计算机科学
COMPUTER SCIENCE

自动驾驶场景下的图像三维目标检测研究进展

周燕, 许业文, 蒲磊, 徐雪妙, 刘翔宇, 周月霞

引用本文

周燕, 许业文, 蒲磊, 徐雪妙, 刘翔宇, 周月霞. 自动驾驶场景下的图像三维目标检测研究进展[J]. 计算机科学, 2024, 51(11): 133-147.

ZHOU Yan, XU Yewen, PU Lei, XU Xuemiao, LIU Xiangyu, ZHOU Yuexia. Research Progress of Image 3D Object Detection in Autonomous Driving Scenario [J]. Computer Science, 2024, 51(11): 133-147.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.231000075

自动驾驶场景下的图像三维目标检测研究进展

周　燕[1,2]　许业文[1]　蒲　磊[1]　徐雪妙[2]　刘翔宇[1]　周月霞[1]
1 佛山大学电子信息工程学院　广东 佛山 528000
2 华南理工大学计算机科学与工程学院　广州 510641
(zhouyan791266@fosu.edu.cn)

摘　要　二维目标检测技术由于缺乏对物理世界尺寸、深度等信息的描述,在自动驾驶场景中应用还存在较大的局限性。许多研究者结合自动驾驶实际需要,在图像三维目标检测上做了许多探索。为了对该领域进行全面研究,文中对近年来国内外发表的相关文献进行综述,介绍了基于图像的三维目标检测以及图像与点云融合的三维目标检测两类方法,并根据网络对输入数据的不同处理方式,对两类方法进一步细分,阐述了各个类别中的代表性方法,对各类方法的优劣进行总结,对比并分析了各算法的性能。此外,详细介绍了自动驾驶场景下三维目标检测的相关数据集和评价指标。最后,对图像三维目标检测领域中存在的挑战和困难进行了分析,并对未来可能的研究方向进行了展望。

关键词:图像三维目标检测;深度学习;自动驾驶;多模态融合;计算机视觉
中图分类号　TP391

Research Progress of Image 3D Object Detection in Autonomous Driving Scenario

ZHOU Yan[1,2],XU Yewen[1],PU Lei[1],XU Xuemiao[2],LIU Xiangyu[1] and ZHOU Yuexia[1]
1 School of Electronic Information Engineering,Foshan University,Foshan,Guangdong 528000,China
2 School of Computer Science and Engineering,South China University of Technology,Guangzhou 510641,China

Abstract　2D object detection techniques have significant limitations when applied to automatic driving scenarios due to the absence of description of the size,depth and other information of the physical environment. Numerous researchers have made extensive explorations in the field of image 3D object detection by aligning with the practical requirements of automatic driving. To conduct a comprehensive study in this domain,this paper reviews recent literature published both domestically and internationally. It introduces two main categories of methods:image-based 3D object detection and 3D object detection by fusing image and point cloud data. Furthermore,it further subdivides these categories based on the different approaches used to process input data by the network. The paper describes representative methods within each category,summarizes the strengths and weaknesses of each method,and conducts a comparative analysis of their performance. Additionally,it provides a detailed introduction to relevant datasets and evaluation metrics for 3D object detection in autonomous driving scenarios. Finally,the paper analyzes the challenges and difficulties in the field of image 3D object detection,and outlines potential future research directions.

Keywords　Image 3D object detection,Deep learning,Automatic driving,Multimodal fusion,Computer vision

1　引言

目标检测作为计算机视觉领域中的基本任务,一直是研究的热点,在自动驾驶中的车辆检测[1]、智能机器人的自动巡航及人脸识别[2]等诸多领域有着广泛的应用。近年来,得益于深度学习理论与技术的快速发展,二维目标检测算法在实时性与准确性上都有较大的突破[3-5]。然而,二维目标检测仅在图像中回归目标的像素坐标,缺乏描述物理世界的尺寸、深度等信息,因此在许多需要感知物理世界的实际应用中还存在局限性。

为了满足自动驾驶、智能机器人等领域的迫切需要,研究者们提出了基于深度学习的三维目标检测相关算法。三维目标检测旨在感知三维空间中的目标,包含定位与识别两个步骤(即在三维空间中对目标进行定位,输出三维边界框,并对

到稿日期:2023-10-12　返修日期:2024-03-28
基金项目:国家自然科学基金(61972091);广东省自然科学基金(2022A1515010101,2021A1515012639);广东省普通高校重点研究项目(2020ZDZX3049);佛山市科技创新项目(2020001003285);广东省教育科学规划课题(2021GXJK445)
This work was supported by the National Natural Science Foundation of China(61972091),Natural Science Foundation of Guangdong Province,China(2022A1515010101,2021A1515012639),Key Research Project of Universities of Guangdong Province(2020ZDZX3049),Science and Technology Innovation Project of FoShan(2020001003285) and Educational Science Planning Project of Guangdong Province,China(2021GXJK445).
通信作者:许业文(2112203032@stu.fosu.edu.cn)

三维边界框内的目标进行类别的判定)。与二维目标检测相比,三维目标检测能获得目标在空间中的位置、尺寸及朝向等几何信息,能更好地感知物理世界,是三维场景感知和理解的基础任务。根据传感器的配置,在实际应用中可以将三维目标检测算法分为基于点云的三维目标检测、基于图像的三维目标检测,以及图像与点云融合的三维目标检测。

一般情况下,激光雷达传感器通过发射激光束,并对三维空间中目标反射回来的信号进行处理,从而获得点云数据。基于点云的三维目标检测方法中较为经典的研究有 VoxelNet[6],PointPillar[7]和 PointRCNN[8]等,这些方法通过直接或间接的方式提取点云特征,并根据提取的特征完成空间中目标的定位与识别。与图像数据相比,点云数据缺乏纹理及颜色特征,且具有稀疏和无序的性质,但它能够提供准确的空间位置信息,因此目前基于点云的目标检测相关算法能实现较高的识别准确率。

然而在实际的自动驾驶应用中,激光雷达传感器由于制造成本较高,其普及受到限制,又得益于深度学习网络在二维目标检测的长足进步,部分研究人员开始转向价格更为低廉的相机传感器,希望通过图像数据来完成对三维目标的感知。但图像数据缺乏深度信息且易受光照条件的影响,深度值计算存在偏差,从而容易导致对目标在空间中的定位误差较大。在实际的自动驾驶测试场景下,安全性是首要考虑的因素。点云数据能够提供准确的三维空间位置信息,而图像数据具有丰富的颜色和纹理特征,因此,一些研究人员将研究的关注点转向对图像与点云数据融合的三维目标检测,将点云数据与图像数据的优势相结合,以实现高精度的三维目标感知。但如何有效地将点云与图像这两种异构的数据进行融合仍需要进一步研究。

本文系统总结了图像三维目标检测领域的研究进展,阐述了基于图像的三维目标检测以及图像与点云融合的三维目标检测两类方法,并对这两类方法进行了新的合理分类。具体来说,在基于图像的三维目标检测方法中,本文将其划分为直接预测的方法、特征升维的方法以及数据升维的方法。直接预测的方法指模型将从图像中提取的二维特征直接用于检测;特征升维的方法指将从图像中提取的特征转化为高维特征或高维特征的变体(如体素特征或鸟瞰图特征)后再进行检测;数据升维的方法指将图像转化为三维数据(即伪点云或者伪点云的变体)后再进行检测。在图像与点云融合的三维目标检测方法中,本文将其划分为特征级融合的方法以及决策级融合的方法。特征级融合的方法指对图像数据和点云数据本身或提取的特征进行融合后再进行检测;决策级融合的方法指两种模态数据先分别进行三维目标检测,然后将两者的检测结果融合。本文通过这一新的划分方式,对图像三维目标检测文献进行清晰归类。

目前已有相关综述对三维目标检测领域进行了梳理总结。Qian 等[9]对三维目标检测方法的各个分支做了较为详尽的综述,包括图像、点云及多种模态数据下的三维目标检测方法,其将基于图像的三维目标检测方法分为基于结果提升的方法和基于特征提升的方法,但这种划分方式不够细致;Cao 等[10]对近年来基于深度学习的视觉目标检测技术进行了详细的综述,主要聚焦于二维目标检测及基于双目的三维目标检测,但其对基于图像的三维目标检测的分析总结还不够完整;文献[11-12]以自动驾驶为背景,仅针对点云数据进行三维目标检测相关方法的综述。现有专门针对自动驾驶场景下的图像三维目标检测的讨论与分析工作还比较少,对图像三维目标检测在自动驾驶领域未来的发展及展望还不够全面。因此,本文专门针对图像三维目标检测在自动驾驶领域的发展进行综述,并在前人的基础上制定了清晰的分类方式。

与前人的工作相比,本文的主要贡献在于:

1)涵盖最新的代表性工作,聚焦于自动驾驶场景下的图像三维目标检测。

2)对图像三维目标检测方法重新进行合理分类,并介绍了不同类别下最具代表性的方法。通过与同类方法及不同类方法的对比,介绍了图像三维目标检测领域目前的发展现状。

3)对自动驾驶场景下的图像三维目标检测所面临的挑战进行分析,并以此引出该领域未来的发展方向,为后续的研究者提供研究思路。

2 相关数据集和评价指标

2.1 相关数据集

自动驾驶场景下的常用数据集主要有:

1)KITTI[13]数据集:目前三维目标检测领域中最早公开的大型自动驾驶数据集。该数据集包含了 7481 个图像训练数据和 7518 个图像测试数据,并且只提供了训练数据的标注,而测试数据的评估仅在官方测试服务器上可用,因此在深度学习网络模型训练阶段通常将训练数据划分为训练集与验证集进行模型性能评估。KITTI 数据集提供了约 50 个场景、超过 8 个类别的标注信息,但只有“汽车”“行人”和“自行车”类别被纳入性能评估的范围,并且该数据集根据目标二维标注框的高度、遮挡程度及截断程度将检测难度划分为简单、中等和困难 3 个等级。

2)nuScenes[14]数据集:第一个提供全套自动汽车传感器数据的大型自动驾驶数据集,包括激光点云数据、图像数据以及毫米波雷达数据等。该数据集拥有 1000 个场景数据,包括 700 个训练场景、150 个验证场景及 150 个测试场景。每个场景获取约 20 s 的数据,共有 40 个关键帧。数据集提供了 23 个类别的标注,不仅包括二维及三维标注框,还标注了目标的可见性、运动和姿态。

3)Waymo Open[15]数据集:由自动驾驶公司 Waymo 发布,它使用 5 个同步的相机对当前场景进行 360 度环视拍摄。Waymo Open 数据集提供了 12.22 万帧训练数据、3 万帧验证数据及 4 万帧测试数据。整个数据集包含 1150 个场景,共有大约 2500 万个三维边界框、2200 万个二维边界框。此外,在数据集多样性上,Waymo Open 数据集也有很大的提升,该数据集涵盖了不同的天气和道路条件,包括白天和夜晚不同的时间段、市中心和郊区的不同地点、行人和自行车等不同的道路对象等。

KITTI 作为自动驾驶场景下三维目标检测领域数据集的先驱,在该领域的发展中起到了不可替代的作用,但其仅在晴天光照良好的情况下进行数据的采集,没有考虑到照明和

天气条件的影响，并且数据规模相对较小，在该数据集上训练的模型在极端天气和照明不佳的情况下难以发挥原本的检测性能。与 KITTI 数据集相比，nuScenes 和 Waymo Open 数据集包含的数据规模更大，并且还在多种天气（如下雨、下雪、雾天等）和照明（如白天和晚上）条件下进行数据的采集，可以更好地评估模型在面临各类天气状况时的表现，为自动驾驶安全提供保障。

2.2　评价指标

判断一个目标检测模型的优劣，通常可以从以下 3 个方面进行评估：目标检测的速度、目标定位的精度和目标分类的精度。

1）目标检测的速度通常用每秒传输帧（Frames Per Second，FPS）来评估，即每秒内可以处理的数据帧数量。FPS 数值越大，说明检测的速度越快，检测的实时性能越好。

2）目标定位的精度一般使用 IoU 来判断，如式（1）所示，其主要是衡量模型生成的预测框与真实框之间的重叠程度，IoU 越接近 1，其定位精度越好，反之越差。

$$IoU=\frac{A\bigcap B}{A\bigcup B} \tag{1}$$

3）目标分类的精度一般使用查准率（Precision）、查全率（Recall）以及平均准确率（Average Precision，AP）等评价指标进行评估。通过设置 IoU 的阈值，可评估检测结果中的真阳性（True Positive，TP）、真阴性（True Negative，TN）、假阳性（False Positive，FP）和假阴性（False Negative，FN），从而计算出模型的查准率 P 与查全率 R，计算式如式（2）、式（3）所示：

$$P=\frac{TP}{TP+FP} \tag{2}$$

$$R=\frac{TP}{TP+FN} \tag{3}$$

以查准率 P 为纵坐标、查全率 R 为横坐标，连接所有点，得到的就是 P-R 曲线。通过对 P-R 曲线进行积分，可以得到平均准确率 AP。在三维目标检测领域，通常会使用两种 AP 指标，分别是 AP_{3D} 和 AP_{BEV}。AP_{3D} 指三维检测框的平均准确率，AP_{BEV} 指鸟瞰图（Bird-Eye View，BEV）下检测框的平均准确率。

为了简化计算，KITTI 使用了基于 40 个查全点插值的方法来计算平均准确率。如式（4）所示，其中 p_{interp} 代表插值点的查准率值。AP 越高，表示对该类的检测精度越高。而对于多类别的整体精度表现，通常使用均值平均精度（Mean Average Precision，mAP）衡量，即对所有类别的平均准确率进行平均。

$$AP=\frac{1}{40}\sum_{r\in\{1/40,2/40,\cdots,1\}}p_{\text{interp}}(r) \tag{4}$$

在三维目标检测任务中，虽然 AP 是用于衡量算法性能的主要指标，但其仅能衡量物体的检测与定位精度，不能对检测结果的方向进行衡量。基于此，KITTI 数据集定义了一个新的指标：平均方向相似性（Average Orientation Similarity，AOS），用于衡量预测框与真实框的航向角相似程度。AOS 的计算方式与 AP 类似，如式（5）、式（6）所示：

$$AOS=\frac{1}{40}\sum_{r\in\{1/40,2/40,\cdots,1\}}\max_{\tilde{r}:\tilde{r}\geqslant r}s(\tilde{r}) \tag{5}$$

$$s(\tilde{r})=\frac{1}{|D(r)|}\sum_{i\in D(r)}\frac{1+\cos\Delta_\theta^{(i)}}{2}\delta_i \tag{6}$$

其中，r 代表查全率，$s(\tilde{r})$ 为方向相似性，其被定义为所有预测框与真实框余弦距离的归一化值；$D(r)$ 表示在查全率 r 下所有预测结果为正样本的集合；$\Delta_\theta^{(i)}$ 表示预测框航向角与真实框之间的差；δ_i 为惩罚项，用于防止多个预测框匹配到同一个真实框，当检测到的目标 i 已经匹配到真实框时 $\delta_i=1$，否则 $\delta_i=0$。

3　基于图像的三维目标检测方法

基于图像的三维目标检测方法仅使用图像数据作为网络的输入，从输入的单张或者多张图像中完成三维目标检测任务。本文根据网络模型对输入图像的不同处理方式，将基于图像的三维目标检测方法分为直接预测的方法、特征升维的方法以及数据升维的方法。

3.1　直接预测的方法

直接预测的方法通常有两类处理方式：1）从图像的二维特征中预测出与目标物体相关的一系列二维信息，然后利用这些信息将预测结果提升到三维空间中；2）结合一些先验信息从图像中预测出三维相关的一些属性，然后得到最终的预测结果。

受到二维目标检测工作的启发，一些方法将二维目标检测网络加以修改并迁移到三维目标检测中。Chen 等[16]对二维目标检测网络 Fast R-CNN[17]进行拓展，提出了 3DOP，将建议框生成的问题定义为马尔可夫随机场的能量最小化问题。针对单张图像输入，Chen 等[18]又在 3DOP 的基础上提出了 Mono3D，结合语义、上下文、目标大小以及位置先验等信息辅助检测，产生最终的三维边界框。类似地，Wang 等将 FCOS[19]迁移到三维目标检测领域，提出了 FCOS3D[20]，整个网络结构如图 1 所示。FCOS3D 将 7 自由度三维目标投影到图像域，将其解耦为二维和三维属性以适应三维场景，根据目标的尺寸将目标分布到特征金字塔的不同级别，并根据训练过程的投影三维中心对目标进行分配，最后利用共享检测头检测出类别和位置信息。

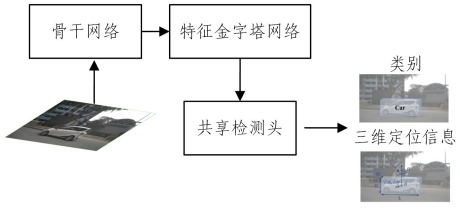

图 1　FCOS3D 网络结构图

Fig. 1　FCOS3D network structure

与迁移二维目标检测方法的思路不同，Liu 等[21]认为二维检测模块是多余的，且多阶段的复杂处理反而会影响网络的学习，因此提出了 SMOKE 网络。整个网络结构如图 2 所示，通过将关键点估计与三维参数回归相结合，从而预测物体的三维边界框。该网络不需要复杂的“前/后处理”或额外的数据和细化阶段，仅通过一个简单的网络结构，就实现了单目三维目标检测任务。

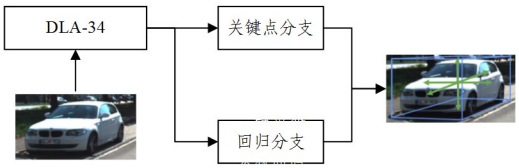

图 2 SMOKE 网络结构图

Fig. 2 SMOKE network structure

在三维目标检测任务中，依靠深度信息能够对空间中的目标进行精准的定位。而图像数据缺乏深度信息，基于图像数据进行三维目标检测一般需要先推断出深度信息，深度信息的准确与否极大地影响了检测的性能，因此许多工作围绕深度估计或者立体匹配展开。Li 等[22]提出了一种基于立体视觉的三维目标检测方法 Stereo R-CNN，其通过扩展 Faster R-CNN[23]用于双目图像输入，能够同时检测和关联左右图像中的目标。Stereo R-CNN 首先预测出粗糙的三维目标边界框，然后使用左右图像中校准后的感兴趣区域进行双目匹配，从而恢复精确的三维边界框。针对像素级别立体匹配导致的高计算消耗问题，Qin 等[24]提出的 TLNet 采用三维锚点来显式构建立体图像对中感兴趣区域(Region of Interest，ROI)之间的物体级别对应关系，从而降低计算量。Sun 等[25]结合学习特定类别的形状先验更准确地预测 ROI 上像素的视差值，然后将其转换为实例伪点云进行三维边界框回归。Xu 等[26]引入自适应缩放模块进一步利用 RGB 图像中丰富的纹理线索进行更精确的视差估计。而考虑到实例深度的耦合性会降低对实例深度预测的精确度，Peng 等[27]提出了 DID-M3D，将实例深度表述为视觉深度和属性深度的组合，将物体三维位置的不确定性解耦为视觉深度不确定性和属性深度不确定性。通过组合不同类型的深度和不确定性，从而获得更精确的实例深度估计值。Huang 等[28]则借助 Transformer 可以有效捕获远程依赖关系的优势，提出了一种端到端的深度感知 Transformer 网络——MonoDTR。该方法可以在不引入额外计算的情况下隐含地学习深度感知特征。针对现有单目三维目标检测网络无法从全局感受野理解场景空间语义信息的问题，MonoDETR[29]引入了 DETR 模型，采用了一种新颖的深度引导机制，首先预测输入图像对应的前景深度图，然后将其作为引导信号，从而引导后续的特征提取和 3D 检测，取得了良好的检测效果。

除了直接从深度估计或立体匹配入手提高三维目标检测性能的思路外，也有一些工作另辟蹊径。Shift R-CNN[30]避开深度估计，将深度学习和几何学的知识相结合，首先预测出目标的二维边界框、三维尺寸和方向信息，然后根据这些预测参数和已知的相机投影矩阵，对三维转换求闭合解。Decoupled-3D[31]将三维目标检测任务分解为一个解耦的结构化多边形预测任务和一个深度恢复任务。Shi 等[32]则对导致距离估计不准确的原因进行了理论和实验分析，从而提出了一个基于几何的距离分解方法。其将目标距离分解为与目标物理高度和在图像平面上投影高度相关的一些量，使得距离的估计具有可解释性。Gu 等[33]提出了在单目三维目标检测中运用全局几何约束的方法，通过该方法建立所有目标物体之间的联系，并全局优化它们的三维位置。Wang 等[34]构建了预测目标物体间的几何关系图，并结合概率表示来捕获深度估计的不确定性，从而提升了深度估计质量。Ma 等[35]利用坐标转化，实现了 Pseudo-LiDAR[36]的等效实现，揭示了伪点云类方法能够发挥作用的原因是图像坐标到世界坐标的转换。Chen 等[37]从伪点云管道的思路中得到启发，为了减小不同模态数据在转化过程中带来的误差，其尝试从单目图像生成伪立体图像进行三维目标检测，极大降低了数据模态转化的难度，取得了良好的检测效果。

针对三维目标检测过程中对目标尺寸、角度以及位置等进行编码时存在的量纲(单位、范围)不一致问题，Simonelli 等[38]将三维目标边界框的属性参数在损失函数层面上分组，解耦它们之间的依赖关系，降低了网络训练的难度。而为了更好地对三维物体的方向角进行预测，Deep3DBox[39]提出了混合离散连续损失，将对三维物体方向角的预测转化为分类加回归的形式，使物体方向角的回归值被限制在一个更小的范围内，从而取得了更好的预测效果。针对三维目标的定位错误问题，Ma 等[40]提出了 3 个改进策略：1)用更精准的定位来解决二维边界框中心和三维边界框投影中心之间的错位问题；2)从训练集中删除利用现有技术难以准确定位的远处物体；3)根据参数对三维 IoU 的贡献率动态地调整样本中参数的损失权重。为了进一步提升检测性能，MonoCon[41]在 MonoDLE 的基础上添加了辅助任务，其关键思想是利用物体的三维边界框标注得到许多映射后的二维信息作为辅助任务的监督信号。通过这种做法，MonoCon 获得了更好的检测效果。

3.2 特征升维的方法

特征升维的方法通常是从二维图像中构造出从点云数据中提取得到的特征，比如鸟瞰图特征或者体素特征，然后基于鸟瞰图特征或者体素特征进行三维目标检测。

基于图像的方法由于缺乏深度信息，目前的检测性能与基于点云的方法相比还比较落后。为了提升检测性能，许多研究开始了不同的尝试。Srivastava 等[42]首次尝试使用生成对抗网络从二维图像中生成鸟瞰图特征，并与现有的以点云作为输入的检测网络 BirdNet[43]和 MV3D[44]相结合来实现三维目标的检测和定位，取得了较为优异的性能。考虑到直接估计深度值难度较大且不准确，CaDDN[45]把深度值离散化为一些区间，预测逐个像素的分类深度分布，将丰富的背景特征信息映射到三维空间的对应深度区间，并结合单阶段检测网络来得到最终的检测结果，有效降低了深度预测的难度。

现有基于图像的三维目标检测算法通常采用基于透视图的表达方式，然而这种表达方式存在一个问题，即目标的形状和尺度会随着深度的变化而急剧变化，因此很难推断出有意义的距离信息。为了解决这个问题，Roddick 等[46]引入了正交特征变换，将图像的特征映射到正交的三维空间来摆脱图像域的限制，使网络在一个比例一致、物体间距离有意义的域中，对目标的参数进行整体推理。Huang 等[47]则提出了一种多相机输入的三维目标检测范式，通过重用现有二维目标检测模块构建了 BEVDet，BEVDet 通过视图转换器将多个图像的特征映射到鸟瞰图中，最后在鸟瞰图中结合现有的算法进行三维目标检测。考虑到单帧数据的信息十分有限并且

在自动驾驶场景中往往可以非常容易地获得历史帧的数据和特征用于辅助当前帧的预测，Huang 等[48]通过升级 BEVDet 范式，提出了 BEVDet4D。BEVDet4D 整个网络结构如图 3 所示，其将提取的特征通过 BEV 编码器转化为鸟瞰图特征，根据自车的运动信息将前一帧特征在世界坐标系中进行对齐(Align)，然后通过分离操作(Detach)产生新的特征向量，将其与当前帧的特征进行通道的拼接(Concatenate)，从而达到融合前一帧特征与当前帧特征的效果，在时空四维空间中进行三维目标的检测。在历史帧信息的辅助下，BEVDet4D 在目标的精度、朝向以及速度各个方面的预测上均有了大幅的提升。

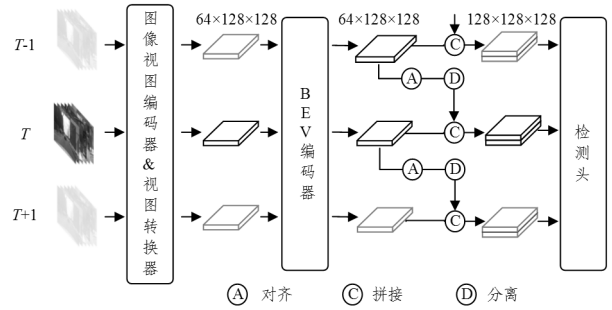

图 3　BEVDet4D 网络结构图

Fig. 3　BEVDet4D network structure

BEVDet 范式对深度值很敏感，但大部分方法对深度值的感知是在没有相机信息的情况下隐式学习的。在没有显式深度信息监督的情况下，准确感知深度值是非常困难的。因此 Li 等提出了 BEVDepth[49]，其利用编码的相机内外参数获得显式的深度监督，并通过融合前一帧的数据来丰富当前帧中的信息，从而辅助对速度和深度的预测。而 BEVFormer[50]利用 Transformer 能捕获远程依赖关系的能力，聚合来自多个视图的时空特征和历史 BEV 特征，实现了不依赖深度信息生成鸟瞰图特征，从而完成三维目标检测。BEVStereo[51]则引入多视角立体视觉技术，试图打破深度估计造成的检测瓶颈，并针对传统多视角立体视觉技术应用于三维目标检测时存在的视图间亲和度测量计算成本高、难以处理室外移动场景的问题做了改进，提出了一种有效的方法来动态选择匹配候选区域的尺度，使网络能够更好地检测移动的目标物体。为了克服 BEV 检测器受图像主干网束缚的问题，BEVFormer v2[52]引入了透视监督，将从透视视角产生的监督信号直接作用于主干网，以帮助其学习二维识别任务中丢失的三维信息。同时，BEVFormer v2 还对 BEV 检测器的复杂结构进行了改进，极大优化了其性能。

为了提升送入预测网络时的特征质量，DSGN[53]将双目图像对作为输入，利用权重共享的孪生网络提取二维特征，并构建一个平面扫描体(Plane-Sweep Volume，PSV)用于学习像素的对应关系，最后将 PSV 转换为三维几何体以供检测。通过这种表示可以更好地学习深度信息和语义线索，从而提升所提取特征的质量。DSGN＋＋[54]在此基础上提出了一种新颖的立体体积表示方法，集成前视图和俯视图特征并利用数据增强策略缓解 DSGN 中存在的类别不均衡问题。而为了更好地建立二维图像和三维空间合理的对应关系，MonoNeRD[55]使用有符号距离函数对场景进行建模，从而生成稠密的三维表示，并将其视为神经辐射场(Neural Radiance Fields，NeRF)，利用体渲染从中恢复出 RGB 图像和像素级别的深度图，极大地提高了单目三维目标检测的精度。

为了缓解深度信息缺失造成的检测瓶颈问题，SGM3D[56]利用立体图像提取的三维特征增强单目图像学习的特征，提出了多粒度域适应模块以及基于 IoU 匹配的对齐模块来引导单目网络在不同特征级别上模仿立体特征，从而有效提升了单目网络的检测效果。而针对立体方法存在的对三维空间几何感知特征表示不足的问题，ESGN[57]利用高效的几何感知特征生成(Efficient Geometry-Aware Feature Generation，EGFG)模块，在相机截锥体空间中构建多尺度立体体积，并使用多尺度鸟瞰图(BEV)投影和融合模块生成多个几何感知特征。在不需要复杂聚合网络的条件下，有效提升了对三维空间的几何感知能力。

3.3　数据升维的方法

数据升维的方法灵感来自于点云的三维目标检测，这类方法将图像转化为伪点云(或其变体)的表示方式，然后利用点云三维目标检测管道完成检测任务。

许多工作认为深度估计不准确是造成检测性能较差的主要原因，而 Wang 等[36]认为更深层次的原因来源于数据表示方式，对此其提出了 Pseudo-LiDAR 算法，整个网络结构如图 4 所示。Pseudo-LiDAR 通过深度估计将预测的深度图转化为伪点云，然后使用基于点云的检测管道对伪点云进行检测，得到了良好的检测效果，为基于图像的三维目标检测开拓了一个新的研究分支。You 等[58]通过分析发现，基于立体视觉的方法对三维目标定位误差的大小主要取决于深度估计误差的大小，因此其改进 Pseudo-LiDAR，提出了 Pseudo-LiDAR＋＋，通过调整立体网络架构和损失函数进行直接深度估计来缓解这个问题。虽然 Pseudo-LiDAR，Pseudo-LiDAR＋＋取得了不错的效果，但是整个网络的训练并不是端到端的，这对网络整体的优化是次优的。因此，Qian 等[59]提出了一个新的框架，它基于可微的表示变化模块，允许对整个伪点云管道进行端到端的训练，且能与大多数最先进的网络兼容。类似地，Kim 等[60]提出了一种使用单目图像序列作为输入，在自监督损失下进行学习的端到端网络框架，有效提升了检测性能。为了应对基于立体视觉的伪点云类方法在复杂立体匹配算法下所面临的计算量大的挑战，Meng 等[61]提出了一种轻量级的检测算法，使用带有二元神经网络的高效深度估计器进行实时的深度预测，在保持实时性的同时实现了检测精度的提升。而为了利用知识蒸馏技术提高单目检测的效果，Sun 等[62]利用真实点云训练教师网络，并使用学生网络接收来自教师网络的提炼知识，然后将单目图像转换为伪点云并将其作为学生网络的输入进行检测，取得了良好的检测效果。

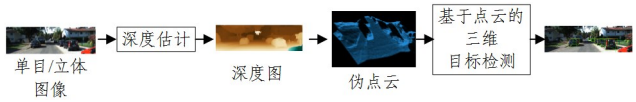

图 4　Pseudo-LiDAR 网络结构图

Fig. 4　Pseudo-LiDAR network structure

针对从单一图像中生成的原始伪点云较为密集且包含大量的背景点，容易导致检测出现歧义的问题，RefinedMPL[63]对原始伪点云进行预处理，将原始伪点云转化为更有利于三维目标检测任务的稀疏表示，在没有太多计算开销的情况下，取得了较好的检测性能。针对伪点云中存在大量噪声的问题，Weng 等[64]从两个方面出发：1)使用“二维—三维”边界框一致性约束调整估计的三维边界框；2)使用实例掩码代替二维建议框的表示，从而减少点云视锥体中不属于目标的点数。这两点改进有效缓解了伪点云中存在的大量噪声问题。为了使伪点云中包含更丰富的特征信息，AM3D[65]根据二维目标建议框对目标区域进行前景分割，得到目标实例的像素，然后利用注意力机制将 RGB 线索嵌入到伪点云的表示中，最后使用 PointNet 作为主干网络回归目标的三维边界框，取得了较好的检测效果。

许多将图像转化为伪点云表示的方法在进行深度估计时会同等对待前景和背景像素，但由于前景像素和背景像素深度分布模式不同，这种处理会导致深度估计的效果较差。因此，CG-Stereo[66]在立体匹配网络中对前景和背景像素使用两个独立的解码器，并利用立体匹配算法估计的置信度作为一种软注意力机制引导目标检测器更多地关注具有高质量深度信息的点，从而提高三维目标检测的精度。

虽然现有的基于伪点云的三维目标检测方法已经取得了明显成效，但仍存在以下问题：1)由于图像的高分辨率以及特征需要额外的维度表示，因此学习图像空间中的三维匹配代价(3D Cost Volume)在内存和计算方面的开销非常大；2)深度估计和下游检测任务不在同个度量空间中执行，由于透视投影，近处的物体比远处的物体占据更多的像素，这种不平衡会影响深度估计的质量。因此，PLUMENet[67]直接在三维度量空间中构造一个伪点云特征体(Pseudo LiDAR Feature Volume)，从中执行深度估计和三维目标检测，以较快的速度取得了良好的检测性能。为了提高检测速度，Meng 等[68]通过低质量的深度估计网络以及点云数据增强方式进行伪点云的构建，以极低的计算延迟取得了与一些基于深度神经网络进行准确深度估计的方法同等甚至更好的性能。

3.4 算法性能的对比分析

基于图像的三维目标检测方法通常以单张或者多张图像作为输入，从中对三维目标进行检测和定位。由于图像缺乏深度信息，目前基于图像的三维目标检测方法的检测效果还比较差。表 1 列出了基于图像的三维目标检测各个类别方法在 KITTI 测试集上对 Car 类别的性能指标。所有数据来源于 KITTI 官方排行版。各个类别中的最佳性能用粗体标出。

直接预测的方法或是从图像中预测出二维信息然后提升到三维空间，或是预测出三维目标相关的一些属性然后得到最终预测结果。单目三维目标检测因其低廉的价格和简单的部署方案受到了极大的欢迎，相关工作如 MonoDLE，MonoCon，DID-M3D 对单目三维目标检测进行了快速的改进。通过分析单目三维目标检测性能低下的原因，MonoDLE 中给出了相关的改进措施。而 MonoCon 则在此基础上，在网络的训练阶段进一步添加了辅助任务，在保证推理速度的同时利用单目上下文信息进一步提高检测性能。为了更好地估计目标实例的深度，DID-M3D 将实例深度解耦得到更好的估计结果从而提升单目检测的性能。与单目深度估计相比，多目(包括双目)深度估计能得到更精确的深度信息，从而使得多目三维目标检测方法的性能比单目的方法高很多。多目的三维目标检测方法如 TLNet 针对在像素级别上进行立体匹配的方法存在的耗时多、计算代价大的缺点，提出采用三维锚点来显式构建立体图像对中 ROI 之间的目标级别对应关系，从而避免计算像素级别的深度图，提高网络的推理速度。由表 1 中可以看出，以单张图像作为输入的方法的检测性能均比以多张图像作为输入的方法性能低，但检测速度大多比以多张图像作为输入的方法快，这主要是多图匹配比较耗时造成的。如何提高深度估计的速度和精度以及二维信息和三维信息的准确映射是直接预测的方法的研究重点。

特征升维的方法的基本思想是将图像坐标系中的二维图像特征转化为世界坐标系中的三维体素特征或进一步压缩垂直维度，生成鸟瞰图特征，然后送入网络预测最终结果。在这类方法中，关键问题是如何将二维图像特征转化为更能表征目标的三维体素特征或者鸟瞰图特征。由于单目深度估计的不准确性，直接预测深度值的方法效果较差。CaDDN 将深度值的预测转化为分类加回归的方式，然后结合鸟瞰图和单阶段检测器来产生最终的检测结果。以多张图像作为输入的作品如 DSGN，其利用权重共享的孪生网络提取二维特征后转化为三维几何体，并进一步利用三维卷积网络获取鸟瞰图特征进行三维目标检测，取得了较优的性能。针对 DSGN 存在的对三维特征表示能力不足和类别不均衡问题，DSGN++采用了深度平面扫描和数据增强策略，大幅提升了 DSGN 的检测性能。虽然利用单张图像的 CaDDN 改进了深度值的预测方式，但由表 1 可以看出，其与使用多张图像作为输入的 CaDDN 以及 CaDDN++相比性能差距仍较大，这主要是单目深度估计的瓶颈所致。如何提高深度估计的准确性以及构造更好的体素或鸟瞰图特征是特征升维的方法的研究重点。

数据升维的方法通过将图像转化为伪点云的表示形式进行三维目标检测，它可以和现有的基于点云的三维目标检测管道结合实现三维目标的检测和定位。一些方法如 Mono3D_PLiDAR，RefinedMPL 从改善生成的伪点云的质量出发进行研究。Mono3D_PLiDAR 以单张图像作为输入，通过“二维—三维”边界框一致性约束和实例掩码策略缓解了生成的伪点云中存在的大量噪声问题；RefinedMPL 提出了稀疏化方案，将伪点云转化为稀疏表示得到了更好的检测效果。另一些方法如 Pseudo-LiDAR++从提高深度估计的精度进行研究，通过调整网络架构和损失函数以及利用稀疏的激光雷达传感器的数据做监督以提高性能。由表 1 可以看出，基于单张图像的方法和基于多张图像的方法的性能差距仍很大，这主要是深度估计的精确度不足造成的。不断改善生成的伪点云的质量、提高深度估计的精确度是数据升维的方法的研究重点。

表 1　基于图像的三维目标检测方法在 KITTI 测试集 Car 类别上的性能指标

Table 1　Performance indicators of image based 3D object detection methods in the Car category of KITTI test set

类别	方法	图像数量	$AP_{3D}/AP_{BEV}(IoU=0.7)/\%$			运行时间/s
			简单	中等	困难	
直接预测	MonoDIS[38]	单张	10.37/17.23	7.94/13.19	6.40/11.12	N
	SHIFT R-CNN[30]	单张	6.88/11.84	3.87/6.82	2.83/5.27	0.25
	TLNet[24]	多张	7.64/13.71	4.37/7.69	3.74/6.73	0.10
	Stereo R-CNN[22]	多张	47.58/61.92	30.23/41.31	23.72/33.42	0.30
	Decoupled-3D[31]	单张	11.08/23.16	7.02/14.82	5.63/11.25	0.08
	SMOKE[21]	单张	14.03/20.83	9.76/14.49	7.84/12.75	**0.03**
	PatchNet[35]	单张	15.68/22.97	11.12/16.86	10.17/14.97	0.40
	Disp R-CNN[25]	多张	37.12/**79.76**	25.80/**58.62**	22.04/**47.73**	0.39
	ZoomNet[26]	多张	**55.98**/72.94	**38.64**/54.91	**30.97**/44.14	0.30
	MonoDETR[29]	单张	24.52/32.20	16.26/21.45	13.93/18.68	0.04
	MonoRCNN[32]	单张	18.36/25.48	12.65/18.11	10.03/14.10	0.07
	MonoDLE[40]	单张	17.23/24.79	12.26/18.89	10.29/16.00	0.04
	HomoLoss[33]	单张	11.87/29.60	7.66/20.68	6.82/17.81	0.04
	DID-M3D[27]	单张	24.40/32.95	16.29/22.76	13.75/19.83	0.04
	MonoDTR[28]	单张	21.99/28.59	15.39/20.38	12.73/17.14	0.04
特征升维	DSGN[53]	多张	73.50/82.90	52.18/65.05	45.14/56.60	0.67
	CaDDN[45]	单张	19.17/27.94	13.41/18.91	11.46/17.19	0.63
	DSGN++[54]	多张	**83.21**/**88.55**	**67.37**/**78.94**	**59.91**/**69.74**	**0.20**
	MonoNeRD[55]	单张	22.75/31.13	17.13/23.46	15.63/20.97	N
数据升维	Pseudo-LiDAR[36]	多张	54.53/67.30	34.05/45.00	28.25/38.40	0.40
	RefinedMPL[63]	单张	18.09/28.08	11.14/17.60	8.94/13.95	0.15
	Mono3D_PLiDAR[64]	单张	10.76/21.27	7.50/13.92	6.10/11.25	**0.10**
	Pseudo-LiDAR++[58]	多张	68.38/**84.61**	**54.88**/**73.80**	**49.16**/**65.59**	0.60
	CG-Stereo[66]	多张	33.22/39.24	24.31/29.56	20.95/25.87	0.57
	PLUMENet[59]	多张	N /82.97	N /66.27	N /56.70	0.15

注:"N"表示无法从 KITTI 官方排行版中查到相关数据;Car 类别 AP_{3D}和 AP_{BEV}的 IoU 阈值设置为 0.7。

从整个基于图像的三维目标检测方法来看,使用多张图像进行三维目标检测的效果明显优于使用单张图像的,这主要是由于多目深度估计能够提供比单目深度估计更精确的深度信息,但相应的对硬件层面的要求也更高。相比直接预测的方法,特征升维和数据升维的方法检测效果更佳,但其检测速度相对较慢,原因主要是将图像数据转化成其他特征形式或转化成伪点云的表示需要一定的计算消耗。表 2 总结了基于图像的几类三维目标检测方法的优劣。

目前基于图像的三维目标检测方法的检测性能相对较差,对于远距离或者受遮挡的小目标的检测效果并不理想。许多研究集中在 KITTI 数据集上进行,在大型、复杂数据集上的研究还相对较少。这些问题都将激发研究界对基于图像的三维目标检测的探索热情。

表 2　基于图像的三维目标检测方法的对比

Table 2　Comparison of image based 3D object detection methods

类型	优点	缺点
直接预测	无需复杂的数据模态转化过程	受透视投影影响,深度估计偏差较大
特征升维	具有更好的三维场景理解力	数据模态转换过程需要额外的计算
数据升维	具有更丰富的几何信息,可以更好地处理目标间的遮挡和重叠情况	检测效果受生成伪点云质量影响,对噪声和误差敏感

4　图像与点云融合的三维目标检测方法

虽然不少工作聚焦于仅使用图像数据进行三维目标检测,但由于图像数据缺乏深度信息,仅基于图像数据进行三维目标检测存在很大的挑战。为了进一步提升检测性能,利用不同传感器的优势弥补单一模态数据存在的不足,许多工作对图像与点云融合的三维目标检测进行了研究。根据网络融合不同模态数据的时机,本文将图像与点云融合的三维目标检测方法分为特征级融合的方法以及决策级融合的方法。

4.1　特征级融合的方法

特征级融合的方法通常是对图像数据和点云数据本身或提取的特征进行融合,然后再进行三维目标的检测。

由于二维卷积神经网络在图像上拥有强大的特征提取能力,一些方法将点云数据转化成不同的视图,然后结合 RGB 图像完成三维目标检测。MV3D[44]以点云的鸟瞰图、前景图及 RGB 图像作为输入进行三维目标检测。首先从鸟瞰图生成一系列精准的三维建议框;然后将其投影至前景图和 RGB 图像的特征图中,并采用深度融合网络对每个视图的 ROI 池化特征进行融合;最后利用融合特征预测目标类别及其三维边界框。Ku 等[69]则仅从点云鸟瞰图和 RGB 图像入手,结合给定的三维锚框提取鸟瞰图和 RGB 图像对应区域的特征,然后利用双线性插值剪裁两类特征图并进行像素级别的融合以获取候选区域,对较小目标类别取得了更好的检测效果。为了更好地融合多视图特征,SCANet[70]引入空间通道注意力模块捕获多尺度特征和全局上下文信息,并结合多级融合方案,将不同视图的特征融合在一起。出于同样的动机,Guo 等[71]提出了一种深度多尺度多模态融合的三维目标检测方法,以点云鸟瞰图和 RGB 图像作为输入,提取两者的多尺度

特征进行融合，最后输入到位置回归和分类网络进行三维目标的分类和定位。而针对 RGB 图像和激光雷达范围图存在噪声信息(如遮挡和截断的情况)的问题，Wang 等[72]提出了 MVAF-Net，以点云鸟瞰图和激光雷达范围图以及 RGB 图像作为输入，通过一个注意力逐点融合(Attentive Pointwise Fusion)模块实现多视图特征的自适应融合。此外，他们还设计了一个注意力逐点加权(Attentive Pointwise Weighting)模块，帮助网络学习结构信息和点特征的重要性，在 KITTI 数据集上实现了速度和精度之间的平衡。

将点云数据转化成不同视图的方法虽可利用成熟的二维卷积神经网络提取特征，但在将点云转化成不同视图的过程中会损失许多空间信息。为了更好地保留点云中的空间信息，Qi 等[73]提出了一种基于视锥体的检测框架 F-PointNets，该框架以 RGB-D 数据作为输入并结合二维检测器对目标进行定位，从而生成三维视锥体候选区域，并利用 PointNet[74](或 PointNet++[75])对视锥体内的点云进行实例分割及边界框回归。类似地，SIFRNet[76]利用前景图和对应的视锥体内的点云来生成三维目标检测的结果。SIFRNet 由 Point-UNet，T-Net 和 Point-SENet 这 3 个子网络组成。Point-UNet用于实现目标尺度的不变性以及捕获目标的朝向信息，并将从图像中提取的特征与视锥体内的点云进行融合；T-Net 用于中心化感兴趣的点，并通过 Point-SENet 预测三维边界框。SIFRNet 可以在点云极其稀疏的情况下取得较好的检测结果，具有一定的泛化能力和鲁棒性。

相比相机设备采集的密集特征，激光雷达采集的点云数据十分稀疏。为了提升检测效果，更好地融合两种异构数据至关重要。针对早期的多模态三维目标检测方法缺少前期模态间交互学习的问题，MVX-Net[77]提出在前期阶段融合图像和点云特征。为了减少融合过程中的信息丢失问题，3D-CVF[78]提出门控特征融合网络以及空间注意力映射机制，从而更好地融合相机视图特征和点云特征。EPNet[79]提出新颖的融合单元，采用逐点的方式使用语义图像特征来扩充点的特征，且不需要任何图像标注(也就是二维检测框)。DVF[80]提出密集体素融合的顺序融合方法，通过生成多尺度密集体素特征表示从而提高低点云密度区域的表现力，进而提高检测效果。为了充分交互来自两种不同模态数据的特征，MSMDFusion[81]首先使用多深度非投影方法增强图像像素的深度质量，然后应用特定模态的稀疏卷积模块聚合局部信息，并将几何和语义特征逐步融合到统一的空间中，共同为检测头提供更全面的信息。

Liang 等[82]提出了一种基于 BEV 的三维目标检测方法，利用连续卷积将不同分辨率的图像和点云特征图融合在一起。而针对基于 BEV 的融合方法存在的对特征融合不准确的问题，Xie 等提出了 PI-RCNN[83]，直接在点云上融合多传感器特征。与直接融合点云和图像数据的做法不同，PointFusion[84]独立处理点云和图像数据，分别用 ResNet[85]和 PointNet 提取图像和点云数据的特征，然后由融合网络组合所得到的输出，并利用输入的点云作为空间锚点来预测目标的三维边界框及其置信度。然而 PointFusion 并不具备对多模态输入进行联合推理的能力，检测结果会受二维检测网络性能的影响。因此，Wang 等[86]提出了多阶段互补融合的三维目标检测网络 MCF3D 来缓解这个问题。MCF3D 将点云数据和 RGB 图像数据作为输入，自适应地加权来自不同视图的候选特征，使融合结果包含有效且关键的信息，更利于高度遮挡或拥挤场景的三维目标的检测。

除了更好地融合多模态数据的思路，一些工作也从学习更好的特征表示出发。Liang 等[87]通过联合解决多个感知任务来学习更好的特征表示，以提升检测性能。PointAugmenting[88]使用预训练好的二维检测模块提取图像上的像素级特征来增强点云数据所表达的信息，在经过编码的点云数据上进行三维目标检测。Zhang 等[89]借助 Transformer 容易学习到全局特征的能力，提出了 CAT-Det。CAT-Det 网络结构如图 5 所示，图中 PTB 和 ITB 分别表示点云 Transformer 编码分支和图像 Transformer 编码分支，CMT 表示跨模态 Transformer 模块，OMDA 表示单向多模态数据增强方法。CAT-Det 采用双分支编码点云与图像模式的长距离上下文特征，并通过跨模态模块对两种模态的特征进行融合，实现了特征的有效学习。Wu 等[90]则利用深度补全生成伪点云并通过特征提取器提取伪点云的二维图像特征和三维几何特征，利用有效的 ROI 融合策略得到更加细粒度和更加精确的特征用于检测。

尽管多模态融合方案在三维目标检测领域越来越流行，但在劣质图像条件下(如照明不佳或传感器未对准)鲁棒性并不好，这主要是由于标定矩阵建立的是点云和图像像素的硬关联，因此很依赖图像的质量。针对这个问题，Bai 等提出了基于 Transformer 结构的 TransFusion[91]，使用软关联机制来处理劣质图像的情况，自适应地确定应该从图像中获取的信息，从而取得了更好的融合效果。大多数融合方案在缺失某一模态数据时无法正常工作。为了应对传感器失效的突发情况，Liang 等[92]提出了 Bevfusion，将多视图相机和雷达传感器输入编码到同一 BEV 空间中，利用所提出的动态融合模块融合两个模态的特征，最后进行下游检测任务，解决了许多融合方法中相机和雷达传感器之间的依赖问题。

为了增强点云数据所表达的语义信息，Vora 等[93]提出了 PointPainting，利用一种简单的顺序多模态融合结构将点云投影到语义分割后的二维图像上，并利用已有的三维检测器实现对目标的检测。类似地，FusionPainting[94]也在语义层面融合点云和图像数据。其通过提出的基于注意力的融合模块，将分割得到的图像和点云的语义信息进行自适应融合，利用带有语义标签的点云进行检测任务。为了更好地对齐图像数据和点云数据，Chen 等[95]提出了 AutoAlign，应用一种自动特征融合策略，通过一个可学习的对齐图(Learnable Alignment Map)来建模“图像-点云”之间的映射关系。为了保持跨模态信息的一致性，Li 等[96]提出了一种名为体素场融合的跨模态三维目标检测方法，使用射线的方式在体素场中表示并融合特征，在 KITTI 和 nuScenes 数据集上均取得了不错的检测效果。

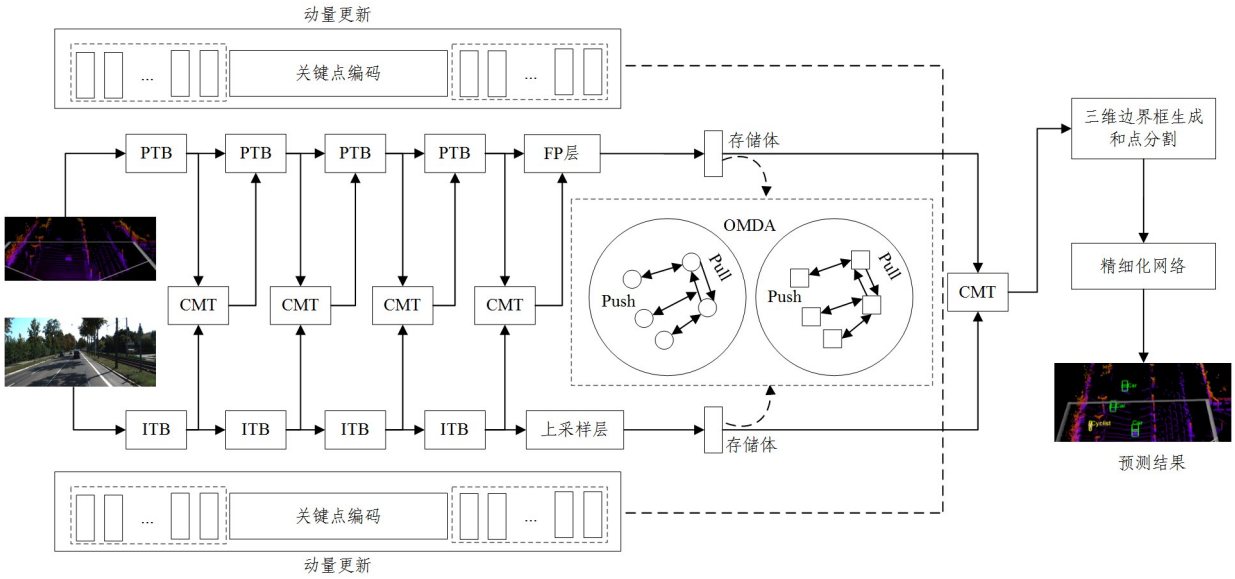

图 5　CAT-Det 网络结构图

Fig. 5　CAT-Det network structure

4.2　决策级融合的方法

决策级融合的方法通常是对图像数据和点云数据分别进行三维目标检测，然后将两个检测结果融合从而得到更好的最终结果。

Kim 等[97]将稀疏的点云输入投影成密集的前视图，并借助 Fast R-CNN 分别在图像及点云的投影视图中生成区域建议框，随后将两个图像上的建议框进行融合形成最终的结果。Asvadi 等[98]将点云转换为相关深度图和反射率密集图，并使用 YOLO 生成点云投影视图及图像的检测结果，然后进行融合操作。但由于当时点云检测技术尚未成熟，因此这些方法都将点云数据降维到图像中进行处理，导致最终检测性能不太理想。

目前在图像及点云等单一模态数据方面已有较好的目标检测方法，然而将各模态在决策级融合以实现更好的检测效果还需要进一步研究。Pang 等提出了一种低复杂度的融合网络 CLOCs[99]，该网络利用现有的图像及点云目标检测方法，基于二维与三维建议框提取几何与语义一致性特征，并学习其概率相关性以进行融合。该方法以低复杂度的融合方式，有效地提升了在多模态数据下的检测性能，然而 CLOCs 需要同时运行二维及三维检测器，这需要较高的计算消耗。为了进一步减少内存占用，Pang 等[100]又进一步提出了 Fast-CLOCs 网络。与 CLOCs 的不同的是，该网络去除了二维检测器，仅使用轻量级的网络提取图像特征，将三维建议框投影到二维中并使用图像特征进行修正以完成二维建议框的生成，最后将二维及三维建议框使用 CLOCs 网络进行融合。该方法由于去除了完整的二维检测器，减少了内存占用及计算消耗并实现了实时检测。

整个网络结构如图 6 所示。

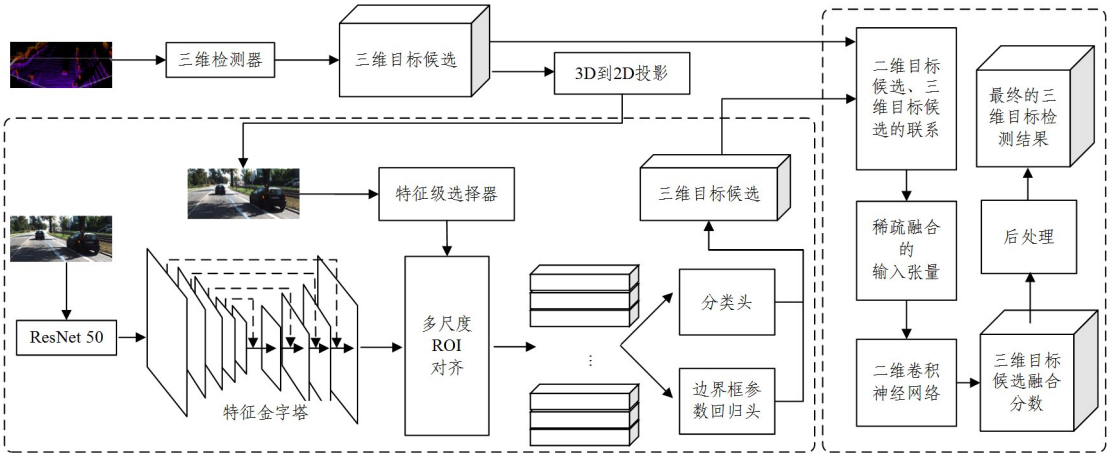

图 6　Fast-CLOCs 网络结构图

Fig. 6　Fast-CLOCs network structure

为了进一步对图像的语义特征和点云的几何特征进行有效融合，Guo 等[101]提出了 LIGA-Stereo 网络模型。该方法首先对输入的双目图像进行二维语义特征提取并生成二维建议框，随后将建议框投影到三维空间中，并在点云三维检测器的指导下进一步学习语义特征，以实现语义特征与几何特征的融合。该方法在大尺度及小尺度目标上的检测效果均有提升。

表 3 图像与激光点云融合的三维目标检测方法在 KITTI 测试集上的性能指标

Table 3 Performance indicators of 3D object detection methods based on image and laser point cloud fusion on KITTI test set

检测类别	方法	图像数量	$AP_{3D}/AP_{BEV}(IoU=0.7)/\%$			运行时间/s
			简单	中等	困难	
Car	特征级融合	MV3D[44]	74.97/86.62	63.63/78.93	54.00/69.80	0.36
	特征级融合	AVOD[69]	83.07/90.99	71.76/84.82	65.73/79.62	0.10
	特征级融合	F-PointNets[73]	82.19/91.17	69.79/84.67	60.59/74.77	0.17
	特征级融合	UberATG-ContFuse[82]	83.68/94.07	68.78/85.35	61.67/75.88	**0.06**
	特征级融合	UberATG-MMF[87]	88.40/93.67	77.43/88.21	70.22/81.99	0.08
	特征级融合	PointPainting[92]	82.11/92.45	71.70/88.11	67.08/83.36	0.40
	特征级融合	EPNet[79]	89.81/94.22	79.28/88.47	74.59/83.69	0.10
	特征级融合	DVF[80]	89.40/93.12	82.45/89.42	77.56/86.50	0.10
	特征级融合	MVAF-Net[72]	87.87/91.95	78.71/87.73	75.48/85.00	**0.06**
	特征级融合	PI-RCNN[83]	84.37/91.44	74.82/85.81	70.03/81.00	0.10
	特征级融合	3D-CVF[78]	89.20/93.52	80.05/89.56	73.11/82.45	0.06
	特征级融合	CAT-Det[89]	89.87/92.59	81.32/90.07	76.68/85.82	0.30
	特征级融合	SFD[90]	**91.73/95.64**	**84.67/91.85**	**77.92/86.83**	0.10
	决策级融合	CLOCs[99]	**89.16/93.05**	**82.28/89.80**	**77.23/86.57**	**0.10**
	决策级融合	LIGA-Stereo[101]	81.39/88.15	64.66/76.78	57.22/67.40	0.40
	决策级融合	Fast-CLOCs[100]	89.10/93.03	80.35/89.49	76.99/86.40	**0.10**
Pedestrian	特征级融合	AVOD[69]	50.46/58.49	42.27/**50.32**	39.04/**46.98**	**0.10**
	特征级融合	F-PointNets[73]	50.53/57.13	42.15/49.57	38.08/45.48	0.17
	特征级融合	PointPainting[92]	50.32/**58.70**	40.97/49.93	37.87/46.29	0.40
	特征级融合	CAT-Det[89]	**54.26**/57.13	**45.44**/48.78	**41.94**/45.56	0.30
	决策级融合	LIGA-Stereo[101]	40.46/44.71	30.00/34.13	27.07/30.42	0.40
	决策级融合	Fast-CLOCs[100]	**52.10/57.19**	**42.72/48.27**	**39.08/44.55**	**0.10**
Cyclist	特征级融合	AVOD[69]	63.76/69.39	50.55/57.12	44.93/51.09	**0.10**
	特征级融合	F-PointNets[73]	72.27/77.26	56.12/61.37	49.01/53.78	0.17
	特征级融合	PointPainting[92]	77.63/83.91	63.78/71.54	55.89/62.97	0.40
	特征级融合	CAT-Det[89]	**83.68/85.35**	**68.81/72.51**	**61.45/65.55**	0.30
	决策级融合	LIGA-Stereo[101]	54.44/58.95	36.86/40.60	32.06/35.27	0.40
	决策级融合	Fast-CLOCs[100]	**82.83/83.34**	**65.31/67.55**	**57.43/59.61**	**0.10**

注:"N"表示无法从 KITTI 官方排行版中查到相关数据;Car 类别 AP_{3D} 和 AP_{BEV} 的 IoU 阈值设置为 0.7;Pedestrian 和 Cyclist 类别 AP_{3D} 与 AP_{BEV} 的 IoU 阈值设置为 0.5。

4.3 算法性能的对比分析

为了结合图像数据和点云数据的优势,进一步提升三维目标检测的性能,许多工作聚焦在图像与点云融合的三维目标检测方法上。该类方法在近年来取得了很大的进展,但其对硬件的要求较高,在实际部署时会受到价格、环境等方面的影响。表 3 列出了图像与激光点云融合的方法在 KITTI 测试集上对各个类别的性能指标。所有数据来源于 KITTI 官方排行版。各个类别中的最佳性能用粗体标出。

特征级融合的方法通常会对图像数据和点云数据及其提取的特征进行融合,通过单次或者多次融合来结合不同模态中对目标物体的有效表征,从而提高三维目标检测的性能。一些方法如 MV3D,AVOD,SCANet,MVAF-Net 等将点云数据转化成不同的视图,然后结合 RGB 图像完成三维目标检测,但这种方式会损失许多的空间信息。为了更好地保留点云中的空间信息,一些作品如 F-PointNets 和 SIFRNet 提出了基于视锥体的检测框架,该框架直接从点云中提取信息,可以有效克服强遮挡问题,但其会依赖二维检测模块。CAT-Det 则采用双分支编码点云与视图特征来实现多模态信息的有效提取和融合,取得了对小目标物体较好的检测效果。决策级融合的方法对数据的处理流程与特征级融合的方法相比有较大的不同,它不需要复杂的数据融合和数据对齐操作。这类方法通常是分别对图像数据和点云数据进行三维目标检测,然后将两者的检测结果相结合得到精度更高的最终结果。由表 3 可以看出,目前图像与点云融合的方法对大目标类别"Car"的检测精度已经很高,在检测速度上不少方法如 MVAF-Net 以及 3D-CVF 也已经很快,但对于小目标类别"Pedestrian""Cyclist"来说,检测精度还有较大的提升空间。对于特征级融合的方法来说,如何提高对不同模态数据的融合效果是该类方法的研究重点;对于决策级融合的方法来说,目前在这一块的研究相对较少,还有许多待探索的研究思路。如何有效利用不同模态的检测结果来提升最终的检测精度是这类方法的研究重点。表 4 总结了图像和点云融合的三维目标检测方法的优劣。

表 4 图像和激光点云融合的三维目标检测方法的对比

Table 4 Comparison of 3D object detection methods based on image and laser point cloud fusion

类型	优点	缺点
特征级融合	综合点云数据中精确的空间信息以及图像数据中丰富的语义和纹理信息构建更好的特征表示	对齐和融合两种模态数据过程中对计算资源的消耗较大
决策级融合	不需要对特征做对齐和融合,计算消耗相对较少	无法综合两种模态数据的优势做互补,只能在检测结果上做融合

从整个图像与点云融合的三维目标检测方法来看,该类方法有着极大的潜力。融合类的方法可以利用点云数据中精确的空间信息以及图像数据中丰富的语义和纹理信息,这对于小目标物体的识别以及大型复杂场景的目标识别来说,会比提供单一模态数据取得的检测性能的上限高很多。相应地,

该类方法也面临挑战，如何合理有效融合两种模态数据、排除噪声信息干扰，以及如何对两类数据的检测结果有效融合至关重要。图像与点云融合的三维目标检测是一个十分有前景的研究领域，该领域存在的问题和巨大的发展潜力将会吸引越来越多研究者的加入。

5　总结与展望

尽管图像三维目标检测技术取得了一定的进展，但仍难以满足自动驾驶场景下的现实需要。图像三维目标检测领域还面临着许多挑战，同时也存在许多可探索的方向。结合对近年来相关文献的分析和总结，本文对图像三维目标检测领域存在的问题和未来研究方向进行总结和展望。

1）提升深度估计质量。由于透视投影的影响，物体在图像中的大小和形状会发生较大变化，直接预测的方法从图像中提取的二维特征很难准确反映物体在实际三维空间中的情况，因此难以准确捕捉深度信息。特征升维的方法可以将图像转换为体素或鸟瞰图表征，从而提供更多的物体几何特征和空间关系信息。但这一转化过程通常依赖深度估计，而数据升维的方法将图像转化为伪点云，同样离不开深度估计。因此，基于图像的方法检测性能的提升与深度估计的质量密切相关。分析和改进深度估计的质量，并更好地将其与三维目标检测网络结合，将有效提高检测性能。这是未来许多研究者致力探索的方向。

2）提升异构数据的融合质量和效率。为了弥补图像数据缺失的精确空间信息，许多工作融合点云数据进行辅助检测，以提升检测性能。虽然特征级融合的方法可以综合利用两种模态数据的优势，产生更能表征目标物体信息的特征，但通常会引入对齐和计算消耗大的问题。另一方面，决策级融合的方法仅对两种模态数据的检测结果进行融合，对于在单种数据中都无法检测出来的目标，决策级融合的方法也无法提供有效的帮助。因此，未来的研究重点将是探索更有效的融合方式，并解决降低计算消耗和避免引入数据对齐等新问题。

3）引入时间序列辅助检测。在现实世界中，驾驶员依赖连续的视觉感知来获取周围环境信息。然而，目前很少有研究工作通过连续感知来辅助目标检测。BEVDet4D 已经证实时间信息对检测性能的提升有实质性帮助。为了进一步突破三维目标检测领域的性能瓶颈，引入时间序列辅助检测是一个值得探索的方向。

4）半监督、无监督学习的探索。三维目标检测数据集的制作成本高昂且耗时，通常涉及不同传感器之间的协作和大量的人力标注，并且标注过程中也不可避免会出现一些错误。与其他视觉领域相比，在三维目标检测领域，半监督和无监督学习的研究相对较少，这限制了该领域的发展。因此，探索半监督和无监督学习在三维目标检测中的应用，具有重要的研究意义。

5）提升模型的泛化性能。目前的三维目标检测方法大多基于特定数据集进行训练，在遇到数据分布显著变化的情况下，很难发挥原有的性能。而在实际的自动驾驶场景中，天气、光照、道路条件等因素可能与模型训练时使用的数据集存在较大差异，这使得现有模型难以胜任实际应用。为了更有利于算法从理论研究阶段走向实际应用，提升模型的泛化性能十分必要。

结束语　三维目标检测在自动驾驶感知环节中扮演着重要的角色，是自动驾驶后续一系列决策的基础，是自动驾驶汽车的“眼睛”。

本文系统总结了近年来的图像三维目标检测方法，深入剖析了各类方法的特点和局限性，并对比分析了各类算法的性能指标。此外，对三维目标检测领域的数据集和评价指标进行了说明，分析了目前图像三维目标检测领域存在的不足，指明了未来可能的研究方向。

随着深度学习和自动驾驶的发展，基于图像的三维目标检测由于其实际部署的价格优势将会得到工业界的青睐；图像与点云融合的三维目标检测方法由于其可结合图像数据和点云数据的优势，检测性能的上限比单一模态数据高很多，在一些对检测性能要求较高的场景中也会有较大的需求。我们相信，图像三维目标检测领域的研究会越来越深入，落地应用的项目会越来越多，前景一片光明。

参 考 文 献

[1] MAO J，SHI S，WANG X，et al. 3D Object Detection for Autonomous Driving：A Comprehensive Survey[J]. arXiv：2206.09474，2022.

[2] GUO G，ZHANG N. A survey on deep learning based face recognition[J]. Computer Vision and Image Understanding，2019，189：102805.

[3] LIU W，ANGUELOV D，ERHAN D，et al. SSD：Single Shot MultiBox Detector[C]// Proceedings of the 14th European Conference on Computer Vision. Amsterdam：Springer，2016：21-37.

[4] REDMON J，FARHADI A. YOLOv3：An Incremental Improvement[J]. arXiv：1804.02767，2018.

[5] BOCHKOVSKIY A，WANG C Y，LIAO H Y M. YOLOv4：Optimal Speed and Accuracy of Object Detection[J]. arXiv：2004.10934，2020.

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

[7] LANG A H，VORA S，CAESAR H，et al. PointPillars：Fast Encoders for Object Detection From Point Clouds[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach：IEEE，2019：12689-12697.

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

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

[10] CAO J L，LI Y L，SUN H Q，et al. A Survey on Deep Learning Based Visual Object Detection[J]. Journal of Image and Graphics，2022，27(6)：1697-1722.

[11] ZAMANAKOS G, TSOCHATZIDIS L, AMANATIADIS A, et al. A comprehensive survey of LIDAR-based 3D object detection methods with deep learning for autonomous driving[J]. Computers & Graphics, 2021, 99: 153-181.

[12] HUO W L, JING T, REN S. Review of 3D Object Detection for Autonomous Driving[J]. Computer Science, 2023, 50(7): 107-118.

[13] GEIGER A, LENZ P, URTASUN R. Are we ready for autonomous driving? the kitti vision benchmark suite[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Providence: IEEE, 2012: 3354-3361.

[14] CAESAR H, BANKITI V, LANG A H, et al. nuscenes: A multimodal dataset for autonomous driving[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020: 11621-11631.

[15] SUN P, KRETZSCHMAR H, DOTIWALLA X, et al. Scalability in perception for autonomous driving: Waymo open dataset [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020: 2446-2454.

[16] CHEN X, KUNDU K, ZHU Y, et al. 3D Object Proposals for Accurate Object Class Detection[C]// Proceedings of the 28th International Conference on Neural Information Processing Systems, 2015: 424-432.

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

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

[19] TIAN Z, SHEN C H, CHEN H, et al. FCOS: Fully Convolutional One-Stage Object Detection[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019: 9626-9635.

[20] WANG T, ZHU X, PANG J, et al. FCOS3D: Fully Convolutional One-Stage Monocular 3D Object Detection[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. Montreal: IEEE, 2021: 913-922.

[21] LIU Z, WU Z, TOTH R. SMOKE: Single-Stage Monocular 3D Object Detection via Keypoint Estimation[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Seattle: IEEE, 2020: 4289-4298.

[22] LI P, CHEN X, SHEN S. Stereo R-CNN Based 3D Object Detection for Autonomous Driving[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 7636-7644.

[23] REN S Q, HE K M, GIRSHICK R, et al. Faster R-CNN: towards real-time object detection with region proposal networks [J]. Advances in Neural Information Processing Systems, 2015, 39(6): 1137-1149.

[24] QIN Z, WANG J, LU Y. Triangulation Learning Network: From Monocular to Stereo 3D Object Detection[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 7607-7615.

[25] SUN J, CHEN L, XIE Y, et al. Disp R-CNN: Stereo 3D Object Detection via Shape Prior Guided Instance Disparity Estimation [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020: 10545-10554.

[26] XU Z, ZHANG W, YE X, et al. ZoomNet: Part-Aware Adaptive Zooming Neural Network for 3D Object Detection[C]// Proceedings of the AAAI Conference on Artificial Intelligence. New York: AAAI, 2020: 12557-12564.

[27] PENG L, WU X, YANG Z, et al. DID-M3D: Decoupling Instance Depth for Monocular 3D Object Detection[C]// Proceedings of the 17th European Conference on Computer Vision. Tel Aviv: Springer, 2022: 71-88.

[28] HUANG K C, WU T H, SU H T, et al. MonoDTR: Monocular 3D Object Detection with Depth-Aware Transformer[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022: 4002-4011.

[29] ZHANG R, QIU H, WANG T, et al. Monodetr: Depth-guided Transformer for Monocular 3D Object Detection[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. Paris: IEEE, 2023: 9155-9166.

[30] NAIDEN A, PAUNESCU V, KIM G, et al. Shift R-CNN: Deep Monocular 3D Object Detection With Closed-Form Geometric Constraints[C]// Proceedings of the IEEE International Conference on Image Processing. Taipei: IEEE, 2019: 61-65.

[31] CAI Y, LI B, JIAO Z, et al. Monocular 3D Object Detection with Decoupled Structured Polygon Estimation and Height-Guided Depth Estimation[C]// Proceedings of the AAAI Conference on Artificial Intelligence. New York: AAAI, 2020: 10478-10485.

[32] SHI X, YE Q, CHEN X, et al. Geometry-Based Distance Decomposition for Monocular 3D Object Detection[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021: 15152-15161.

[33] GU J, WU B, FAN L, et al. Homography Loss for Monocular 3D Object Detection[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022: 1070-1079.

[34] WANG T, XINGE Z H U, PANG J, et al. Probabilistic and geometric depth: Detecting objects in perspective[C]// Conference on Robot Learning. New York: PMLR, 2022: 1475-1485.

[35] MA X, LIU S, XIA Z, et al. Rethinking Pseudo-LiDAR Representation[C]// Proceedings of the 16th European Conference on Computer Vision. Glasgow: Springer, 2020: 311-327.

[36] WANG Y, CHAO W L, GARG D, et al. Pseudo-LiDAR From Visual Depth Estimation: Bridging the Gap in 3D Object Detection for Autonomous Driving[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 8437-8445.

[37] CHEN Y N, DAI H, DING Y. Pseudo-stereo for monocular 3d object detection in autonomous driving[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022: 887-897.

[38] SIMONELLIA,BULO S R,PORZI L,et al. Disentangling Monocular 3D Object Detection[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul:IEEE,2019:1991-1999.

[39] MOUSAVIAN A,ANGUELOV D,FLYNN J,et al. 3D Bounding Box Estimation Using Deep Learning and Geometry[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Honolulu:IEEE,2017:5632-5640.

[40] MA X,ZHANG Y,XU D,et al. Delving into Localization Errors for Monocular 3D Object Detection[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville:IEEE,2021:4719-4728.

[41] LIU X,XUE N,WU T,et al. Learning Auxiliary Monocular Contexts Helps Monocular 3D Object Detection[C]// Proceedings of the AAAI Conference on Artificial Intelligence. Online:AAAI,2022:1810-1818.

[42] SRIVASTAVA S,JURIE F,SHARMA G. Learning 2D to 3D Lifting for Object Detection in 3D for Autonomous Vehicles [C]// Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems. Macau:IEEE,2019:4504-4511.

[43] BELTRAN J,GUINDEL C,MORENO F M,et al. BirdNet:A 3D Object Detection Framework from LiDAR Information[C]// Proceedings of the 21st International Conference on Intelligent Transportation Systems. Maui:IEEE,2018:3517-3523.

[44] CHEN X,MA H,WAN J,et al. Multi-View 3D Object Detection Network for Autonomous Driving[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu:IEEE,2017:6526-6534.

[45] READING C,HARAKEH A,CHAE J,et al. Categorical Depth Distribution Network for Monocular 3D Object Detection[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville:IEEE,2021:8551-8560.

[46] RODDICK T,KENDALL A,CIPOLLA R. Orthographic Feature Transform for Monocular 3D Object Detection[J]. arXiv:1811.08188,2018.

[47] HUANG J,HUANG G,ZHU Z,et al. BEVDet:High-Performance Multi-Camera 3D Object Detection in Bird-Eye-View[J]. arXiv:2112.11790,2021.

[48] HUANG J,HUANG G. BEVDet4D:Exploit Temporal Cues in Multi-Camera 3D Object Detection[J]. arXiv:2203.17054,2022.

[49] LI Y,GE Z,YU G,et al. BevDepth:Acquisition of Reliable Depth for Multi-View 3D Object Detection[C]// Proceedings of the AAAI Conference on Artificial Intelligence. Washington:AAAI,2023:1477-1485.

[50] LI Z,WANG W,LI H,et al. BEVFormer:Learning Bird's-Eye-View Representation from Multi-Camera Images via Spatiotemporal Transformers[C]// European Conference on Computer Vision. Cham:Springer Nature Switzerland,2022:1-18.

[51] LI Y,BAO H,GE Z,et al. BEVStereo:Enhancing Depth Estimation in Multi-View 3D Object Detection with Temporal Stereo [C]// Proceedings of the AAAI Conference on Artificial Intelligence. Washington:AAAI,2023:1486-1494.

[52] YANG C,CHEN Y,TIAN H,et al. BEVFormer v2:Adapting Modern Image Backbones to Bird's-Eye-View Recognition via Perspective Supervision[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver:IEEE,2023:17830-17839.

[53] CHEN Y,LIU S,SHEN X,et al. DSGN:Deep StereoGeometry Network for 3D Object Detection[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle:IEEE,2020:12533-12542.

[54] CHEN Y,HUANG S,LIU S,et al. DSGN++:Exploiting Visual-Spatial Relation for Stereo-Based 3D Detectors[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2022,45(4):4416-4429.

[55] XU J,PENG L,CHENG H,et al. MonoNeRD:NeRF-like Representations for Monocular 3D Object Detection[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. Paris:IEEE,2023:6814-6824.

[56] ZHOU Z,DU L,YE X,et al. SGM3D:Stereo guided monocular 3D object detection[J]. IEEE Robotics and Automation Letters,2022,7(4):10478-10485.

[57] GAO A,PANG Y,NIE J,et al. Esgn:Efficient stereo geometry network for fast 3d object detection[J]. IEEE Transactions on Circuits and Systems for Video Technology,2024,34(4):2000-2009.

[58] YOU Y,WANG Y,CHAO W L,et al. Pseudo-LiDAR++:Accurate Depth for 3D Object Detection in Autonomous Driving [J]. arXiv:1906.06310,2019.

[59] QIAN R,GARG D,WANG Y,et al. End-to-End Pseudo-LiDAR for Image-Based 3D Object Detection[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle:IEEE,2020:5880-5889.

[60] KIM C,KIM U H,KIM J H. Self-supervised 3D Object Detection from Monocular Pseudo-LiDAR[C]// IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems. Bedford:IEEE,2022:1-6.

[61] MENG H,LI C,CHEN G,et al. Efficient 3D Object Detection Based on Pseudo-LiDAR Representation[J]. IEEE Transactions on Intelligent Vehicles,2024,9(1):1953-1964.

[62] SUN C,XU C,FANG W,et al. Monocular 3D Object Detection from Comprehensive Feature Distillation Pseudo-LiDAR[J]. IEEE Access,2023,11:98969-98976.

[63] VIANNEY J M U,AICH S,LIU B. RefinedMPL:Refined Monocular Pseudo LiDAR for 3D Object Detection in Autonomous Driving[J]. arXiv:1911.09712,2019.

[64] WENG X,KITANI K. Monocular 3D Object Detection with Pseudo-LiDAR Point Cloud[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision Workshop. Seoul:IEEE,2019:857-866.

[65] MA X,WANG Z,LI H,et al. Accurate Monocular 3D Object Detection via Color-Embedded 3D Reconstruction for Autonomous Driving[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul:IEEE,2019:6850-6859.

[66] LI C Y,KU J,WASLANDER S L. Confidence Guided Stereo 3D

Object Detection with Split Depth Estimation[C]// Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems. Las Vegas: IEEE, 2020: 5776-5783.
[67] WANG Y, YANG B, HU R, et al. PLUMENet: Efficient 3D Object Detection from Stereo Images[C]// Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems. Prague: IEEE, 2021: 3383-3390.
[68] MENG H, LI C, CHEN G, et al. Accurate and Real-Time Pseudo Lidar Detection: Is Stereo Neural Network Really Necessary? [J]. arXiv: 2206.13858, 2022.
[69] KU J, MOZIFIAN M, LEE J, et al. Joint 3D Proposal Generation and Object Detection from View Aggregation[C]// Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems. Madrid: IEEE, 2018: 5750-5757.
[70] LU H, CHEN X, ZHANG G, et al. Scanet: Spatial-Channel Attention Network for 3D Object Detection[C]// Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. Brighton: IEEE, 2019: 1992-1996.
[71] GUO R, LI D, HAN Y. Deep multi-scale and multi-modal fusion for 3D object detection[J]. Pattern Recognition Letters, 2021, 151: 236-242.
[72] WANG G, TIAN B, ZHANG Y, et al. Multi-View Adaptive Fusion Network for 3D Object Detection[J]. arXiv: 2011.00652, 2020.
[73] QI C R, LIU W, WU C, et al. Frustum PointNets for 3D Object Detection from RGB-D Data[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 918-927.
[74] CHARLES R Q, SU H, KAICHUN M, et al. PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017: 77-85.
[75] QI C R, YI L, SU H, et al. PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space[C]// Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017: 5105-5114.
[76] ZHAO X, LIU Z, HU R, et al. 3D Object Detection Using Scale Invariant and Feature Reweighting Networks[C]// Proceedings of the AAAI Conference on Artificial Intelligence. Honolulu: AAAI, 2019: 9267-9274.
[77] SINDAGI V A, ZHOU Y, TUZEL O. MVX-Net: Multimodal VoxelNet for 3D Object Detection[C]// Proceedings of the International Conference on Robotics and Automation. Montreal: IEEE, 2019: 7276-7282.
[78] YOO J H, KIM Y, KIM J, et al. 3D-CVF: Generating Joint Camera and LiDAR Features Using Cross-View Spatial Feature Fusion for 3D Object Detection[C]// Proceedings of the 16th European Conference on Computer Vision. Glasgow: Springer, 2020: 720-736.
[79] HUANG T, LIU Z, CHEN X, et al. EPNet: Enhancing Point Features with Image Semantics for 3D Object Detection[C]// Proceedings of the 16th European Conference on Computer Vision. Glasgow: Springer, 2020: 35-52.
[80] MAHMOUD A, HU J S K, WASLANDER S L. Dense Voxel Fusion for 3D Object Detection[C]// Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa: IEEE, 2023: 663-672.
[81] JIAO Y, JIE Z, CHEN S, et al. MSMDfusion: Fusing LIDAR and Camera at Multiple Scales With Multi-Depth Seeds for 3D Object Detection[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Paris: IEEE, 2023: 21643-21652.
[82] LIANG M, YANG B, WANG S, et al. Deep Continuous Fusion for Multi-Sensor 3D Object Detection[C]// Proceedings of the 15th European Conference on Computer Vision. Munich: Springer, 2018: 663-678.
[83] XIE L, XIANG C, YU Z, et al. PI-RCNN: An Efficient Multi-Sensor 3D Object Detector with Point-Based Attentive Cont-Conv Fusion Module[C]// Proceedings of the AAAI Conference on Artificial Intelligence. New York: AAAI, 2020: 12460-12467.
[84] XU D, ANGUELOV D, JAIN A. PointFusion: Deep Sensor Fusion for 3D Bounding Box Estimation[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 244-253.
[85] HE K, ZHANG X, REN S, et al. Deep Residual Learning for Image Recognition[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016: 770-778.
[86] WANG J, ZHU M, SUN D, et al. MCF3D: Multi-Stage Complementary Fusion for Multi-Sensor 3D Object Detection[J]. IEEE Access, 2019, 7(5): 90801-90814.
[87] LIANG M, YANG B, CHEN Y, et al. Multi-Task Multi-Sensor Fusion for 3D Object Detection[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 7337-7345.
[88] WANG C, MA C, ZHU M, et al. PointAugmenting: Cross-Modal Augmentation for 3D Object Detection[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021: 11789-11798.
[89] ZHANG Y, CHEN J, HUANG D. CAT-Det: Contrastively Augmented Transformer for Multimodal 3D Object Detection[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022: 898-907.
[90] WU X, PENG L, YANG H, et al. Sparse Fuse Dense: Towards High Quality 3D Detection with Depth Completion[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022: 5408-5417.
[91] BAI X, HU Z, ZHU X, et al. TransFusion: Robust LiDAR-Camera Fusion for 3D Object Detection with Transformers[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022: 1080-1089.
[92] LIANG T, XIE H, YU K, et al. Bevfusion: A simple and robust lidar-camera fusion framework[J]. Advances in Neural Information Processing Systems, 2022, 35: 10421-10434.
[93] VORA S, LANG A H, HELOU B, et al. PointPainting: Sequential Fusion for 3D Object Detection[C]// Proceedings of the

IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle：IEEE，2020：4603-4611.

[94] XU S，ZHOU D，FANG J，et al. Fusionpainting：Multimodal fusion with adaptive attention for 3d object detection[C]// IEEE International Intelligent Transportation Systems Conference. Indianapolis：IEEE，2021：3047-3054.

[95] CHEN Z，LI Z，ZHANG S，et al. AutoAlign：Pixel-Instance Feature Aggregation for Multi-Modal 3D Object Detection[C]// Proceedings of the 31st International Joint Conference on Artificial Intelligence. Vienna：IJCAI，2022：827-833.

[96] LI Y，QI X，CHEN Y，et al. Voxel Field Fusion for 3D Object Detection[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans：IEEE，2022：1110-1119.

[97] KIM T，GHOSH J. Robust Detection of Non-Motorized Road Users Using Deep Learning on Optical and LIDAR Data[C]// Proceedings of the 19th International Conference on Intelligent Transportation Systems. Rio de Janeiro：IEEE，2016：271-276.

[98] ASVADI A，GARROTE L，PREMEBIDA C，et al. Multimodal vehicle detection：fusing 3d-LIDAR and color camera data[J]. Pattern Recognition Letters，2018，115：20-29.

[99] PANG S，MORRIS D，RADHA H. CLOCs：Camera-LiDAR Object Candidates Fusion for 3D Object Detection[C]// Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems. Las Vegas：IEEE，2020：10386-10393.

[100] PANG S，MORRIS D，RADHA H. Fast-CLOCs：Fast Camera-LiDAR Object Candidates Fusion for 3D Object Detection[C]// Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa：IEEE，2022：3747-3756.

[101] GUO X，SHI S，WANG X，et al. LIGA-Stereo：Learning LiDAR Geometry Aware Representations for Stereo-Based 3D Detector [C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal：IEEE，2021：3133-3143.

ZHOU Yan，born in 1979，master，professor，master supervisor，is a member of CCF（No. 60294M）. Her main research interests include machine vision，graphics and image processing，and its applications in public safety，intelligent manufacturing and so on.

XU Yewen，born in 1999，postgraduate. His main research interests include computer vision and 3D object detection.

（责任编辑：何杨）