、多模态及多视角数据集,由装载有激光雷达、高分辨率相机和GPS/IMU惯性导航系统等传感器的自动驾驶车辆端和路侧设备端采集的数据组成,主要数据精度在表10中列出,各采集设备的参数如表11所示。数据集覆盖了10 km的城市道路、10 km的高速公路、28个十字路口和38平方公里的不同天气与照明变化的驾驶区域。包含71 254帧点云数据和71 254帧图像数据,其中40%的数据从路侧端传感器采集,60%帧的数据从车端传感器采集,并在基准算法测验验证中按照5:2:3的比例将数据集划分成训练集、验证集和测试集。所有数据均进行了人工修正。标记的目标类别涵盖小汽车(car)、卡车(truck)、面包车(van)、公交车(bus)、行人(pedestrian)、自行车(cyclist)、三轮车(tricyclist)、摩托车(motorcyclist)、手推车(barrowlist)和交通锥桶(trafficcone),具体的数据集标注信息主要分为单侧标注(表12)和融合标注(表13)。

表10 DAIR-V2X数据集数据精度

Table 10 Data accuracy of DAIR-V2X dataset

采集设备	数据模态	数据精度
相机(车端)	图像	1 920×1 080像素
相机(路端)	图像	1 920×1 080像素
激光雷达(车端)(DAIR-V2X-C)	点云	0.72 Mpts/s
激光雷达(车端)(DAIR-V2X-V)	点云	2.4 Mpts/s

表11 DAIR-V2X数据集采集设备信息

Table 11 Sensors used to setup DAIR-V2X dataset

设备名称	数量	型号	参数
激光雷达(路端)	4	Jaguar Prime from Innovusion	300线;采样频率10 Hz;水平视场角100°;竖直视场角-30°~10°;测距范围≤280 m;测距精度±3 cm
相机(路端)	4	-	色彩空间RGB;采样频率25 Hz;分辨率1 920×1 080像素;压缩方式JPEG
激光雷达(车端)(DAIR-V2X-C)	1	Hesai Pandar 40P	40线;采样频率10 Hz;水平视场角360°;竖直视场角-30°~10°;测距范围≤200 m;竖直分辨率±0.33°
激光雷达(车端)(DAIR-V2X-V)	1	Velodyne128 LiDAR	128线;采样频率10 Hz;水平视场角360°;垂直视场角-30°~10°;测距范围≤245 m;测距精度≤3 cm
相机(车端)	1	-	颜色空间RGB;采样频率20 Hz;分辨率1 920×1 080像素;压缩方式JPEG
GPS/IMU惯性导航(车端)	1	-	更新频率1 000 Hz

注:"-"表示官方没有提供详细信息。

表12 DAIR-V2X数据集单侧标注信息

Table 12 One-sided annotation information in DAIR-V2X dataset

字段	字段长度	单位	含义
Type	-	-	目标类型(10种)
Truncated_state	1	-	障碍物截断从[0, 1, 2]中取值,分别表示不截断、横向截断、纵向截断
Occluded_state	1	-	障碍物遮挡从[0, 1, 2]中取值,0代表不遮挡、1代表0%~50%遮挡,2代表50%~100%遮挡
Alpha	1	弧度	目标的观测角度,范围[-π,π]
2D_box	4	像素	图像中目标的2D边界框,标注为左上顶点与右下顶点的像素坐标(x_{min},y_{min},x_{max},y_{max})
3D_dimension	3	m	3D边界框的尺寸高(h)、宽(w)、长(l)
3D _location	3	m	3D边界框中心点坐标(x,y,z),基于LiDAR坐标系
Rotation	1	弧度	表示障碍物绕Z轴旋转角度,范围[-π,π],基于LiDAR坐标系

注:"-"表示没有统一规定或者单位。

表13　DAIR-V2X数据集融合标注信息

Table 13　Fused annotation information in DAIR-V2X dataset

字段	类型	单位	含义
Type	字段	–	4类目标类型：小汽车(car)、卡车(truck)、面包车(van)、公交车(bus)
World_8_points	数组	m	障碍物3D边界框8个顶点的坐标(x,y,z)，基于世界坐标系
System_error_offset	数组	m	路端与车端相对标定存在的系统误差(*delta_x*,*delta_y*)，经过人工二次修正；未融合标注的为空("")

注："–"表示没有单位。

1.2.5　其他数据集

除上述4个常用的数据集外，还有能为3D目标检测提供额外信息的其他数据集及基准。如，ApolloCar3D(Song等，2019)数据集包含5 277幅驾驶图像和超过6万辆汽车样本，并且为每辆汽车标注了模型尺寸和配备行业级3D CAD(computer aided design)模型。该数据集还结合车辆的3维位姿和3维形状开发了一种新的3维度量。

1.3　评价基准 —— 常用评价指标

在3D目标检测领域，平均精度(average precision，AP)(Everingham等，2010)是考察算法性能的最主要评价指标。然而，根据预测结果与真值之间的匹配标准，不同的数据集关于平均精度的计算方式有所差别。本节详细介绍KITTI 3D数据集、nuScenes数据集以及Waymo Open Dataset数据集的3D目标检测算法评价基准。

1.3.1　KITTI 3D数据集评价基准

在KITTI 3D数据集中，对3D边界框匹配和方向回归分别设立了不同的评价指标。边界框匹配性能采用AP来衡量，而方向回归性能则采用平均方向相似性(average orientation similarity，AOS)来评估。在计算AP时，要求真正例(true positive，TP)的交并比(intersection-over-union，IoU)(Everingham等，2010)超过50%，其中对小汽车(car)类别要求正例的IoU超过70%。KITTI 3D数据集根据3D预测框和3D真值框之间的交并比计算IoU。

平均方向相似性(AOS)定义为

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

式中，$r = TP/TP + FN$是PASCAL(pattern analysis，statical modeling and computational learning)目标检测的召回率(recall)(Geiger等，2012)；$\tilde{r}$是大于等于r的召回率值；$s \in [0, 1]$是方向相似性，其定义为

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

式中，$\boldsymbol{D}(r)$表示在召回率为r时的集合，$\Delta_{\theta}^{(i)}$表示目标i的预测值与真值之间的角度差。为了惩罚多个检出匹配到同一个真值，如果检出目标i已经匹配到真值(IoU至少50%)则设定$\delta_i = 1$，否则$\delta_i = 0$。

1.3.2　nuScenes数据集评价基准

在nuScenes数据集中，定义目标检测是在t时刻进行，并使用$[t-0.5\ \text{s}, t]$时间内的传感器数据进行检测。nuScenes数据集的检测性能只使用10类目标进行计算。这10个类别是nuScenes数据集标注的23个类别的子集。

同样，nuScenes数据集也使用平均精度AP作为目标检测的主要评价指标。然而，与KITTI 3D不同，nuScenes数据集通过预测框和真值框在鸟瞰图投影的2D中心距离d来评价目标匹配度。这样做的目的是将目标定位指标独立出来，与尺寸回归和方向估计指标进行分离。通过计算召回率和准确率(precision)均超过10%的PR(precision-recall)曲线下的归一化面积来计算平均精度均值(mean average precision，mAP)。去除召回率或准确率低于10%的部分，能够将低精度和低召回率区间的噪声的影响降至最低。设定集合$\boldsymbol{D}$为$\{\boldsymbol{D} | d = 0.5\ \text{m}, 1\ \text{m}, 2\ \text{m}, 4\ \text{m}\}$，集合$\boldsymbol{C}$为$\{\boldsymbol{C} | d \leqslant \boldsymbol{D}\}$，即可计算最终的mAP指标。具体为

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

除了平均精度AP度量之外，该数据集中还提出多个TP误差度量指标，对满足$d \leqslant 2$ m的所有TP检测结果进行计算。这些指标对每个类别独立进行计算，首先对每个类计算各自的PR曲线，去掉召回率小于10%的部分，再求平均。如果某一个类别的召回率达不到10%，则该类别的所有TP误差度量指标

误差均为1。

TP误差度量指标主要有5个。1)平均平移误差(average translation error，ATE)，使用2维欧几里得距离进行定义；2)平均比例误差(average scale error，ASE)，在方向对齐和平移对齐后，采用预测框和真值框的鸟瞰图投影IoU来定义，即“1-IoU”；3)平均方位误差(average orientation error，AOE)，其定义为预测和真值之间的最小偏航角，单位为弧度，除了那些只在180°测量(对称)的障碍物以外，所有的角度都是在360°的范围内进行测量；4)平均速度误差(average velocity error，AVE)，其定义为绝对速度误差，使用的是2维平面的速度差的L_2范数；5)平均属性误差(average attribute error，AAE)，其定义为1减去分类精度。

对每一个TP误差度量指标，可以计算所有类别的平均TP误差度量指标(mean true positive，mTP)。具体为

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

式中，TP_c指类别为c的TP误差度量指标，$\boldsymbol{C}$是类别集合。如果一些指标在某些类别物体上没有明确定义，那么不进行计算。比如，对于一些静止的障碍物不计算其速度误差。

在nuScenes数据集基准中，将不同的误差类型合并成一个标量分数，即nuScenes检测分数(nuScenes detection score，NDS)，其定义为

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

尽管mAVE、mAOE和mATE的计算结果都可能大于1，但NDS将每一个度量都限制在[0，1]范围之间。NDS既考虑了目标检测性能，又结合了基于边界框的位置、大小、方向、属性和速度的量化检测质量。

1.3.3 Waymo Open Dataset数据集评价基准

与其他数据集相比，Waymo Open Dataset数据集关于目标检测的评价指标除了AP之外，还提出了一个新的度量APH，将航向信息合并到目标检测度量AP中，定义为

$$APH = 100\int_0^1 \max\left\{h(r') \mid r' \geqslant r\right\}\mathrm{d}r \tag{6}$$

式中，$h(r')$是一个类似PR曲线中的$p(r)$的参数，其融合了加权的航向信息。每个TP目标的航向信息取$\frac{\min\left(\left|\tilde{\theta}-\theta\right|, 2\pi - \left|\tilde{\theta}-\theta\right|\right)}{\pi}$，$\tilde{\theta}$和$\theta$分别为预测航向角和真值航向角，单位为弧度，且定义在$[-\pi, \pi]$之间。

1.3.4 评价指标比较

由上述定义可以看出，KITTI 3D数据集将3D边界框匹配与方向回归分离开来，设置了两个不同的评价指标。nuScenes数据集则认为应当分离每一个损失项并独立优化，将目标定位、目标尺寸回归和方向预测分别独立并设置了不同的评价指标。另外，nuScenes数据集认为通过IoU指标来评价算法的定位性能是不恰当的，特别是当鸟瞰投影为长条状矩形时。如图3所示，其中实线边界框为真值框，虚线框为预测框。当采取2维边界框中心距离作为定位性能指标时，预测框的定位与真值框完全匹配；当采取IoU作为定位性能指标时，预测框定位与真值框相差较大。Waymo Open Dataset数据集则认为不应将方向回归与目标定位和尺寸回归分开计算平均精度，应该考虑每个组成部分与最终结果的相关性，将它们合并到一起建立新的评价标准来评估3D目标检测。

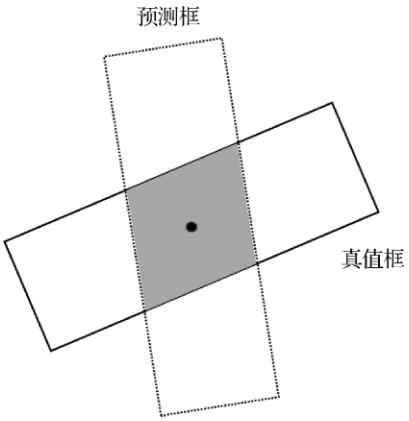


图3 基于交并比进行定位评价的缺点

Fig. 3 Disadvantage of IoU-based location evaluation

2 制约因素与误差分析

2.1 数据制约因素

数据是制约基于图像3D目标检测方法性能的一个重要因素，主要体现在训练数据的精度、样本差异、标注数据量和标注规范等方面。本节将分析和讨论这些因素带来的影响。

2.1.1　数据精度

数据精度涉及两个问题，即采集设备综合成像性能和可检测范围。受硬件设施的制约，车载相机的分辨率大多为百万级像素，远低于其他一些领域应用的高分辨率相机。由于目标成像尺寸与距离成反比，当目标成像面积小于一定值（如面积 < 50像素）时，目标检测准确率将大幅下降。对于车载相机，成像的范围（即视场角、可视距离和景深）取决于成像芯片尺寸、镜头焦距和光圈等综合设备参数。在固定分辨率和光圈的条件下，焦距越长，可视距离越远，视场角越小，能够采样的范围越小；焦距越短，可视距离越近，视场角越大，能够采样的范围越大。为了平衡视场和目标成像尺度的关系，进行全视角的感知，目前的数据采集平台使用全景摄像机或装载多个相机解决采样视场角的问题。

2.1.2　样本差异

图像是现实世界物体通过成像系统在传感器像平面上的投影。由于现实世界中物体间的几何关系和投影变换的限制，3维世界中的物体投影到图像平面时会导致诸多样本差异。其中，较为明显的样本差异有成像尺寸不一、成像角度变化、物体间遮挡、自遮挡以及截断（即部分区域在图像之外）。图4给出了一些示例（红框框出）。图4（a）（b）为同一车辆在不同距离上的视觉表现。图4（a）中车辆距离设备较近，成像尺寸较大；图4（b）中车辆距离设备较远，成像尺寸较小。图4（c）中靠近图像边界的车辆被图像截断；图4（d）中路边的车辆被其他车辆遮挡。为了体现这些差异，一些数据集添加了相应的属性标注。在KITTI数据集中，根据目标在图像上2D边界框的高所对应的最大像素值、遮挡程度以及截断程度，将数据集分为简单、中等以及困难3个子集。Waymo Open Dataset数据集中，则根据目标对应的激光雷达点云数目，划分为LEVEL_1和LEVEL_2两个级别。由于点云是由激光雷达扫描得到，同样大小的物体，在近处的点云数较多，而在远处的点云数会较少，因此可以认为这是按照物体与收集设备之间的距离进行简单划分。

在方法论上，很多学者针对样本间的差异提出了自己的见解和解决方案，这些将在2.2.2节进行详细的探讨。

(a) 近处的目标　(b) 远处的目标　(c) 被截断的目标　(d) 被遮挡的目标

图4　物体远近大小变化、截断和遮挡示例

Fig. 4　Examples of object distance and size change, truncation and occlusion ((a) close object; (b) distant object; (c) truncated object; (d) occluded object)

2.1.3　标注数据量

多方面原因造成已标注的3D数据集的数据量远少于已标注的2D数据集（如ImageNet等）。首先，采集3D数据集需要构建包含高分辨率相机、激光雷达、毫米波雷达和惯性导航系统等传感器的专用采集平台。相比于只需要一台摄像机的2D数据集采集系统，3D数据采集系统造价昂贵。其次，2D数据集常在设备固定的情况下进行数据采集，而用于自动驾驶的3D数据集则是在设备随车运动的情况下采集，成像质量容易受到运动、气候和光照等诸多因素影响。最后，相比于2D数据的简单标注（使用矩形框标注等），3D数据需要在3维空间中进行标注，标注信息更多，并且需要对不同传感器的数据进行时间和空间上的匹配校正，对标注人员的技能要求较高。因此3D数据集的数据采集和标注都十分困难，需要投入大量的人力物力。这也是目前制约3D目标检测发展的重要原因。表14展示了一些数据集的数据量对比情况，其中×n表示有n个相机的数据，2D数据集中的数据规模为ISLVRC2012比赛的数据。

2.1.4　标注规范

现有3D数据集的标注方式主要包含点云分割标注和3D边界框标注。点云分割主要针对的是点云数据，将扫描得到的点云数据归属于不同的物体，这种方法适用于利用点云数据进行目标检测（Aghdam等，2021；Xu等，2022；Ku等，2018；Chen等，2017；Dou等，2019；Chen等，2020a；Nabati和Qi，2021；Lu等，2019），但不适合基于图像的3D目标检测。由于点云数据能够提供精确的深度信息，现有部分算法尝试将点云数据作为辅助数据应用到基于

表 14　数据集规模对比

Table 14　Size comparison of datasets

数据集	类别	场景数	类别数目	数据类型及数量				数据规模				场景
				图像/幅	立体视觉	激光雷达/帧	雷达/帧	训练集/幅	验证集/幅	测试集/幅	边界框/个	夜晚/雨天
KITTI 3D	3D	–	8	15 K	有	15 K	0	7 418×1	–	7 518×1	200 K	无/无
nuScenes	3D	1 000	23	1.4 M	有	400 K	1.3 M	28 130×6	6 019×6	8 006×6	1.4 M	有/有
Waymo Open Dataset	3D	1 150	4	1 M	有	200 K	0	122 200×5	30 407×5	40 077×5	12 M	有/有
Argoverse(Chang等,2019)	3D	113	15	490 K	有	44 K	0	39 384×7	15 062×7	12 507×7	993 K	有/有
Lyft L5(Houston等,2020)	3D	366	9	323 K	无	46 K	0	22 690×6	–	27 468×6	1.3 M	无/无
H3D(Patil等,2019)	3D	160	8	83 K	无	27 K	0	8 873×3	5 170×3	13 678×3	1.1 M	无/无
A*3D(Song等,2019)	3D	–	7	39 K	有	39 K	–	39 179×1	–	–	230 K	有/有
CityScapes 3D(Cordts等,2016)	3D	–	8	25 K	有	0	0	2 975×1	500×1	1 525×1	40 K	无/无
ImageNet(Deng等,2009)	2D	–	21 841	14 M	–	–	–	1 281 167	50 000	100 000	–	–

注："–"表示官方没有提供详细信息。

图像的3D目标检测的训练阶段(Feng等,2021)。

基于图像的3D目标检测使用的图像标注主要为3D边界框。现实世界中的物体运动可以看做6个自由度的问题,一般的3D边界框的标注信息包括3D边界框的中心点坐标(x,y,z)、长宽高(l,w,h),以及偏转角(俯仰角、滚动角、偏航角θ)。由于交通场景下的目标是在道路平面上运动的,因此不考虑目标的俯仰角和滚动角(认为取值为0),通常使用7个参数(x,y,z,l,w,h,θ)标注3D边界框的3D位置、尺寸和姿态信息(Geiger等,2012;Caesar等,2020;Sun等,2020)。3D目标检测可以认为是一个多任务学习的问题,其目标就是对目标类别以及3D边界框的位置、尺寸和姿态信息进行回归预测。

在实际的检测任务中,由于不同类别目标的外观和运动差异性,采用以上标注方式会出现不同程度的方向估计偏差。Ku等人(2019a)对KITTI数据集中目标方向估计进行统计,发现一般算法对于小汽车和骑行者的方向预测精度远高于对于行人的方向预测精度。小汽车和骑行者的平均方向估计偏差(average angular errors,AAE)分别小于7°和20°,而行人的接近56°,其原因可能是3D边界框的标注方式忽略了目标类别之间的差异。具体来说,小汽车和摩托车/自行车作为刚性物体,其外形一般不会发生太大的改变,很容易定义其朝向,适合用一个有明确方向的3D立方体框进行标注;而行人是非刚性的,姿态多变,很多情况下行人的朝向难以定义,使用立方体状的3D边界框标注会导致不同个体间方向差异变大,从而影响到标注数据的质量和检测算法的方向预测结果。

考虑到行人的特殊性,采取新的标注方式(Dong和Isler,2020;Li等,2021a)可能提高其预测精度。比如,Dong和Isler(2020)通过椭球框回归预测物体的3维位置和尺寸(如图5所示)。结合几何图形形状特性和行人特性,可以对行人采取圆柱体加方向矢量的标注方式(如图6所示)。这种标注的潜在优势包括:1)采取圆柱体或者椭球体的标注方式可以适应行人的外形多变的特性,克服立方体标注导致方向预测误差较大的问题;2)采取圆柱体的标注方式可以减少回归参数,标注的参数只涉及物体类别c、圆柱体底部圆心(x,y)、半径r和圆柱体的高h,将立方体边界框的7个参数(x,y,z,w,h,l,θ)减少到圆柱体标注的4个参数(x,y,r,h),参数量得到削

图5　使用椭球进行物体3D目标检测示例(Dong和Isler,2020)

Fig. 5　Examples of 3D object detection using ellipsoid in the literature (Dong and Isler, 2020)

图6　使用圆柱体对行人进行3D标注示例

Fig. 6　Example of 3D annotation of pedestrians using cylinder

减，同时不影响对行人定位和尺寸的预测；3）采取圆柱体的标注方式还可以进一步减少不同标注者对同一行人进行标注产生的差异。

2.2　方法论制约因素和误差分析

2.2.1　通用框架

与2D目标检测算法的分类类似（Redmon等，2016；Girshick等，2014；Girshick，2015；Ren等，2017；Liu等，2016；Zhou等，2019），基于图像的3D目标检测算法根据在算法执行过程中是否使用中间表示，可以将算法分类为单阶段（one-stage）算法（Brazil和Liu，2019；Chen等，2020b；Wang等，2021b；Luo等，2021；Liu等，2020）和两阶段（two-stage）算法（Weng和Kitani，2019；You等，2020），如图7所示。

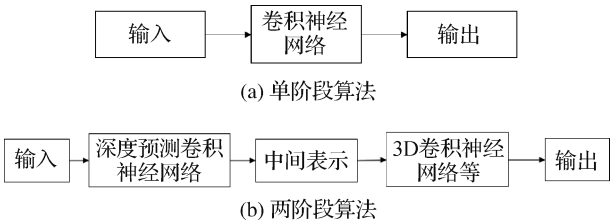


图7　单阶段和两阶段算法流程图

Fig. 7　Flow chart of one-stage and two-stage algorithms
((a) one-stage algorithm; (b) two-stage algorithm)

单阶段算法根据投影关系，利用卷积神经网络（convolutional neural network，CNN）直接在图像平面回归目标的3D信息。在算法执行过程中可能会使用深度提示（Brazil和Liu，2019；Gao等，2022；Wang等，2021a；Ku等，2019b）等信息来辅助目标3D信息的回归，但不会引入深度图或者伪激光雷达点云等中间表示。

两阶段算法在算法过程中会对图像域内的所有物体或者对2D推荐区域内的物体进行深度预测，生成深度图，利用深度图辅助3D目标检测（Reading等，2021；Badki等，2020；Chen等，2020c）。某些算法（Weng和Kitani，2019；You等，2020；Ma等，2019）还将深度图反投影到3维世界，生成伪激光雷达点云表示，然后利用目前比较成熟的基于激光雷达点云的3D目标检测算法（Qi等，2017a，b；Zhou和Tuzel，2018；Lang等，2019；Zhang等，2022；Mao等，2021；Xu等，2021；Zheng等，2021）在生成的伪激光雷达点云上对3D目标进行检测。

单阶段算法直接在2维图像平面上预测目标的3维信息，两阶段算法的第1阶段从图像反向预测深度图或伪激光雷达点云。由于存在3维空间到2维像平面的投影，如何有效利用投影变换和物体位置等几何关系成为提高基于图像的3D目标检测性能的关键因素之一。

由于维度坍塌导致图像缺少深度信息，单阶段目标检测算法会生成利用深度提示等辅助3D目标检测，两阶段算法则会进一步生成深度图或伪激光雷达点云等中间表示。无论是单阶段算法还是两阶段算法都会隐式或显式地利用深度信息。因此，对深度预测的精度直接影响3D目标检测算法的性能。

依靠深度学习算法生成的深度提示或深度信息精度不佳，而点云数据或者立体图像能够提供更准确的深度信息。在算法过程中模拟激光雷达点云（两阶段算法）或立体视觉，或直接使用立体视觉作为输入，或使用低线数激光雷达点云数据作为监督训练信号，都可以提高基于图像的3D目标检测的性能。此外，通过使用时空序列数据，如视频等手段也能够提高3D目标检测的性能。因此，选择合适的数据模态成为提高3D目标检测性能的一个途径。

以下将从几何关系、深度预测精度和数据模态3方面制约因素总结分析基于图像的3D目标检测的研究现状，以期能够启迪未来工作。

2.2.2　几何关系

要提高基于图像的3D目标检测性能，可以利用的几何关系包括3D目标投影到2D图像上的2D—

3D几何约束和物体间的成像位置关系。2D—3D几何约束主要表现为2D图像与3D透视图直接的几何关系互相约束。由于投影变换，2维图像缺少距离维信息，信息维度的提升面临一对多、甚至无解的情况，这属于视觉领域的不适定问题（杨步一 等，2021）。物体间的成像位置关系主要包括目标之间的遮挡、目标截断和成像尺寸大小变化。

有许多学者对如何利用2D—3D几何约束来提高3D目标检测精度提出了独特的见解。Brazil和Liu（2019）提出了一个独立的基于区域推荐的3D目标检测网络M3D-RPN（monocular 3D region proposal network），利用2D和3D透视图的几何关系，允许3D检测框利用图像空间中生成的强大的卷积特征。Mousavian等人（2017）提出使用深度卷积神经网络回归相对稳定的3维目标属性，然后将这些估计与目标的2维边界框提供的几何约束相结合，生成完整的3维边界框。Ku等人（2019b）提出一种利用候选框和形状重建的单目3D目标检测方法MonoPSR，使用2D目标检测器的检测来生成场景中每一个目标的3D候选框，回归以实例为中心的3D候选框，生成3维边界框，同时估计实例点云以恢复局部形状与比例，增强2D—3D的一致性。

使用2D—3D几何约束会存在特征失配等问题。为了缓解这些问题，Luo等人（2021）提出了一种具有特征对齐和非对称非局部注意的单目3D单级目标检测器M3DSSD（monocular 3D single stage object detector）。M3DSSD采取两步特征对齐的策略：第1步，执行形状对齐以便特征图的感受野能够聚焦于具有高度置信度的预定义锚框；第2步，使用中心对齐来执行2D/3D中心特征对齐。Li等人（2019a）提出了一个单目3D目标检测框架GS3D，旨在从2D图像中提取底层的3D信息，在没有点云或立体视觉数据的情况下确定物体的精确3维边界框。该方法可以为每个预测的2D长方形获取一个粗长方体，用来指导细化确定目标对象的3D预测框输出。与仅使用2D边界框提取的特征进行预测框细化的方法不同，该方法通过利用可见表面的视觉特征来探索目标对象的3D结构信息，利用曲面的新特征来消除仅使用2维边界框带来的表示模糊的问题。张峻宁等人（2020）提出一种基于透视投影的单目3D目标检测网络。首先，利用世界坐标系、相机坐标系以及目标坐标系三者之间的转换关系，建立一种利用消失点（vanishing point，VP）求解目标3维边界框的模型。其次，运用空间几何关系和先验尺寸信息，将其简化为方位角、目标尺寸与3维边界框的约束关系。最终，根据物体尺寸约束的单峰和易回归优势，进一步提出一种学习型的方位角—尺寸的损失函数，提高网络学习效率和检测精度。严娟等人（2020）针对3D检测任务中存在的特征间依赖关系利用不足的问题，提出了一种结合混合域注意力与空洞卷积的3D目标检测方法。该方法使用了融入混合注意力机制的特征提取器，即关注特征的通道域和空间域的注意力机制，有效突出了特征的通道与空间两个方面的关键特征。同时利用特征空洞卷积，增大了特征的感受野，获得了高分辨率的特征图。

目前，大多数涉及2D—3D几何约束的3D目标检测算法将从3维边界框到2维边界框的投影约束作为一个重要的辅助检测手段。考虑到2维边界框4个边缘仅提供了4个约束，微小的误差都会导致性能的降低，加入更多的约束能够促进3D目标检测网络的性能。Liu等人（2021a）提出了一种具有启发性的方法，引入了地平面预测作为额外的先验知识指导深度预测和下游的3D目标检测任务。Li等人（2020）提出了一个关键点检测辅助3D目标检测的算法，预测图像空间中3维边界框的9个透视关键点，然后利用3维和2维透视的几何关系来恢复3维空间中的尺寸、位置和方向。实验证明，即使在关键点估计非常嘈杂的情况下也可以稳定地预测目标的属性，这使得该方法能够以较小的模型实现快速的检测。

然而，有学者认为在3D目标检测过程中使用2D—3D先验是非必要的。Wang等人（2021b）通过建立在全卷积单级检测器上的实践来研究2D检测器在3D目标检测中的应用，并提出了一个通用框架FCOS3D（fully convolutional one-stage monocular 3D object detection）。该框架将通常定义的7个自由度3维目标转换到图像域，并将其解耦为2维和3维属性。然后，考虑物体的2维比例，仅根据训练过程的投影3维中心将物体分布到不同的特征级别。此外，还使用基于3维中心的2维高斯分布重新定义中心度，以适应3D目标检测公式。这些操作使得该框架脱离了任何2D检测或者2D—3D对应先验。Liu等人（2020）认为2D检测网络是冗余的，并且会为3D检测引入不可忽略的噪声。为此，提出了一种新的3D目标检测方法SMOKE（single-stage monocular

3D object detection via keypoint estimation)，通过将单个关键点估计与回归的3维变量相结合来预测每个检测到的目标的3D边界框。

由于投影关系的限制以及图像视域有限，图像中的物体相比现实中的物体往往存在截断、遮挡、自遮挡和成像尺寸变化等问题。有许多学者针对这些问题提出了改进办法。目前，大多数检测器将每个3D物体视为独立的训练目标，导致缺少遮挡样本的有用信息。Chabot等人(2017)提出使用车辆的特征点对3维车辆信息进行编码，其基本思想是使用单目图像恢复3维车辆信息。首先，使用由真实尺寸的3维网格组成的3维车辆数据集，为每一个3维模型标注了几个顶点，这些顶点对应车辆的不同的零部件(如车轮、照明灯等)，并且为每一个3D模型定义了3D形状。然后，通过恢复每一个检测到的车辆输入图像中的3D点的投影，为每一个检测框选择最佳对应的3维模型，在2D—3D之间进行匹配，最终恢复车辆方向、3维位置和尺寸(如图8所示)。Xu等人(2020)提出了一种新的基于立体图像的3D目标检测框架ZoomNet(part-aware adaptive zooming neural network)，并引入了学习零部件位置作为补充特征来提高抗遮挡能力。Chen等人(2020b)则提出了一种具有启发性的方法，通过考虑配对样本关系来改进单目3D目标检测。首先，对相邻样本对的位置和3D距离计算不确定性预测；然后，将单阶段不确定性感知预测结构和后优化模块集成在一起，以保证运行效率。如图9所示，对于任意的样本对，通过将其2维边界框中心的距离设置为直径来定义范围圆，如果该样本对包含其他物体中心，则忽略该样本对(如图9(a)所示)。实验证明，在难识别样本检测性能上，该方法取得较好的效果。

Ma等人(2021)通过密集诊断实验发现，基于现有技术水平，精确定位远处的物体是几乎不可能的，而这些样本的存在将会误导网络的学习。因此，建

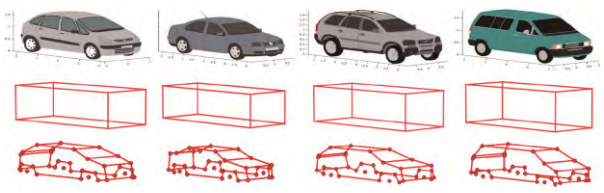


图8　车辆关键部分特征点编码示例(Chabot等，2017)

Fig. 8　Examples of feature point coding for key parts of the vehicle(Chabot et al., 2017)

(a) 用于训练和推理的配对匹配策略

(b) 有效样本对示例

图9　样本对及范围圆示例(Chen等，2020b)

Fig. 9　Examples of sample pair and range circle (Chen et al., 2020b)((a)pair matching strategy for training and interence;(b)example image with all effective sample pairs)

议从训练集中删除远处的小目标样本以提高检测器的性能。在自动驾驶领域往往更关注近处的物体，所以该建议具有一定的现实意义。Xu等人(2020)引入了一种自适应缩放模块，将2D实例边界框调整为统一的分辨率，并相应地调整相机的内在参数，通过这种方式可以从调整大小的长方形图像中估计出更高质量的视差图，进而为附近和远处的物体构建密集的点云。Roddick等人(2018)认为现有的系统大多受基于透视图像的表示法的局限，目标的外观和比例随深度而急剧变化，很难在图像域中推断出有意义的距离(即深度预测)。他们认为3D目标检测的一个基本任务是在3D空间中理解世界，因此引入了正交特征变换，通过将基于图像的特征映射到一个正交的3维空间，跳出图像域的限制，从而能够在尺度一致且目标之间的距离有意义的域中整体地理解场景的空间结构。

由于大多数现有方法对所有目标采用相同的处理方式，而不管其在图像中位置的分布不同，从而导致对于截断物体的检测性能有限。为了解决这一问题，Zhang等人(2021)提出了一种灵活的单目3D目标检测框架，解耦了截断物体，并能自适应地结合多种物体深度估计方法。该方法根据物体投影的3D中心是在图像内部还是外部，将物体分为内部物体和外部物体两组，并将内部物体和外部物体的表示和偏移解耦。对于投影中心在图像内的对象，由其中心投影x_c直接识别；对于投影中心在图像外的物体，为了解耦表示，该方法通过2维边界框中心x_b与3维边界框中心投影x_c的连线和图像边缘的交点x_i来识别表示外部物体(如图10(a)所示)。其中，x_i的

预测通过 1 维高斯分布生成的边缘热图来实现(如图 10(b)所示),对于严重截断的物体,x_i比x_b有更好的表示(如图 10(c)所示)。

尽管目前在利用几何关系辅助目标检测方面取得了不错的效果,但是图像中蕴含的道路标线标志等大量细节所提供的几何信息、不同距离物体成像大小隐含的深度信息等仍有待挖掘。当前基于图像的 3D 目标检测算法精度仍然较低,2D—3D 投影关系的有效利用和难识别样本的检测仍是限制目前算法性能的关键因素和今后研究的重点方向。

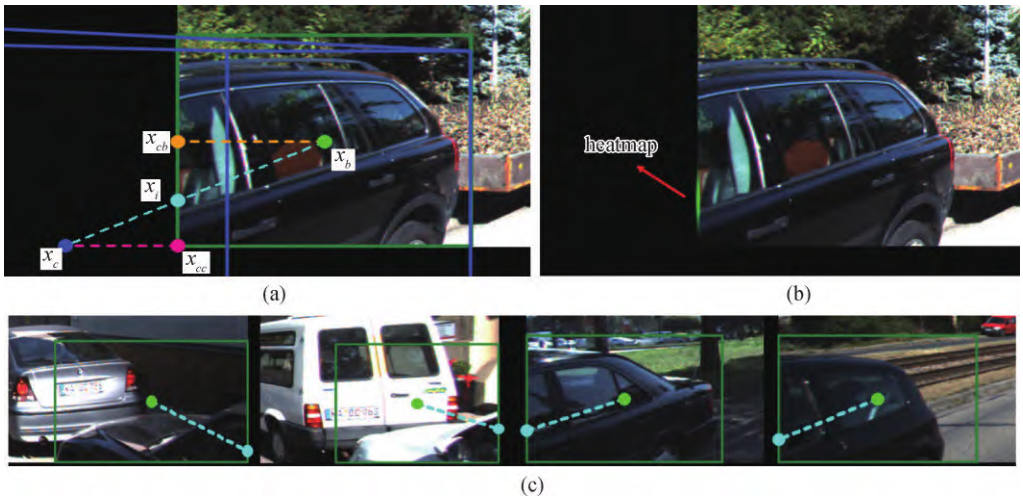


图 10　外部物体表示(Zhang 等,2021)

Fig. 10　External object representation(Zhang et al. ,2021)

2. 2. 3　深度预测精度

无论是在算法过程中引入深度信息的单阶段 3D 目标检测算法,还是生成深度图或者伪激光雷达点云等中间表示的两阶段 3D 目标检测算法,其中关键一步是对于目标深度的预测。然而,图像数据天然地缺少深度信息,因此深度预测是基于图像的 3D 目标检测算法最为重大的挑战,也是造成与基于激光雷达点云的 3D 目标检测算法检测性能差异的主要原因。Ma 等人(2021)的研究表明,深度预测的准确与否将会极大地影响下游检测任务的精度。当前,基于图像的深度预测算法主要存在两方面的问题。1)受投影关系的影响,深度预测往往不够准确;2)连续性的深度预测在图像的深度突变处对深度预测性能往往不佳,而对预测深度离散化之后,则存在着深度的分类比较粗糙,精度类别不能任意划分的问题。

为了解决第 1 个问题,学者们考虑增强深度特征信息,或者使用深度提示线索,进一步对场景进行深度预测,生成深度图来辅助下游的目标检测任务。Brazil 和 Liu(2019)在所提出的单阶段目标检测网络 M3D-RPN 中引入了深度感知卷积层,以开发特定位置的特征,从而提高网络对 3D 场景的理解。Luo 等人(2021)提出了一种具有多尺度采样的非对称非局部注意模块来提取深度特征。周大可等人(2021)提出了一种基于局部平面参数预测的方法。该方法以 U-Net 为基础网络,主要功能模块有 3 个:1)基于 ResNet50(residual network 50)的编码网络;2)加入了串联双注意力机制的解码网络;3)利用局部平面参数估计的多尺度预测模块。预测模块能将深度图预测任务转化为平面参数预测问题,即将多个尺度的特征图输入转换成多个标准尺度的深度图,最后计算深度图合成的双目视图来与真实视图的重建误差。Gao 等人(2022)提出了一个深度提示增强策略,为了提供特征化的深度模式作为深度估计的提示,设计了一个专用的深度提示模块来生成称为深度提示的行特征,以分格方式显式监督模型训练。Liu 等人(2021b)提出了一种具有启发性的方法,旨在确定地平面如何在驾驶场景的 3 维检测中提供深度推理的额外线索。Eigen 等人(2014)提出了一种从单目图像中预测深度的方法,通过使用两个深度网络堆栈来解决深度预测的问题,其中一个深度网络基于整个图像进行粗略的全局预测,另一个深度网络则局部细化此预测,应用比例不变误差来帮助测量深度关系。王泉德等人(2022)提出了一种强化

边缘的单目图像深度预测方法，利用单目RGB图像通过深度估计网络得到预测深度图。由于预测深度图边缘变化过于平缓，再将预测深度图和RGB图像输入至深度补偿网络中，基于输入的单目RGB图像得到对应的深度补偿图。之后将预测深度图与深度补偿图进行融合，得到边缘清晰的深度图。

为了能够更好地从2D图像中获取深度信息，Ding等人(2020)改进了传统的2维卷积，提出了一个深度引导动态深度扩展局部卷积网络D^{4}LCN(depth-guided dynamic-depthwise-dilated LCN)，该网络可以从基于图像的深度图中自动学习滤波器及其感受域，使不同图像的不同像素具有不同的滤波器。该方法克服了传统2维卷积的限制，缩小了图像表示和3D点云表示之间的差距。Qin等人(2019a)提出了一个由4个特定子任务网络构成的单目3D目标检测网络MonoGRNet(monocular geometric reasoning network)。该网络包含一个实例深度估计(instance depth estimation, IDE)子网络，使用稀疏监督直接预测目标3维边界框的深度，通过估计水平和垂直维度中的位置，进一步实现了3维定位，最后在全局环境中优化3维边界框和位姿联合学习。Peng等人(2020)在所提方法中引入一个实例深度感知模块(instance depth aware, IDA)来进行深度预测，用于辅助基于立体视觉的3D目标检测。该模块能够通过实例深度感知、视差自适应和匹配加权来准确预测3维边界框中心的深度。Zhang等人(2021)则将直接回归的目标深度和来自不同关键点组的求解深度进行组合，作为目标的深度估计输出。Wang等人(2021a)提出一种深度调节的动态信息传播网络DDMP(denoising diffusion probabilistic models)，有效地将多尺度深度信息与图像上下文信息集成。该方法首先在图像中自适应地采样上下文感知节点，然后动态预测用于传播信息的混合深度相关滤波器权重和相似性矩阵，并通过增加中心感知深度编码(center-aware depth encoding, CDE)任务，成功地缓解了不准确的深度先验。王秋晨等人(2022)提出一个递归特征融合的单目深度累积估计方法。单目图像通过特征提取模块产生不同尺度的特征，之后通过递归特征融合模块充分融合多尺度信息。该递归特征融合模块借助卷积的GRU(gated recurrent units)的更新和遗忘机制，按照潜在的顺序对特征进行有规律的融合。在解码器阶段，深度重建过程被分解为多层，不同层分别预测不同细粒度的深度图，最终将不同细粒度的深度图累计生成最终的深度估计结果。张竞澜等人(2022)提出一种基于动态空间金字塔池化(dynamic spatial pyramid pooling, DSPP)的单目图像深度估计方法。首先针对单目图像提取不同分辨率的特征图，之后通过一个动态密集的DSPP模块进行特征融合，最后再通过解码器将最终的特征图变为场景深度图。其中，DSPP模块基于空洞空间金字塔池化(atrous spatial pyramid pooling, ASPP)思想，通过结合通道注意力充分利用每一层特征，在减少网络参数量的前提下，提升了模型整体的准确率。阮晓钢等人(2022)提出一种基于双鉴别器生成对抗网络的单目深度估计方法，利用生成对抗网络来生成准确的视差。该方法的生成对抗网络包含一个生成器和两个判别器，生成器用来生成视差图，真实图像与视差图合成重建图像。重建的左右目图像与真实左右目图像分别作为判别器的输入，由判别器来辨别输入是真实图像还是重建图像，交替训练生成器和判别器，直到判别器无法辨别哪些是重建图像和真实图像，这样网络可得到较为准确的深度。张聪等人(2022)提出一种基于通道注意力机制的单目深度估计网络SE-DenseDepth(squeeze-and-excitation dense depth)，解决细节估计不准确、同一平面距离估计错误的问题。该网络在编码器中嵌入通道注意力机制，依据不同通道对深度信息不同的贡献度对通道进行编码，提高编码器对图像特征的表征能力。同时，为了获取精细的图像深度信息，建立编码器到解码器的跳跃连接，从而可以挖掘更多的低层信息。杨蕙同等人(2022)针对复杂场景深度估计常出现的误匹配现象，提出一种多尺度注意力特征融合立体匹配算法MGNet。该算法提出了一个轻量级的相关注意力模块，捕获全局上下文信息和远距离通道依赖关系；设计了一个多尺度卷积全局注意力模块，捕获多尺度上下文和全局上下文信息；在代价聚合阶段引入通道注意力，抑制具有歧义的匹配信息，提取有区别性的特征。通过以上操作，该算法在具有反射区域、重复纹理等复杂场景中均有优异的表现。

因为在物体的轮廓边界处的深度会发生突变，采用连续的深度值进行深度预测会产生较大的误差，导致轮廓边界处模糊不清，不利于将前景的目标物体与背景分离开。Weng和Kitani(2019)发现，由

预测深度生成的伪激光雷达点云信号通常存在着“长尾”等问题，如图 11 所示。图 11(a)为激光雷达点云数据，图 11(b)为根据深度图生成的伪激光雷达点云数据，图 11(c)为激光雷达点云数据与伪激光雷达点云数据对比。根据深度图生成的伪激光雷达点云数据在目标的轮廓边界处存在“长尾”和错位，即图 11(b)(c)中椭圆框出的区域。其原因之一就是估计的深度在目标边界不准确。考虑到这种不准确可能是由在 2D 目标检测时使用 2 维边界框导致的，因此建议使用实例掩膜代替 2 维边界框作为目标的候选区域，通过这样的处理，生成伪激光雷达点云的“长尾”问题可以得到一定的缓解。

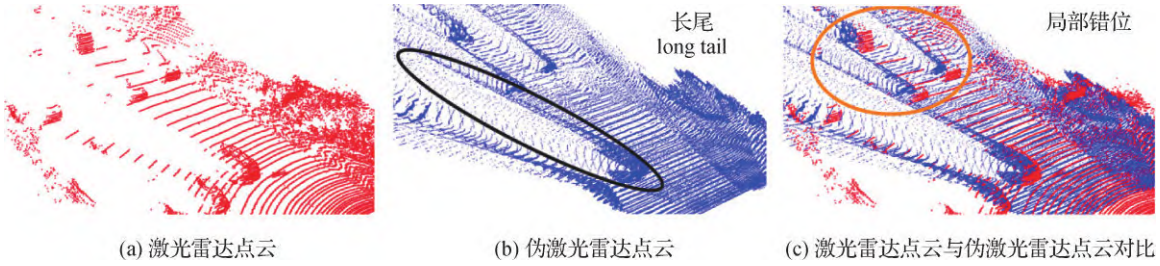


图 11　伪激光雷达点云“长尾”等问题可视化(Weng 和 Kitani，2019)

Fig. 11　Visualization of the “long tail” of the pseudo-LiDAR point cloud(Weng and Kitani, 2019)((a)LiDAR point cloud; (b)pseudo-LiDAR point cloud; (c)comparison between the LiDAR point cloud and pseudo-LiDAR point cloud)

也有学者提出将深度离散化处理，将深度预测问题转变为深度分类问题。Reading 等人(2021)提出了一个分类深度分布网络 CaDDN(categorical depth distribution network)。该方法将深度进行离散化，将深度预测问题变为一个深度分类问题，对每一个像素预测其离散化的深度分类，将丰富的上下文特征信息投影到 3 维空间中的适当深度间隔，然后使用计算效率高的鸟瞰投影和单级检测器来产生最终的输出检测。不过，这类深度离散化策略的精度难以任意调整。因此，Badki 等人(2020)提出了一种通过一系列二元分类来估计深度的方法 Bi3D 。该方法不预测目标是否处于特定的深度 D，而是通过二分类的方法，将目标分类为比深度 D 更近或者更远。使用二分类的方法大幅提高了网络的检测速度，而且通过多次二分类，该算法能够实现任意精度的深度估计。Garg 等人(2021)使用了一个能够输出任意深度值的新神经网络构架，提出了一个从真实分布和预测分布之间的 Wasserstein 距离导出的新损失函数来缓解这个问题。

由于 3D 目标检测通常包含有深度预测和目标检测两个相对独立的部分，目前大多数的算法都是独立优化这两个子任务。一些学者认为，不应该将深度预测与下游的检测任务分离开来，而应该在两者之间建立联系，以根据任务需求联合优化。Wang 等人(2021c)提出了一个基于立体视觉的 3D 目标检测方法 PLUMENet，对于基于立体图像进行深度估计生成伪激光雷达点云和基于伪激光雷达点云进行 3D 目标检测两个任务，直接在 3 维空间中构建一个伪激光雷达特征体，用于解决两个子任务在不同的度量空间中分别优化导致整体结果次优的问题。Qian 等人(2020)提出了一种基于可微变化(change of representation，CoR)模块的新框架，对基于伪激光雷达点云表示的 3D 目标检测算法进行端到端的训练，弥补了两个任务之间的割裂性，并且能够与大多数先进算法兼容。

虽然深度预测在单目 3D 目标检测中具有重要的辅助作用，且目前大多数的算法都是在深度预测的基础上进一步进行目标检测的，然而，Simonelli 等人(2020)认为深度预测辅助 3D 目标检测是非必要的，并在所提方法中引入了一种用于单目 3D 目标检测的新型单级深度结构 MoVi-3D(monocular visual-inertial 3D object detection)。该方法利用几何信息在训练和测试时生成虚拟视图，并对目标外观与距离关系进行了标准化，显著减少距离带来的视觉外观变化影响。通过生成虚拟视图这种方式，深度模型不再需要学习深度特定表示，大幅降低了算法的复杂性。

2.2.4　数据模态

数据模态涉及两个方面，1)在算法过程中模拟激光雷达点云(两阶段算法)或者立体视觉；2)使用

立体视觉、多模态数据等作为辅助输入或者监督信号。

一般认为，基于图像的3D目标检测算法性能较差的原因是基于图像的深度预测性能不佳，但是Wang等人(2019)认为造成基于图像的3D目标检测算法与基于激光雷达点云的3D目标检测算法性能差异的主要原因不是数据的质量，而是数据的表示，提出将基于图像的深度图转换为伪激光雷达点云表示，这实际上是模拟激光雷达点云信号。通过这样的转换表示，目前性能较好的基于激光雷达点云的3D目标检测算法可以应用到伪激光雷达点云信号处理中。Weng和Kitani(2019)也使用伪激光雷达点云作为3D目标检测中间表示。该方法按照两阶段3D目标检测的流程，首先检测输入图像的2D目标候选区域，从生成的伪激光雷达点云中为每一个推荐区域提取点云截锥体，并为每个截锥体检测一个定向的3维边界框。为了处理伪激光雷达点云中的大量噪声，提出了两点创新，1)使用2D/3D边界框一致性约束，保证3维边界框投影与对应的2D候选区域具有较高的重叠度；2)使用实例掩膜而不是2维边界框作为2D候选区域的表示，以减少不属于点云截锥体的数据。Ma等人(2019)利用一个独立的模块将输入数据从2维图像平面转换到3维点云空间，以获得更好的输入表示；然后使用点云主干网络进行3D目标检测，以获得目标的3维位置、尺寸和方向。除此之外，为了增强点云的识别能力，还提出一种多模式的特征融合模块，将互补的图像RGB特征线索嵌入到生成的点云表示中。

目前使用伪激光雷达点云表示的基于图像的3D目标检测算法显示出了强大的能力，但是该类方法过度依赖独立的深度估计器，在训练阶段需要大量的像素注释，在推理阶段则需要大量计算，限制了其现实应用。Li等人(2021b)提出一种高效、准确的立体图像3D目标检测方法RTS3D(real-time stereo 3D detection)。与伪激光雷达方法中的3维表征空间不同，该方法引入了一种新的4D特征一致傅里叶轮廓嵌入(Fourier contour embedding，FCE)空间作为3维场景的中间表示。FCE空间通过探索从立体对扭曲的多尺度特征一致性来编码目标的结构和语义信息，从而无需深度监督。此外，该方法还设计了语义引导径向基函数(radial basis functions，RBF)和结构感知注意模块，以减少FCE空间噪声的影响，而无需实例掩膜监督。

除了模拟3D点云数据以外，还有学者提出模拟立体视觉的方法。Zhou等人(2022)提出一种立体引导的单目3D目标检测框架SGM3D(stereo-guided monocular 3D object detection network)，使用从立体输入中学习到的鲁棒的3维特征来增强单目检测的特征。该方法提出了一种多粒度域自适应(multi-granularity domain adaptation，MG-DA)机制，使网络能够从单目图像生成模拟立体视觉的特征，包括粗鸟瞰图级和精细的锚框级特征。该方法还引入了一种基于IoU匹配的对齐方式，补偿锚级别域自适应过程中存在的失配问题，以在立体和单目预测之间实现目标级域自适应。Chen等人(2022)提出一种伪立体3D目标检测框架，从单目图像输入中模拟立体图像特征，其中包含3种新的虚拟视图生成方法，分别是图像级生成、特征级生成和特征克隆，用于辅助从单个图像中检测3维目标。此外，还提出一种基于视差特征映射的动态核的视差方向动态卷积，用于自适应地从单个图像中过滤特征，以生成虚拟图像特征，从而缓解了深度估计误差引起的特征退化。Chen等人(2020c)提出一个深度立体几何网络DSGN(deep stereo geometry network)，通过可微体积表示法，有效编码了3D规则空间中的3D几何结构，从而可以同时学习到深度信息和语义线索。

提升基于单目视觉的3D目标检测算法性能的一个重要方法是使用多模态数据。激光雷达点云数据能够提供精确的深度信息，但是目前激光雷达价格昂贵，不适用于在大规模的自动驾驶车辆上进行部署。一个经济实惠的折中方案是使用立体视觉图像或者使用低线数的激光雷达点云数据作为辅助数据输入。

基于立体视觉进行3D目标检测主要利用的是双目视差信息。双目视差原理如图12所示。

根据成像原理，可以得到深度估计与视差估计之间的函数关系，它们之间的关系为

$$D(u,v)=\frac{f\times b}{z(u,v)} \tag{7}$$

式中，(u,v)代表图像中的像素坐标，$D(u,v)$为像素点(u,v)的视差估计，f为相机焦距(左右相机使用同一焦距的镜头)，b为左右相机的水平距离(默认左右相机在同一水平线上)，$z(u,v)$为像素点(u,v)的深度估计。

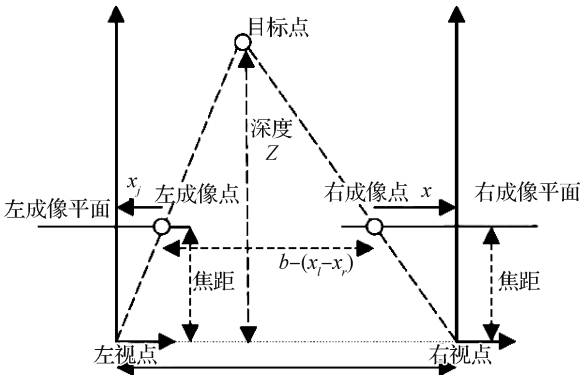


图 12　双目视差成像原理图

Fig. 12　Binocular parallax imaging schematic

从立体图像中检测 3 维物体的关键挑战是如何有效利用立体图像中密集的语义和几何信息。Li 等人(2019b)提出一种基于立体视觉的 3D 目标检测算法 Stereo R-CNN (stereo region convolutional neural network),扩展了 Faster RCNN。该方法同时检测和关联左右图像中的目标对象,并在立体区域推荐网络(region proposal network,RPN)之后添加额外的分支来预测稀疏关键点、视点和目标维度。接着,把这些关键点、视点和目标维度与左右 2 维边界框相结合,粗略计算目标 3 维边界框。然后,通过使用左右感兴趣区域(region of interest,ROI)的基于区域的光度对齐来恢复精确的 3 维边界框。Chen 等人(2018)提出通过利用立体图像实现 3D 目标检测的方法,构建了一个能量函数,对目标的大小先验、目标在地平面上的位置以及几个与自由空间、点云密度和到地面距离有关的深度信息特征进行编码,并利用此函数最小化结果生成一组高质量的 3D 目标候选。然后,在此基础上使用卷积神经网络进行目标检测,利用上下文和深度信息联合回归 3D 边界框坐标和目标位姿。与使用像素级的深度图的方法不同,Qin 等人(2019b)建议使用 3D 锚框来明确构建立体图像中感兴趣区域的目标级对应关系,指导深度神经网络从中学习检测和三角测量 3 维空间中的目标对象。该方法中还介绍了一种经济高效的特征通道重加权策略,增强了代表性特征并削弱了噪声信号以促进学习的过程。Liu 等人(2021a)为了解决基于立体视觉的 3D 目标检测速度较慢的问题,提出了一个基于立体视觉的 3D 目标检测框架 YOLOStereo3D。该方法从基于 2D 图像的检测框架中获得特征信息,并且使用立体特征增强它们。YOLOStereo3D 结合了实时的单级 2D/3D 目标检测器的知识和推理结构,引入了一个轻量级的立体匹配模块,使算法具有较高的运行速度。王一强和陶洋(2022)提出一种基于立体区域卷积神经网络改进的立体视觉 3D 目标检测网络 FR-CNN。该方法首先在特征提取网络中加入频域通道注意力模块,使算法可以关注更多与目标相关的语义信息,减少深层残差网络权重变化所带来的影响。此外,网络中还加入了统一动态样本加权策略,在网络训练时为“困难”样本和“简单”样本进行合理的损失权重分配,提取目标更为全面的关键特征信息。王康如等人(2020)提出一种迭代式自主学习的立体视觉 3D 目标检测算法。首先,利用迭代式自主学习的视差估计算法来进行目标区域的视差估计。随后,通过相机内外参将视差信息转换为场景点云,利用自适应特征融合模块将 RGB 信息和点云进行融合以实现精准的目标检测。张羽丰等人(2021)提出一个基于立体图像的目标检测与目标距离估计网络。首先,利用 R-CNN 网络搭建基本网络,然后使用双目候选框提取网络代替原有的候选框提取网络。为解决直接利用边界框计算目标距离的误差较大问题,该网络加入了一个专门的视差回归分支,合并了基准框内的左右视图的图像特征,最后对左右视图合并的目标框与该基准框的视差进行回归。于洁潇等人(2021)认为利用左右目视图和校准信息之间的关系以及左右视差图的一致性可以获取更加精准的深度信息,提出一种改进的立体区域卷积神经网络算法,以确定性网络 DetNet(deterministic networking)作为骨干网络,提升对远景目标的检测效果,并基于左右目视图的关键点,建立左右视图关键点一致性损失函数,提高潜在关键点的位置精度和算法的检测准确性。赵邢等人(2019)提出一种基于立体视觉的车辆目标检测算法。首先利用基于深度学习的目标检测方法获取车辆在 2 维图像上的信息,结合深度相机利用立体视觉获取车辆的关键 3 维空间信息;然后综合 2 维与 3 维信息建立 3 维空间坐标,计算实现车辆的 3 维边界框绘制,辅助区分车辆空间方位。该方法为端到端方法,不需要其他额外的输入信息。迟旭然等人(2022)针对自动驾驶场景提出一种基于 Stereo-RCNN 的 Fast Stereo-RCNN 的 3D 目标检测算法。首先,用单分支网络获得目标 3 维边界框的多个角点来重构目标 3 维中心点,利用轻量级区域生成网络固化 3 维关键

点，采用二分支关键点检测网络锐化算法的目标辨别能力，再结合双层特征融合网络缩短底层特征到高层特征的传递路径。曹杰程和陶重犇（2021）提出一种基于锚框引导的立体视觉两阶段3D目标检测算法FGAS RCNN。在第1阶段，双目RGB图像分别生成前景概率图并根据获得的稀疏锚点预测物体形状，根据预测锚点未知和锚框形状输出相应的ROI建议框。第2阶段的关键点生成网络利用稀疏锚点信息生成关键点热力图，并结合3维立体回归生成目标3维预选框。苏凯祺等人（2022）提出一种基于立体图像的多路径特征金字塔3D目标检测网络MpFPN（multi-path feature pyramid network）。该网络在特征提取模块增加自底向上，由上至下的路径以及输入特征图至输出特征图之间的连接，为后续联合区域推荐网络提供了更高级的语义信息以及细粒度的多尺度空间信息，从而提高网络的检测精度。

除了使用深度图作为监督信号外，有学者提出利用价格低廉的低线数激光雷达点云数据作为深度预测的监督信号。Feng等人（2021）提出一种新型的两级网络FusionDepth，通过利用低成本的4线激光雷达点云数据，促进自监督单目视觉密集深度学习。该方法首先融合单目图像特征和稀疏激光雷达点云特征来预测初始深度图，然后设计一个高效的前馈优化网络，实时纠正伪3维空间中初始深度图的错误。秦超等人（2022）提出一个利用低线数激光雷达和相机实现的3D目标检测算法。首先将64线激光雷达点云采样量降至原始点云数量的10%，将稀疏点云和RGB图像输入至深度补全网络中获得深度图，以生成激光点云。之后将深度图转为点云俯视图，再通过基于关键点特征金字塔的3D目标检测网络得到物体边界框的几何信息、类别信息等。

运动信息是人类视觉用于检测、跟踪和深度感知的重要依据。目前的大多数基于图像的3D目标检测算法都是基于静态的图像，没有充分利用运动的时间序列信息。Brazil等人（2020）提出了一种新的基于单目视频的3D目标检测方法，利用运动来提取场景动力学特征并提高了定位精度。

此外，赵华卿等人（2019）认为对于3D目标检测来说，其方向角的先验信息并未得到充分利用，据此提出了一种基于2维图像估计先验方向角的3D目标检测算法。算法通过颜色信息和深度信息得到2维的分割实例，并在分割实例上提取关键点。然后，通过关键点的优化过程排除不确定点和修正误判点，通过点云重建得到关键点的3维坐标，根据关键点坐标估计目标的方向角，并将其作为初始化3维边界框的方向角。最后，将2维RGB图像与深度图进行特征融合，对初始化的3维边界框以及方向角进行回归，以得到更好的检测效果。

事实上，基于图像的3D目标检测方法中的各种限制因素并不是各自完全独立的，很多学者通过综合克服或者缓解多方面的算法难点，获得了性能的提升。

3 国内外研究对比

如果从双目视觉领域看，3D目标检测的研究很早就开始了，但是在自动驾驶的3D目标检测的研究和应用领域上，国内外学者更倾向于采用以图像为主的多模态数据融合方式来实现。从这个角度来看，虽然国外的研究起步较早，但国内外的研究差距并不大。

在数据集方面，本文所列的4个主要数据集发布时间依次为2012、2019、2019、2022年，跨度约10年。KITTI 3D数据集发布于2012年，是最为广泛使用的数据集，由于其发布时间早，也是各种算法研究论文中使用最多的数据集。而我国研究人员于2022年2月发布的DAIR-V2X数据集，则有着更加丰富的标注信息。表15对各个数据集的发布时间、场景多样性、数据精度、图像数据量和标注信息等各方面进行了对比。可以看出，数据精度基本上与采集设备的发展水平正相关，研究者都采用了当时主流的、较为高端的采集设备。

从评价基准看，各个数据集给出的评价指标不尽相同。我国的DAIR-V2X沿用了目标检测领域常用的mAP指标，数据也类似KITTI那样分为简单、中等和困难3级。针对车路协同应用，增加了数据传输消耗这个指标，使用位数（bit）评估通信成本，以衡量在车路协同3D目标检测中使用了多少路端数据。目标检测方面，没有给出更多的评价指标。

在算法研究上，国内外学者的差距很小，各个技术路线上都有国内学者发表顶级论文。以KITTI数据集3D目标检测Car类别的中等检测结果为基准，

表 15　4个数据集的图像数据对比

Table 15　Comparison of image data from 4 datasets

名称	发布时间	发布单位	图像数据量/幅	数据精度/像素	标注数据量/幅	标注信息
KITTI 3D	2012	Karlsruhe Institute of Technology	15 K	1 224 × 370	80 K/80 K	3D边界框/2D边界框
nuScenes	2019	Motional团队	1.4 M	1 600 × 900	1.4 M/–	3D边界框/2D边界框
Waymo Open Dataset	2019	谷歌Waymo公司	1 M	1 920 × 1 280 /1 920 × 1 040	12 M/9.9 M	3D边界框/2D边界框
DAIR-V2X	2022.2	清华大学智能产业研究院	71 K	1 920 × 1 080	1.2 M/–	3D边界框/2D边界框

注："–"表示没有2D边界框标注信息具体数据。

根据统计(Paperswithcode，2023b)，至2023年1月30日，基于单目3D目标检测方法性能top20的论文中，第一作者为国内单位的论文12篇，第一作者为国外单位的论文8篇。性能最佳的方法为浙江大学的Hong等人(2022)提出的CMKD(cross-model knowledge distillation)，使用了约4.2万幅未标注的KITTI数据作为额外的训练数据，Car类别的AP指标检测结果在简单、中等和困难数据上分别可以达到28.55%、18.69%和16.77%，中等的AP指标比第2名(Liu等，2022)高出2.23%。可以看出，目前基于单目3D目标的检测精度仍处于较低水平。在基于立体视觉3D目标检测性能top10的论文中(Paperswithcode，2023a)，第一作者为国内单位的论文5篇，第一作者为国外单位的论文5篇。性能最佳的方法是香港中文大学计算机科学与工程系的Chen等人(2023)提出的DSGN++，Car类别的AP指标检测结果在简单、中等和困难数据上分别为82.21%、67.37%和59.91%，检测结果的AP_{75}为67.37%，比第2名(Guo等，2021)高2.71%。可以看出，目前基于立体视觉3D目标的检测精度处于一个比较有限的水平。从以上研究成果看，国内外研究水平相差不大。表16给出了一些3D目标检测算法的性能对比，表中算法性能均是在KITTI数据集上对Car类别进行测试得到，取IoU阈值为0.7，-/-/-表示在验证集Val1(Chen等，2015)/Val2(Xiang等，2017)/Test上测试得到的性能，带*数据表示$AP_{|R40}$指标，不带*数据表示$AP_{|R11}$指标。数据模态栏中，M表示monocular、S表示stereo、M1表示monocular+ LiDAR supv、M2表示monocular video、PC表示point cloud。

在产业应用上，目前国内在自动驾驶领域领先的企业或研究机构主要有百度(Baidu，2022)、小马智行(Pony. ai，2022)和华为(WorldAuto，2022)等，其车载感知设备主要为激光雷达、高分辨率相机和毫米波雷达等，在感知模块使用的数据以激光雷达点云为主，图像等数据为辅。百度公司最新的Apollo RT6第6代无人驾驶车辆配备了38个激光雷达和高相机等传感器硬件，配合最新一代的Apollo自动驾驶系统和1 200 TOPS高算力计算单元，具备了L4级自动驾驶能力，实现了在复杂的城市道路自动驾驶。小马智行与上汽集团推出的Aion LX车型配备了激光雷达、高分辨率相机等17个传感器硬件，车辆的可视范围达到360°和最远200 m，其第3代自动驾驶软硬件方案PonyAlpha已经具备了L4级自动驾驶能力。极狐阿尔法S华为HI版车型配备了3路126线车规级激光雷达、13个高分辨率相机、6路毫米波雷达和12个超声波雷达传感器，搭载了华为ADS高阶自动驾驶全栈解决方案，芯片运算能力可达400 TOPS，具备了L4级别的自动驾驶能力，能够在复杂的城市环境中无人驾驶。

国外在自动驾驶领域领先的企业或研究机构主要有谷歌Waymo(池娟，2021)和Tesla(Talpes等，2020)等。谷歌Waymo的自动驾驶车辆配备了激光雷达、高分辨率相机和毫米波雷达等20多个传感器硬件，实现了对周围环境的全面感知，其第5代Waymo driver自动驾驶方案已经具备了L4级别的自动驾驶能力，在公共道路上完成了2 000万英里的自动驾驶测试和超过100亿公里的模拟驾驶测试。Tesla自动驾驶车辆采用纯视觉的自动驾驶感知方案，其自动驾驶车辆配备了8个相机，车辆可视范围达到360°和最远250 m，基于其自研的多头神经网络HydraNet的FSD构架具备了L2+的自动驾驶能力。

表16　部分3D目标检测算法性能表

Table 16　Performance of some 3D object detection algorithms

模型	模态	2D目标检测			鸟瞰图3D目标检测			3D目标检测		
		简单	中等	困难	简单	中等	困难	简单	中等	困难
Mono3D	M	–	–	–	5.22/–/–	5.19/–/–	4.13/–/–	2.53/–/–	2.31/–/–	2.31/–/–
Deep3DBox	M	–	–	–	–/9.99/–	–/7.71/–	–/5.30/–	–/5.85/–	–/4.10/–	–/3.84/–
MultiFusion	M	–	–	–	22.03/19.20/17.73	13.63/12.17/9.62	11.60/10.89/8.22	10.53/7.85/7.08	5.69/5.39/5.18	5.39/4.73/4.68
3DOP	S	–	–	–	12.63/–/–	9.49/–/–	7.59/–/–	6.55/–/–	5.07/–/–	4.10/–/–
M3D-RPN	M	90.24/–/84.34	83.67/–/83.78	67.69/–/67.85	25.94/26.86/26.43	21.18/21.15/18.36	17.90/17.14/16.24	20.27/20.40/20.65	17.06/16.48/15.70	15.21/13.34/13.32
GS3D	M	–	88.85/90.02/–	–	–	–	–	8.71/9.12/–	6.64/6.71/–	6.11/6.31/–
GS3D (scls)	M	–	–	–	–	–	–	11.63/13.46/–	10.51/10.97/–	10.51/10.38/–
MonoGRNet	M	–	–	–	24.97/–/–	19.44/–/–	16.30/–/–	13.88/–/–	10.19/–/–	7.62/–/–
MonoDIS	M	90.23/–/–	88.64/–/–	79.10/–/–	24.26/–/–	18.43/–/–	16.95/–/–	18.05/–/–	14.98/–/–	13.42/–/–
MonoPair*	M	–/–/96.61	–/–/93.55	–/–/83.55	24.12/–/–	18.17/–/–	15.76/–/–	16.28/–/–	12.30/–/–	10.42/–/–
RTM3D (ResNet18)	M	–	–	–	20.81/21.34/–	16.60/16.48/–	15.80/15.45/–	18.13/18.38/–	14.14/14.66/–	13.33/12.35/–
RTM3D (DLA34)	M	–	90.14/91.85/–	–	25.56/24.74/–	22.12/22.03/–	20.91/18.05/–	20.77/19.47/14.41	16.86/16.29/10.34	16.63/15.57/8.77
MoVi-3D*	M	–	–	–	–/–/22.76	–/–/17.03	–/–/14.85	–/–/15.19	–/–/10.90	–/–/9.26
M3DSSD	M	–	–	–	34.51/–/24.15	26.20/–/15.93	23.40/–/12.11	27.77/–/17.51	21.67/–/11.46	18.28/–/8.98
MonoDLE	M	–	–	–	24.97/–/24.79	19.33/–/18.89	17.01/–/16.00	17.45/–/17.23	13.66/–/12.26	11.68/–/10.29
MonoFlex*	M	–	–	–	–	–	–	23.64/–/19.94	17.51/–/13.89	14.83/–/12.07
MonoPSR	M	–	–	–	20.63/21.52/20.25	18.67/18.90/17.66	14.45/14.94/15.78	12.75/13.94/12.57	11.48/12.24/10.85	8.59/10.77/9.06
MonoRUn	M	–	–	–	–	–	–	17.26/–/16.04	12.27/–/10.53	10.41/–/9.11
MonoRUn	M1	–	–	–	–	–	–	20.02/–/19.65	14.65/–/12.30	12.61/–/10.58
CaDDN*	M	–	–	–	–	–	–	–/–/19.17	–/–/13.41	–/–/11.46
AM3D	M	–	–	–	43.75/–/27.91	28.39/–/22.24	23.83/–/18.62	32.23/–/21.48	21.09/–/16.08	17.26/–/15.26
D^4LCN	M	93.59/–/–	85.51/–/–	68.81/–/–	34.82/–/–	25.83/–/–	23.53/–/–	26.97/24.29/16.65*	21.71/19.54/11.72*	18.22/16.38/9.51*
DDMP	M	–	–	–	–/–/28.08	–/–/17.89	–/–/13.44	31.14/30.66/19.71	23.12/22.92/12.78	19.45/18.75/9.80
DD3D (Park等,2021)*	M	–	–	–	–/–/30.98	–/–/22.56	–/–/20.03	–/–/23.22	–/–/16.34	–/–/14.20
ROI-10D	M	89.04/–/–	88.39/–/–	78.77/–/–	10.74/–/–	7.46/–/–	7.06/–/–	7.79/–/–	5.16/–/–	3.95/–/–
ROI-10D (syn.)	M	85.32/–/75.33	77.32/–/69.64	69.70/–/61.18	14.50/–/16.77	9.91/–/12.40	8.73/–/11.39	9.61/–/12.30	6.63/–/10.30	6.29/–/9.39

注:“–”表示官方没有提供详细信息。

续表 16　部分 3D 目标检测算法性能表

Table 16　Performance of some 3D object detection algorithms (continued)

模型	模态	2D 目标检测			鸟瞰图3D目标检测			3D 目标检测		
		简单	中等	困难	简单	中等	困难	简单	中等	困难
Kinematic3D*	M2	–	–	–	27.83/–/26.69	19.72/–/17.52	15.10/–/13.10	19.76/–/19.07	14.10/–/12.72	10.47/–/9.17
Stereo R-CNN	S	98.53/–/–	88.27/–/–	71.14/–/–	68.50/–/61.67	48.30/–/43.87	41.47/–/36.44	54.11/–/49.23	36.69/–/34.05	31.07/–/28.39
RTS3D	S	–	–	–	77.50/–/72.17	58.65/–/51.79	50.14/–/43.19	64.76/–/58.51	46.70/–/37.38	39.27/–/31.12
DSGN	S	89.25/–/95.53	83.59/–/86.43	78.45/–/78.75	83.24/–/82.90	63.91/–/65.05	57.83/–/56.60	72.31/–/73.50	54.27/–/52.18	47.71/–/45.14
LIGA-Stereo	S	–/–/96.43*	–/–/93.82*	–/–/86.19*	89.35/–/88.15*	77.26/–/76.78*	69.05/–/67.40*	84.92/–/81.39*	67.06/–/64.66*	63.80/–/57.22*
SMOKE	M	–/–/92.88*	–/–/86.95*	–/–/77.04*	19.99/–/20.83*	15.61/–/14.49*	15.28/–/12.75*	14.76/–/14.03*	12.85/–/9.76*	11.50/–/7.84*
ZoomNet	S	–	–	–	78.68/–/72.94	66.19/–/54.91	57.60/–/44.14	62.96/–/55.98	50.47/–/38.64	43.63/–/30.97
PL++: AVOD	S	–	–	–	77.0/–/–	63.7/–/–	56.0/–/–	63.2/–/–	46.8/–/–	39.8/–/–
PL++: PIXOR	S	–	–	–	79.7/–/–	61.1/–/–	54.5/–/–	–	–	–
PL++: P-RCNN	S	–	–	–	82.0/–/–	64.0/–/–	57.3/–/–	67.9/–/–	50.1/–/–	45.3/–/–
AVOD	M	–	–	–	33.7/–/–	24.6/–/–	20.1/–/–	19.5/–/–	17.2/–/–	16.2/–/–
F-POINTNET	M	–	–	–	40.6/–/–	26.3/–/–	22.9/–/–	28.2/–/–	18.5/–/–	16.4/–/–
AVOD	S	–	–	–	74.9/–/–	56.8/–/–	49.0/–/–	61.9/–/–	45.3/–/–	39.0/–/–
F-POINTNET	S	–	–	–	72.8/–/–	51.8/–/–	44.0/–/–	59.4/–/–	39.8/–/–	33.5/–/–
PointPillars	PC	–	–	–	–/–/88.35	–/–/86.10	–/–/79.83	–/–/79.05	–/–/74.99	–/–/68.30
BtcDet*	PC	–	–	–	–	–	–	93.15/–/90.64	86.28/–/82.86	83.86/–/78.09
GLENet*	PC	–	–	–	–	–	–	93.51/–/91.67	86.10/–/83.23	83.60/–/78.43
VoTr	PC	–	–	–	–	–	–	89.04/–/89.90*	84.04/–/82.09*	78.68/–/79.14*
SPG	PC	–	–	–	94.09*/–/–	91.11*/–/–	88.86*/–/–	92.53*/–/90.50	85.31*/–/82.13	82.82*/–/78.90
SE-SSD*	PC	–	–	–	96.59/–/95.68	92.28/–/91.84	89.72/–/86.72	93.19/–/91.49	86.12/–/82.54	83.31/–/77.15
MonoCon	M	–	–	–	–/–/31.12*	–/–/22.10*	–/–/19.00*	26.33*/–/22.50*	19.01*/–/16.46*	15.89*/–/13.95*
MonoEF	M	–/–/96.32	–/–/90.88	–/–/83.27	–/–/29.03*	–/–/19.70*	–/–/17.26*	–/–/21.29*	–/–/13.87*	–/–/11.71*

注：“–”表示官方没有提供详细信息。

目前部分自动驾驶车辆已经在不同路况、天气和环境的复杂场景下实现高度的自动化，如百度Apollo自动驾驶车辆在雄安、上海多地完成了自动无人驾驶测试，小马智行在北京、广州等地完成了自动驾驶测试，谷歌Waymo自动驾驶车辆在美国凤凰城、旧金山等地区完成了自动无人驾驶测试。虽然部分车型已经具备了L4级别的自动驾驶能力，但是目前已经产业化的自动驾驶水平基本都在L2—L3级别，即辅助驾驶的水平，实现高度自动化仍任重道远。

4　结　语

本文回顾了基于图像的自动驾驶领域3D目标检测常用的数据集及基准，总结了目前制约该领域发展的一些主要因素及学者在此方面的研究，并对该领域涉及到的误差进行了分析。

现有的3D目标检测框架都是在特定的数据集

(KITTI、nuScenes、Waymo Open Dataset等)上训练并在该数据集的测试集上进行测试。然而,由于各个国家的道路设计、交通规则、驾驶习惯甚至车辆尺寸都有差异,所以我国的研究发展应该更加重视本土的道路场景,并在国内多种复杂场景中采集构建更加适宜的数据集。

研究中采用的评测指标大多沿袭2D目标检测的指标,往往是在一个确定数据集的测试效果。针对自动驾驶的落地应用,还面临更多的挑战,需要更多指向关键影响要素的评测指标,以适应复杂的应用场景。

目前,基于图像进行深度感知或预测仍是限制基于图像的3D目标检测算法性能的关键因素。由于使用的单目图像天然地缺少深度的信息,所以目前基于单目图像的3D目标检测性能仍然较差。除此之外,在实际操作过程中还会遇到物体大小变化、遮挡、自遮挡和截断的问题,这也是目前影响算法性能的一个重要因素。虽然使用立体视觉能够提高深度预测的准确性,但立体视觉在实际操作过程中仍会遇到特征匹配不准确导致视差估计不准的问题。如何提高从图像中进行深度预测的精度仍是一个亟待解决的难题和未来研究的主要方向。

致　谢　本文由中国图象图形学学会交通视频专业委员会组织撰写,该专委会链接为http://www.csig.org.cn/detail/2392。

参考文献(References)

Aghdam H H, Heravi E J, Demilew S S and Laganiere R. 2021. RAD: realtime and accurate 3D object detection on embedded systems// Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Nashville, USA: IEEE: 2869-2877 [DOI: 10.1109/CVPRW53098.2021.00322]

AutoX. 2021. AutoX releases the fifth generation Gen5 fully driverless system. Automobile Parts, (7): #7 (AutoX. 2021. AutoX发布第五代Gen5全无人驾驶系统. 汽车零部件, (7): #7)

Badki A, Troccoli A, Kim K, Kautz J, Sen P and Gallo O. 2020. Bi3D: stereo depth estimation via binary classifications//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 1597-1605 [DOI: 10.1109/CVPR42600.2020.00167]

Baidu. 2022. ApolloAuto/apollo [EB/OL]. [2022-10-26]. https://github.com/ApolloAuto/apollo

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. 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, Baldam G and Beijbom O. 2020. nuScenes: a multimodal dataset for autonomous driving//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 11621-11631 [DOI: 10.1109/CVPR42600.2020.01164]

Cao J C and Tao C B. 2021. An anchor-guided 3D target detection algorithm based on stereo RCNN. Chinese Journal of Scientific Instrument, 42(12): 191-201 (曹杰程, 陶重犇. 2021. 基于Stereo RCNN的锚引导3D目标检测算法. 仪器仪表学报, 42(12): 191-201) [DOI: 10.19650/j.cnki.cjsi.J2107801]

Chabot F, Chaouch M, Rabarisoa J, Teulière C and Chateau T. 2017. Deep MANTA: a coarse-to-fine many-task network for joint 2D and 3D vehicle analysis from monocular image//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 1827-1836 [DOI: 10.1109/CVPR.2017.198]

Chang M F, Lambert J, Sangkloy P, Singh J, Bak S, Hartnett A, Wang D, Carr P, Lucey S, Ramanan D and Hays J. 2019. Argoverse: 3D tracking and forecasting with rich maps//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 8740-8749 [DOI: 10.1109/CVPR.2019.00895]

Chen F, Wu F, Huang Q H, Feng Y J, Ge Q, Ji Y M, Hu C H and Jing X Y. 2020a. Semantic frustum based VoxelNet for 3D object detection//Proceedings of 2020 Chinese Automation Congress (CAC). Shanghai, China: IEEE: 7629-7634 [DOI: 10.1109/CAC51589.2020.9327549]

Chen N H. 2020. Challenging Tesla FSD, Baidu Apollo launches pilot assisted driving ANP. Business Observer, (12): 66-67 (陈念航. 2020. 挑战特斯拉FSD, 百度Apollo推出领航辅助驾驶ANP. 企业观察家, (12): 66-67)

Chen X Z, Kundu K, Zhu Y K, Berneshawi A, Ma H M, Fidler S and Urtasun R. 2015. 3D object proposals for accurate object class detection//Proceedings of the 28th International Conference on Neural Information Processing Systems. Montreal, Canada: MIT Press: 424-432

Chen X Z, Kundu K, Zhu Y K, Ma H M, Fidler S and Urtasun R. 2018. 3D object proposals using stereo imagery for accurate object class detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 40(5): 1259-1272 [DOI: 10.1109/TPAMI.2017.2706685]

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. Honolulu, USA: IEEE: 6526-6534 [DOI: 10.1109/CVPR.2017.691]

Chen Y J, Tai L, Sun K and Li M Y. 2020b. MonoPair: monocular 3D object detection using pairwise spatial relationships//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 12090-12099 [DOI: 10.1109/CVPR42600.2020.01211]

Chen Y L, Huang S J, Liu S, Yu B and Jia J Y. 2023. DSGN++: exploiting visual-spatial relation for stereo-based 3D detectors. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(4): 4416-4429 [DOI: 10.1109/TPAMI.2022.3197236]

Chen Y L, Liu S, Shen X Y and Jia J Y. 2020c. DSGN: deep stereo geometry network for 3D object detection//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 12533-12542 [DOI: 10.1109/CVPR42600.2020.01255]

Chen Y N, Dai H and Ding Y. 2022. Pseudo-stereo for monocular 3D object detection in autonomous driving//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 877-887 [DOI: 10.1109/CVPR52688.2022.00096]

Cheng Z M. 2022. Analysis of Tesla autopilot software system. For Repair and Maintenance, (1): 33-35 (程增木. 2022. 特斯拉自动驾驶软件系统解析. 汽车维修与保养, (1): 33-35) [DOI: 10.3969/j.issn.1008-3170.2022.01.010]

Chi J. 2021. Analysis of Google's patent technology for unmanned driving. Popular Standardization, (4): 162-164 (池娟. 2021. 关于Google无人驾驶的专利技术分析. 大众标准化, (4): 162-164) [DOI: 10.3969/j.issn.1007-1350.2021.04.053]

Chi X R, Pei W, Zhu Y Y, Wang C L, Shi L Y and Li J F. 2022. Fast Stereo-RCNN 3D target detection algorithm. Journal of Chinese Mini-Micro Computer Systems, 43(10): 2157-2161 (迟旭然, 裴伟, 朱永英, 王春立, 史良宇, 李锦峰. 2022. Fast Stereo-RCNN三维目标检测算法. 小型微型计算机系统, 43(10): 2157-2161) [DOI: 10.20009/j.cnki.21-1106/TP.2021-0167]

Cordts M, Omran M, Ramos S, Rehfeld T, Enzweiler M, Benenson R, Franke U, Roth S and Schiele B. 2016. The cityscapes dataset for semantic urban scene understanding//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 3213-3223 [DOI: 10.1109/CVPR.2016.350]

Deng J, Dong W, Socher R, Li L J, Li K and Li F F. 2009. ImageNet: a large-scale hierarchical image database//Proceedings of 2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami, USA: IEEE: 248-255 [DOI: 10.1109/CVPR.2009.5206848]

Ding M Y, Huo Y Q, Yi H W, Wang Z, Shi J P, Lu Z W and Luo P. 2020. Learning depth-guided convolutions for monocular 3D object detection//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 11669-11678 [DOI: 10.1109/CVPR42600.2020.01169]

Dong W B and Isler V. 2020. Ellipse regression with predicted uncertainties for accurate multi-view 3D object estimation. [EB/OL]. [2023-01-01]. https://arxiv.org/pdf/2101.05212.pdf

Dou J, Xue J R and Fang J W. 2019. SEG-VoxelNet for 3D vehicle detection from RGB and LiDAR data//Proceedings of 2019 International Conference on Robotics and Automation (ICRA). Montreal, Canada: IEEE: 4362-4368 [DOI: 10.1109/ICRA.2019.8793492]

Eigen D, Puhrsch C and Fergus R. 2014. Depth map prediction from a single image using a multi-scale deep network [EB/OL]. [2023-01-01]. https://arxiv.org/pdf/1406.2283.pdf

Everingham M, van Gool L, Williams C K I, Winn J and Zisserman A. 2010. The pascal visual object classes (VOC) challenge. International Journal of Computer Vision, 88(2): 303-338 [DOI: 10.1007/s11263-009-0275-4]

Feng Z Y, Jing L D, Yin P, Tian Y L and Li B. 2021. Advancing self-supervised monocular depth learning with sparse LiDAR [EB/OL]. [2023-01-01]. https://arxiv.org/pdf/2109.09628.pdf

Gao T Z, Pan H H and Gao H J. 2022. Monocular 3D object detection with sequential feature association and depth hint augmentation. IEEE Transactions on Intelligent Vehicles, 7(2): 240-250 [DOI: 10.1109/TIV.2022.3143954]

Garg D, Wang Y, Hariharan B, Campbell M, Weinberger K Q and Chao W L. 2021. Wasserstein distances for stereo disparity estimation [EB/OL]. [2023-01-01]. https://arxiv.org/pdf/2007.03085.pdf

Geiger A, Lenz P, Stiller C and Urtasun R. 2022. The KITTI vision benchmark suite [EB/OL]. [2022-10-26]. https://www.cvlibs.net/datasets/kitti/index.php

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, 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]

Girshick R. 2015. Fast R-CNN//Proceedings of 2015 IEEE International Conference on Computer Vision. Santiago, Chile: IEEE: 1440-1448 [DOI: 10.1109/ICCV.2015.169]

Guo W J. 2021. From manned testing tono safety officers, unmanned taxis are gradually approaching. Intelligent Connected Vehicles, (1): 21-25 (郭文佳. 2021. 从载人测试到取消安全员无人驾驶出租车渐行渐近. 智能网联汽车, (1): 21-25)

Guo X Y, Shi S S, Wang X G and Li H S. 2021. LIGA-Stereo: learning

LiDAR geometry aware representations for stereo-based 3D detector//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 3133-3143 [DOI: 10.1109/ICCV48922.2021.00314]

Hong Y, Dai H and Ding Y. 2022. Cross-modality knowledge distillation network for monocular 3D object detection//Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer: 87-104 [DOI: 10.1007/978-3-031-20080-9_6]

Houston J, Zuidhof G, Bergamini L, Ye Y W, Chen L, Jain A, Omari S, Lglovikov V and Ondruska P. 2020. One thousand and one hours: self-driving motion prediction dataset [EB/OL]. [2023-01-01]. https://arxiv.org/pdf/2006.14480.pdf

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, Walsh S and Waslander S L. 2019a. Improving 3D object detection for pedestrians with virtual multi-view synthesis orientation estimation//Proceedings of 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Macau, China: IEEE: 3459-3466 [DOI: 10.1109/IROS40897.2019.8968242]

Ku J, Pon A D and Waslander S L. 2019b. Monocular 3D object detection leveraging accurate proposals and shape reconstruction//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 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. Long Beach, USA: IEEE: 12689-12697 [DOI: 10.1109/CVPR.2019.01298]

Li B Y, Ouyang W L, Sheng L, Zeng X Y and Wang X G. 2019a. GS3D: an efficient 3D object detection framework for autonomous driving//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 1019-1028 [DOI: 10.1109/CVPR.2019.00111]

Li H R, Duan Z C, Ma M J, Chen Y R, Li J Q and Zhao D B. 2021a. MVM3Det: a novel method for multi-view monocular 3D detection. [EB/OL]. [2023-01-01]. https://arxiv.org/pdf/2109.10473.pdf

Li L. 2022. Toyota "defected" to Tesla. Automotive Observer, (4): 14-15 (李琳. 2022. 丰田"投靠"特斯拉. 汽车观察, (4): 14-15) [DOI: 10.3969/j.issn.1673-145X.2022.04.005]

Li P L, Chen X Z and Shen S J. 2019b. Stereo R-CNN based 3D object detection for autonomous driving//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 7636-7644 [DOI: 10.1109/CVPR.2019.00783]

Li P X, Su S and Zhao H C. 2021b. RTS3D: real-time stereo 3D detection from 4D feature-consistency embedding space for autonomous driving. Proceedings of 2021 AAAI Conference on Artificial Intelligence, 35(3): 1930-1939 [DOI: 10.1609/aaai.v35i3.16288]

Li P X, Zhao H C, Liu P F and Cao F D. 2020. RTM3D: real-time monocular 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]

Liao Y Y, Xie J and Geiger A. 2023. KITTI-360: a novel dataset and benchmarks for urban scene understanding in 2D and 3D. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(3): 3292-3310 [DOI: 10.1109/TPAMI.2022.3179507]

Liu A Z. 2022. Ideal ONE: enhanced by intelligent driving. Intelligent and Connected Vehicles, (1): 91-93 (刘岸泽. 2022. 理想ONE: 智能驾驶加持. 智能网联汽车, (1): 91-93)

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 P, Xue N and Wu T F. 2022. Learning auxiliary monocular contexts helps monocular 3D object detection. Proceedings of 2022 AAAI Conference on Artificial Intelligence, 36(2): 1810-1818 [DOI: 10.1609/aaai.v36i2.20074]

Liu Y X, Wang L J and Liu M. 2021a. YOLOStereo3D: a step back to 2D for efficient stereo 3D detection//Proceedings of 2021 IEEE International Conference on Robotics and Automation (ICRA). Xi'an, China: IEEE: 13018-13024 [DOI: 10.1109/ICRA48506.2021.9561423]

Liu Y X, Yuan Y X and Liu M. 2021b. Ground-aware monocular 3D object detection for autonomous driving. IEEE Robotics and Automation Letters, 6(2): 919-926 [DOI: 10.1109/LRA.2021.3052442]

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 Recognition Workshops. Seattle, USA: IEEE: 4289-4298 [DOI: 10.1109/CVPRW50498.2020.00506]

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 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Brighton, UK: IEEE: 1992-1996 [DOI: 10.1109/ICASSP.2019.8682746]

Luo S J, Dai H, Shao L and Ding Y. 2021. M3DSSD: monocular 3D single stage object detector//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 6141-6150 [DOI: 10.1109/CVPR46437.2021.00608]

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. Seoul, Korea (South): IEEE: 6850-6859 [DOI: 10.1109/ICCV.2019.00695]

Ma X Z, Zhang Y M, Xu D, Zhou D Z, Yi S, Li H J and Ouyang W L. 2021. Delving into localization errors for monocular 3D object detection//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 4719-4728 [DOI: 10.1109/CVPR46437.2021.00469]

Mao J G, Shi S S, Wang X G and Li H S. 2022. 3D object detection for autonomous driving: a review and new outlooks [EB/OL]. [2023-01-01]. https://arxiv.org/pdf/2206.09474.pdf

Mao J G, Xue Y J, Niu M Z, Bai H Y, Feng J S, Liang X D, Xu H and Xu C J. 2021. Voxel transformer for 3D object detection//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 3144-3153 [DOI: 10.1109/ICCV48922.2021.00315]

Meng X. 2021. Meng Xing: self-evolution of Didi autonomous driving. Intelligent and Connected Vehicles, (3): 42-44 (孟醒. 2021. 孟醒: 滴滴自动驾驶的自我进化. 智能网联汽车, (3): 42-44)

Mousavian A, Anguelov D, Flynn J and Košecká J. 2017. 3D bounding box estimation using deep learning and geometry//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 5632-5640 [DOI: 10.1109/CVPR.2017.597]

Nabati R and Qi H R. 2021. CenterFusion: center-based radar and camera fusion for 3D object detection//Proceedings of 2021 IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa, USA: IEEE: 1526-1535 [DOI: 10.1109/WACV48630.2021.00157]

Novák L. 2017. Vehicle Detection and Pose Estimation for Autonomous Driving. Prague: Czech Technical University in Prague

Paperswithcode. 2023a. 3D object detection from stereo images on KITTI cars moderate [EB/OL]. [2023-01-30]. https://paperswithcode.com/sota/3d-object-detection-from-stereo-images-on-1

Paperswithcode. 2023b. Monocular 3D object detection on KITTI cars Moderate [EB/OL]. [2023-01-30]. https://paperswithcode.com/sota/monocular-3d-object-detection-on-kitti-cars

Park D, Ambruş R, Guizilini V, Li J and Gaidon A. 2021. Is pseudo-lidar needed for monocular 3D object detection?//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 3122-3132 [DOI: 10.1109/ICCV48922.2021.00313]

Patil A, Malla S, Gang H and Chen Y T. 2019. The H3D dataset for full-surround 3D multi-object detection and tracking in crowded urban scenes//Proceedings of 2019 International Conference on Robotics and Automation (ICRA). Montreal, Canada: IEEE: 9552-9557 [DOI: 10.1109/ICRA.2019.8793925]

Peng W L, Pan H, Liu H and Sun Y. 2020. IDA-3D: instance-depth-aware 3D object detection from stereo vision for autonomous driving//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 13012-13021 [DOI: 10.1109/CVPR42600.2020.01303]

Pony.ai. 2022. Technology [EB/OL]. [2022-10-26]. https://www.pony.ai/tech?lang=en (小马智行. 2022. 核心技术)[EB/OL]. [2022-10-26]. https://www.pony.ai/tech?lang=zh

Qi C R, Su H, Mo K and Guibas L J. 2017a. PointNet: deep learning on point sets for 3D classification and segmentation//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 77-85 [DOI: 10.1109/CVPR.2017.16]

Qi C R, Yi L, Su H and Guibas L J. 2017b. 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.: 5100-5109

Qian R, Garg D, Wang Y, You Y, Belongie S, Hariharan B, Campbell M, Weinberger K Q and Chao W L. 2020. End-to-end pseudo-LiDAR for image-based 3D object detection//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 5880-5889 [DOI: 10.1109/CVPR42600.2020.00592]

Qin C, Wang Y F, Zhang Y C and Yin C L. 2022. 3D object detection based on extremely sparse laser point cloud and RGB images. Laser and Optoelectronics Progress, 59(18): 447-458 (秦超, 王亚飞, 张宇超, 殷承良. 2022. 基于极端稀疏激光点云和RGB图像的3D目标检测. 激光与光电子学进展, 59(18): 447-458) [DOI: 10.3788/LOP202259.1828004]

Qin Z Y, Wang J L and Lu Y. 2019a. MonoGRNet: a geometric reasoning network for monocular 3D object localization. Proceedings of the AAAI Conference on Artificial Intelligence, 33(1): 8851-8858 [DOI: 10.1609/aaai.v33i01.33018851]

Qin Z Y, Wang J L and Lu Y. 2019b. Triangulation learning network: from monocular to stereo 3D object detection//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 7607-7615 [DOI: 10.1109/CVPR.2019.00780]

Reading C, Harakeh A, Chae J and Waslander S L. 2021. Categorical depth distribution network for monocular 3D object detection//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 8551-8560 [DOI: 10.1109/CVPR46437.2021.00845]

Redmon J, Divvala S, Girshick R and Farhadi A. 2016. You only look once: unified, real-time object detection//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 779-788 [DOI: 10.1109/CVPR.2016.91]

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. 2018. Orthographic feature transform for monocular 3D object detection [EB/OL]. [2023-01-01]. https://arxiv.org/pdf/1811.08188.pdf

Ruan X G, Yan W J, Huang J and Guo P Y. 2022. Monocular depth estimation method based on dual-discriminator generative adversarial networks. Journal of Beijing University of Technology, 48(9): 928-934 (阮晓钢, 颜文静, 黄静, 郭佩远. 2022. 基于双鉴别器生成对抗网络的单目深度估计方法. 北京工业大学学报, 48(9): 928-934) [DOI: 10.11936/bjutxb2021050001]

Simonelli A, Buló S R, Porzi L, Ricci E and Kontschieder 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]

Song X B, Wang P, Zhou D F, Zhu R, Guan C Y, Dai Y C, Su H, Li H D and Yang R G. 2019. ApolloCar3D: a large 3D car instance understanding benchmark for autonomous driving//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 5447-5457 [DOI: 10.1109/CVPR.2019.00560]

Su K Q, Yan W Q and Xu J D. 2022. 3D object detection based on multi-path feature pyramid network for stereo images. Journal of Beijing University of Aeronautics and Astronautics, 48(8): 1487-1494 (苏凯祺, 阎维青, 徐金东. 2022. 基于立体图像的多路径特征金字塔网络3D目标检测. 北京航空航天大学学报, 48(8): 1487-1494) [DOI: 10.13700/j.bh.1001-5965.2021.0525]

Sun P, Kretzschmar H, Dotiwalla 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. Seattle, USA: IEEE: 2443-2451 [DOI: 10.1109/CVPR42600.2020.00252]

Talpes E, Sarma D D, Venkataramanan G, Bannon P, McGee B, Floering B, Jalote A, Hsiong C, Arora S, Gorti A and Sachdev G S. 2020. Compute solution for Tesla's full self-driving computer. IEEE Micro, 40(2): 25-35 [DOI: 10.1109/mm.2020.2975764]

Wang K R, Tan J G, Du Q, Chen L L, Li J M and Zhang X L. 2020. 3D object detection based on iterative self-training. Acta Optica Sinica, 40(9): 133-145 (王康如, 谭锦钢, 杜量, 陈利利, 李嘉茂, 张晓林. 2020. 基于迭代式自主学习的三维目标检测. 光学学报, 40(9): 133-145) [DOI: 10.3788/AOS202040.0915005]

Wang L, Du L, Ye X Q, Fu Y W, Guo G D, Xue X Y, Feng J F and Zhang L. 2021a. Depth-conditioned dynamic message propagation for monocular 3D object detection//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 454-463 [DOI: 10.1109/CVPR46437.2021.00052]

Wang Q C, Shuai H and Liu Q S. 2022. Monocular accumulated depth estimation with recursive feature fusion. Journal of Computer-Aided Design and Computer Graphics, 34(10): 1533-1541 (王秋晨, 帅惠, 刘青山. 2022. 递归特征融合的单目深度累积估计. 计算机辅助设计与图形学学报, 34(10): 1533-1541) [DOI: 10.3724/SP.J.1089.2022.19728]

Wang Q D, Wang Q K, Cheng K and Liu Z H. 2022. Monocular depth estimation with enhanced edge. Journal of Huazhong University of Science and Technology (Natural Science Edition), 50(3): 36-42 (王泉德, 王奇坤, 程凯, 刘子航. 2022. 强化边缘的单目图像深度估计. 华中科技大学学报(自然科学版), 50(3): 36-42) [DOI: 10.13245/j.hust.220307]

Wang T, Zhu X G, Pang J M and Lin D H. 2021b. 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. 2019. 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. Long Beach, USA: IEEE: 8437-8445 [DOI: 10.1109/CVPR.2019.00864]

Wang Y, Yang B, Hu R, Liang M and Urtasun R. 2021c. PLUMENet: efficient 3D object detection from stereo images//Proceedings of 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Prague, Czech Republic: IEEE: 3383-3390 [DOI: 10.1109/IROS51168.2021.9635875]

Wang Y Q and Tao Y. 2022. Research on 3D object detection algorithm based on binocular vision. Microelectronics and Computer, 39(2): 19-25 (王一强, 陶洋. 2022. 基于双目视觉的三维目标检测算法研究. 微电子学与计算机, 39(2): 19-25) [DOI: 10.19304/j.issn1000-7180.2021.0730]

Weng X S and Kitani K. 2019. Monocular 3D object detection with pseudo-lidar point cloud//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision Workshop. Seoul, Korea (South): IEEE: 857-866 [DOI: 10.1109/ICCVW.2019.00114]

WorldAuto. 2022. 2022 WIDC officially ends-the new HI edition of Polar Fox Alpha S won two gold awards. WorldAuto, (7): 60-63 (WorldAuto. 2022. 2022 WIDC正式落幕 极狐阿尔法S全新HI版荣获两大项金奖. 世界汽车, (7): 60-63) [DOI: 10.3969/j.issn.1005-9008.2022.07.010]

Xiang Y, Choi W, Lin Y Q and Savarese S. 2017. Subcategory-aware convolutional neural networks for object proposals and detection//Proceedings of 2017 IEEE Winter Conference on Applications of Computer Vision (WACV). Santa Rosa, USA: IEEE: 924-933 [DOI: 10.1109/WACV.2017.108]

Xu Q G, Zhong Y Q and Neumann U. 2022. Behind the curtain: learning occluded shapes for 3D object detection. Proceedings of the AAAI Conference on Artificial Intelligence, 36(3): 2893-2901 [DOI: 10.1609/aaai.v36i3.20194]

Xu Q G, Zhou Y, Wang W Y, Qi C R and Anguelov D. 2021. SPG: unsupervised domain adaptation for 3D object detection via semantic point generation//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 15426-15436 [DOI: 10.1109/ICCV48922.2021.01516]

Xu Z B, Zhang W, Ye X Q, Tan X, Yang W, Wen S L, Ding E R, Meng A J and Huang L S. 2020. ZoomNet: part-aware adaptive zooming neural network for 3D object detection. Proceedings of the AAAI Conference on Artificial Intelligence, 34(7): 12557-12564 [DOI: 10.1609/aaai.v34i07.6945]

Yan J, Fang Z J and Gao Y B. 2020. 3D object detection based on domain attention and dilated convolution. Journal of Image and Graphics, 25(6): 1221-1234 (严娟, 方志军, 高永彬. 2020. 结合混合域注意力与空洞卷积的3维目标检测. 中国图象图形学报, 25(6): 1221-1234) [DOI: 10.11834/jig.190378]

Yang B Y, Du X P, Fang Y Q, Li P Y and Wang Y. 2021. Review of rigid object pose estimation from a single image. Journal of Image and Graphics, 26(2): 334-354 (杨步一, 杜小平, 方宇强, 李佩阳, 王阳. 2021. 单幅图像刚体目标姿态估计方法综述. 中国图象图形学报, 26(2): 334-354) [DOI: 10.11834/jig.200037]

Yang H T, Lei L and Lin Y C. 2022. Binocular depth estimation algorithm based on multi-scale attention feature fusion. Laser and Optoelectronics Progress, 59(18): 259-267 (杨蕙同, 雷亮, 林永春. 2022. 基于多尺度注意力特征融合的双目深度估计算法. 激光与光电子学进展, 59(18): 259-267) [DOI: 10.3788/LOP202259.1815005]

You Y R, Wang Y, Chao W L, Garg D, Pleiss G, Hariharan B, Campbell M and Weinberger K Q. 2020. Pseudo-LiDAR++: accurate depth for 3D object detection in autonomous driving [EB/OL]. [2023-01-01]. https://arxiv.org/pdf/1906.06310.pdf

Yu H B, Luo Y Z, Shu M, Huo Y Y, Yang Z B, Shi Y F, Guo Z L, Li H Y, Hu X, Yuan J R and Nie Z Q. 2022. DAIR-V2X: a large-scale dataset for vehicle-infrastructure cooperative 3D object detection//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 21329-21338 [DOI: 10.1109/CVPR52688.2022.02067]

Yu J X, Zhang M Q and Su Y T. 2021. Three-dimensional vehicle detection algorithm based on binocular vision. Laser and Optoelectronics Progress, 58(2): 301-306 (于洁潇, 张美琪, 苏育挺. 2021. 基于双目视觉的三维车辆检测算法. 激光与光电子学进展, 58(2): 301-306) [DOI: 10.3788/LOP202158.0215004]

Zhang C, Ma Y X, Wan J W, Xu K and Xu G Q. 2022. Multi-scale monocular depth estimation network based on channel attention. Journal of Signal Processing, 38(11): 2332-2341 (张聪, 马燕新, 万建伟, 许可, 徐国权. 2022. 基于通道注意力机制的单目深度估计. 信号处理, 38(11): 2332-2341) [DOI: 10.16798/j.issn.1003-0530.2022.11.010]

Zhang J L, Wei M and Wen W. 2022. Monocular depth estimation based on DSPP. Application Research of Computers, 39(12): 3837-3840 (张竞澜, 魏敏, 文武. 2022. 基于DSPP的单目图像深度估计. 计算机应用研究, 39(12): 3837-3840) [DOI: 10.19734/j.issn.1001-3695.2022.05.0212]

Zhang J N, Su Q X, Liu P Y, Gu H Q and Wang W. 2020. A monocular 3D target detection network with perspective projection. Robot, 42(3): 278-288 (张峻宁, 苏群星, 刘鹏远, 谷宏强, 王威. 2020. 一种基于透视投影的单目3D目标检测网络. 机器人, 42(3): 278-288) [DOI: 10.13973/j.cnki.robot.190221]

Zhang Y F, Li Y X, Zhao M B, Yu X Y, Zhan Y L and Lin W Y. 2021. Object distance estimation based on stereo regional disparity regression. Journal of Image and Graphics, 26(7): 1604-1613 (张羽丰, 李昱希, 赵明璧, 喻晓源, 占云龙, 林巍峣. 2021. 局部双目视差回归的目标距离估计. 中国图象图形学报, 26(7): 1604-1613) [DOI: 10.11834/jig.200511]

Zhang Y F, Zhang Q J, Zhu Z Y, Hou J H and Yuan Y X. 2022. GLENet: boosting 3D object detectors with generative label uncertainty estimation [EB/OL]. [2023-01-01]. https://arxiv.org/pdf/2207.02466.pdf

Zhang Y P, Lu J W and Zhou J. 2021. Objects are different: flexible monocular 3D object detection//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 3288-3297 [DOI: 10.1109/CVPR46437.2021.00330]

Zhao H Q, Fang Z J and Gao Y B. 2019. Prior direction angle estimation in 3D object detection. Transducer and Microsystem Technologies, 38(6): 35-38 (赵华卿, 方志军, 高永彬. 2019. 三维目标检测中的先验方向角估计. 传感器与微系统, 38(6): 35-38) [DOI: 10.13873/J.1000-9787(2019)06-0035-04]

Zhao X, Liang H R and Liang R H. 2019. Combining object detection and binocular vision for 3D car pose estimation. Journal of Computer-Aided Design and Computer Graphics, 31(9): 1518-1527 (赵邢, 梁浩然, 梁荣华. 2019. 结合目标检测与双目视觉的三维车辆姿态检测. 计算机辅助设计与图形学学报, 31(9): 1518-1527) [DOI: 10.3724/SP.J.1089.2019.17625]

Zheng W, Tang W L, Jiang L and Fu C W. 2021. SE-SSD: self-ensembling single-stage object detector from point cloud//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 14489-14498 [DOI: 10.1109/CVPR46437.2021.01426]

Zhou D K, Tian J and Yang X. 2021. Unsurpervised monocular image depth estimation based on the prediction of local plane parameters. Journal of Image and Graphics, 26(1): 165-175 (周大可, 田径, 杨欣. 2021. 结合局部平面参数预测的无监督单目图像深度估计. 中国图象图形学报, 26(1): 165-175) [DOI: 10.11834/jig.200364]

Zhou X Y, Wang D Q and Krähenbühl P. 2019. Objects as points [EB/OL]. [2023-01-01]. https://arxiv.org/pdf/1904.07850.pdf

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 Z Y, Du L, Ye X Q, Zou Z K, Tan X, Zhang L, Xue X Y and Feng J F. 2022. SGM3D: stereo guided monocular 3D object detection. IEEE Robotics and Automation Letters, 7(4): 10478-10485 [DOI: 10.1109/LRA.2022.3191849]

作者简介

李熙莹,女,教授,博士生导师,主要研究方向为视频图像交通信息处理、视频大数据等技术与应用。
E-mail: stslxy@mail.sysu.edu.cn

韦世奎,通信作者,男,教授,博士生导师,主要研究方向为跨媒体智能、计算机视觉、机器学习和智能交通。
E-mail: shkwei@bjtu.edu.cn

叶芝桧,男,硕士研究生,主要研究方向为3D目标检测与3维重建。E-mail: 2020427933@qq.com

陈泽,男,本科生,主要研究方向为智能交通工程。
E-mail: 1149007952@qq.com

陈小彤,女,博士研究生,主要研究方向为跨模态融合及3D目标检测。E-mail: 22110091@bjtu.edu.cn

田永鸿,男,教授,主要研究方向为视频大数据分析与处理、视觉神经信息编码与识别、机器学习与计算机视觉。
E-mail: yhtian@pku.edu.cn

党建武,男,教授,主要研究方向为交通信息工程及控制、智能信息处理。E-mail: dangjw@mail.lzjtu.cn

付树军,男,教授,主要研究方向为图像处理和计算机视觉、医学图像分析和大数据计算。E-mail: shujunfu@163.com

赵耀,男,教授,主要研究方向为图像编码、数字水印、跨媒体内容分析与内容理解、多媒体信息处理。
E-mail: yzhao@bjtu.edu.cn