文章编号:1007－6735(2024)02－0103－17　DOI: 10.13255/j.cnki.jusst.20231101004

自动驾驶中基于深度学习的 3D 目标检测方法综述

梁振明，　黄影平，　宋卓恒，　丁建华
（上海理工大学，光电信息与计算机工程学院，上海 200093）

摘要：随着激光雷达传感器和深度学习技术的快速发展，针对自动驾驶 3D 目标检测算法的研究呈现爆发式增长。为了探究 3D 目标检测技术的发展和演变，对该领域中基于深度学习的 3D 检测算法进行了综述。根据车载传感器的不同，将当前基于深度学习的自动驾驶 3D 目标检测算法分为基于相机 RGB 图像、基于激光雷达点云、基于 RGB 图像–激光雷达点云融合的 3D 目标检测 3 种类型。在此基础上，分析了各类算法的技术原理及其发展历程，并根据平均检测精度（mAP）指标，对比了它们的性能差异与模型优缺点。最后，总结和展望了当前自动驾驶 3D 目标检测中仍然面临的技术挑战及未来发展趋势。

关键词：3D 目标检测；深度学习；自动驾驶；RGB 图像；激光雷达点云；多传感器融合
中图分类号：TP 391.4　　**文献标志码：**A

Survey on deep learning-based 3D object detection methods in autonomous driving

LIANG Zhenming,　HUANG Yingping,　SONG Zhuoheng,　DING Jianhua
(*School of Optical–Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China*)

Abstract: With the rapid developments of LiDAR sensors and deep learning technology, researches on 3D object detection for autonomous driving had witnessed remarkable growth. In order to explore the evolution of 3D object detection technology in autonomous driving, the existing deep learning-based 3D object detection methods were summarized. Based on the rely-on sensors, these methods could be divided into three classes: Camera RGB image-based, LiDAR point cloud-based and RGB image-LiDAR point cloud fusion-based 3D object detection. On this basis, the methodology and chronological overview of different kinds of methods were analyzed, and their 3D detection performance and characteristics were compared according to the mean average precision (mAP) metrics. Finally, the principal technical challenges and potential development trends in future autonomous driving 3D object detection researches were summarized and explored.

收稿日期: 2023−11−01
基金项目: 国家自然科学基金资助项目 (62276167)；上海市自然科学基金资助项目 (20ZR1437900)
第一作者: 梁振明（1991−），男，博士研究生．研究方向：深度学习、自动驾驶 3D 目标检测．E-mail: liangzhm@163.com
通信作者: 黄影平（1966−），男，教授．研究方向：智能汽车环境辨识与自主定位、汽车电子系统仿真测试、视觉检测技术等．E-mail: huangyingping@usst.edu.cn

Keywords: *3D object detection*; *deep learning*; *autonomous driving*; *RGB image*; *LiDAR point cloud*; *multi-sensors fusion*

1 概念及算法分类

目标检测是自动驾驶环境感知中的一项重要任务，旨在精确地找出无人车行驶场景中的感兴趣物体(障碍物)，并确定它们的类别、形状、位置等信息，为无人车的避障、交互、动作调整等提供信息指导。

当前自动驾驶中的目标检测技术分为两种：2D 目标检测和 3D 目标检测。2D 目标检测通常以相机生成的 RGB(红绿蓝)图像为数据源，以卷积神经网络(convolutional neural networks, CNN)等深度学习技术为工具，从图像平面获取目标物体的 2D 像素位置信息(u, v)和 2D 像素尺寸信息(w, h)。3D 目标检测主要以激光雷达(light detection and ranging, LiDAR)生成的点云为数据源，从三维立体的角度检测出目标物体的 3D 位置(x, y, z)、3D 尺寸(w, l, h)，以及航向角 θ 等信息。

与 2D 目标检测相比，3D 目标检测过程由于引入了对场景深度信息和物体航向角信息的测量，可以帮助无人车更准确地感知现实三维环境和目标障碍物，是对 2D 检测能力的升级。近年来，随着激光雷达传感器的快速发展，以及以 CNN 和 Transformer 等为代表的深度学习技术在交通场景感知领域取得的巨大成功，使得自动驾驶 3D 目标检测算法的研究迎来爆发式增长，并逐步发展成为一种趋势[1-2]。

如图 1 所示，根据车载传感器的不同，本文将当前自动驾驶中基于深度学习的 3D 目标检测算法分为基于相机 RGB 图像、基于激光雷达点云、基于 RGB 图像–激光雷达点云融合的 3D 目标检测 3 种类型，详细地介绍了这些算法的技术特点、发展历程，并对比分析了它们的优缺点和性能差异。

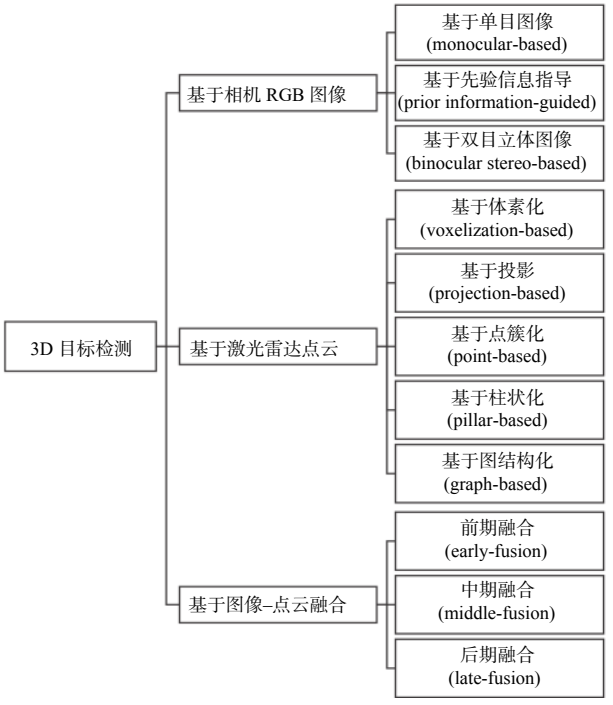

图 1 基于深度学习的自动驾驶 3D 目标检测算法分类

Fig.1 Taxonomy of deep learning-based 3D object detection methods in autonomous driving

2 常用数据集与性能评价指标

2.1 数据集

大型模型训练数据集是深度学习神经网络模型工作的基础，因此，表 1 首先总结了当前基于深度学习的自动驾驶 3D 目标检测研究领域中最常用的 3 款公共数据集。

表 1 常用的自动驾驶 3D 目标检测数据集

Tab.1 Commonly used 3D object detection datasets in autonomous driving

数据集	传感器数量	标注帧数	3D框标注数量	物体类别标注数量
KITTI [3]	相机-2, LiDAR-1	1.5万	8万	8类
nuScence [4]	相机-6, LiDAR-1	39万	140万	23类
Waymo [5]	相机-5, LiDAR-5	23万	1 200万	4类

a. KITTI 数据集。作为最常用的自动驾驶 3D 目标检测数据集，KITTI 共包括了 50 个场景下 1.5 万帧 RGB 图像与激光雷达点云数据。在 KITTI 中，多种自动驾驶中常见的目标障碍物，如：车辆(cars)、行人(pedestrians)、骑自行车的人(cyclists)等被规范地标注出，并且每种障碍物还根据其在场景中的可视区域大小、被遮挡程度，以及是否被截断等情况被分为简单(easy)、中等(moderate)和困难(hard)3 种检测难度级别。

b. nuScene 数据集。该数据集包含 1 000 个场景下 140 万帧 RGB 图像和 39 万帧激光雷达点云数据，每个场景都是由一段长 20 s、10 Hz(即每秒

10 帧）的连续帧构成，并且每个场景都标注出了 23 类常见的目标物体，共计 140 万个物体 3D 边界框信息。

c. Waymo 数据集。该数据集包含 1 150 个场景 23 万帧 RGB 图像和激光雷达点云数据，每个场景都是由以 10 Hz 捕获的 20 s 数据组成，总共标注了 4 类常见的目标物体共 1 200 万个 3D 边界框。

2.2　评价指标

性能评价指标是衡量 3D 检测模型性能好坏程度的度量标准，不同的数据集中，3D 目标检测模型的性能评价指标往往也不相同。

a. KITTI 数据集使用多类别平均检测精度（mean average precision, mAP）来评价模型 3D 检测性能的好坏，其计算方式如式（1）所示：

$$S_{\mathrm{mAP}}=\frac{1}{m}\frac{1}{N}\int_0^1 P(R)\mathrm{d}R=\frac{1}{m}\frac{1}{N}\int_0^1 \max\left[p(r')|r'\geqslant r\right]\mathrm{d}r \tag{1}$$

式中：P 和 R 分别代表了模型 3D 检测过程中 Precision-Recall（P-R）曲线的纵横坐标，通过以 3D IoU（intersection over union）作为模型预测 3D 边界框正确与否的判别标准，就可以绘制出模型 3D 检测过程的 Precision-Recall 曲线，进而可以计算出模型的 3D 检测评分 S_{mAP}。

b. nuScence 数据集使用 nuScence 检测评分（nuScenes detection score, NDS）作为评价指标来衡量模型 3D 检测性能的好坏，其计算方式如式（2）所示：

$$S_{\mathrm{NDS}}=\frac{1}{10}\left[5S_{\mathrm{mAP}}+\sum(1-\min(1,S_{\mathrm{mTP}}))\right] \tag{2}$$

式中：S_{mTP} 指的是模型的平均平移误差（average translation error, ATE）、平均尺度误差（average scale error, ASE）、平均朝向误差（average orientation error , AOE）、平均速度误差（average velocity error, AVE），以及平均属性误差（average attribute error, AAE）等 5 项指标的平均值（mean truth positive, mTP）。

c. Waymo 数据集使用航向加权平均检测精度（average precision weighted by heading , APH）作为评价指标来衡量模型 3D 检测性能的好坏，其计算方式如式（3）所示：

$$S_{\mathrm{APH}}=\frac{1}{N}\int_0^1 H(R)\mathrm{d}R=\frac{1}{N}\int_0^1 \max\left[h(r')|r'\geqslant r\right]\mathrm{d}r \tag{3}$$

H-R（precision weighted by heading-recall）曲线和 P-R 曲线基本一致，只不过在检测准确率和召回率的计算过程中，真阳性案例均要通过定义为 $\min(|\tilde{\theta}-\theta|,2\pi-|\tilde{\theta}-\theta|)/\pi$ 的航向精度进行加权，即将其规范到 [–π，π] 范围内，其中 $\tilde{\theta}$ 和 θ 分别指的是航向角的预测值和真值。

3　基于 RGB 图像的 3D 目标检测

车载相机体积小，价格低廉，其生成的 RGB 图像能够提供稠密的纹理与色彩信息，可以很好地还原出与人眼所见一致的场景语义表示。然而，图像是三维场景的二维投影，成像过程丢失了场景的深度信息，因此在基于 RGB 图像的 3D 目标检测过程中，对场景深度信息的恢复成为了该类方法的关键所在。当前，基于 RGB 图像的自动驾驶 3D 目标检测算法可以分为 3 类：基于单目图像（monocular-based）的 3D 目标检测、基于先验信息指导（prior information-guided）的单目 3D 目标检测，以及基于双目立体图像（binocular stereo-based）的 3D 目标检测。

3.1　基于单目 RGB 图像的 3D 目标检测

单目 RGB 图像无法提供准确的场景深度信息，因此，从单目图像中还原出三维场景的深度信息，成为基于单目图像 3D 目标检测算法的首要任务。

CenterNet 模型[6] 首先提出了一阶无锚点式的单目 3D 检测模式，该模型首先利用成熟的一阶无锚点 2D 检测器在单目图像上检测出目标物体的类别 c、中心点 (u, v)、观测角度 α，以及图平面中 (u, v) 与物体三维中心点 (x, y, z) 之间的相对位置偏置关系 (δ_x, δ_y) 和相对深度偏置关系 δ_d，然后再利用相机的内外参，即可大致估计出目标物体在世界坐标系下的真实三维位置信息 (x, y, z) 及航向角信息 θ 等，即：$d^c\cdot[\ u^c\quad v^c\quad 1\ ]^{\mathrm{T}}=KT\cdot[\ x\quad y\quad z\quad 1\ ]^{\mathrm{T}}$，$\theta=\alpha+\arctan(x/z)$。其中：$u^c=u+\delta_x$；$v^c=v+\delta_y$；$d^c=\sigma^{-1}(1/\delta_d+1)$，$\sigma$ 表示 sigmoid 函数；K，T 分别表示相机的内外参。在此基础上，SMOKE 模型[7] 和 MonoFlex 模型[8] 先后对此作出改进：在 SMOKE 模型中，除了基于 $(u, v, \alpha, \delta_x, \delta_y, \delta_d)$ 等关键点的物体 3D 边界框估计方式以外，还提出了一种“参数多步骤解”方法来优化物体 3D 边界框的估计过程，并加速模型的训练；MonoFlex 模型在基于 $(u, v, \alpha, \delta_x, \delta_y, \delta_d)$ 等关键点的 3D 边界框估计过程中加

入了对物体深度信息的估计，通过对网络提取的单目图像数据特征进行解耦，以多检测头的形式同时对物体的边界框关键点和深度信息进行估计，然后再以估计的深度信息反哺物体3D边界框关键点的回归过程，以此来优化模型的3D检测结果。

M3D-RP模型[9]首次提出了一阶有锚点式的单目3D检测模式，该方法依赖一系列预定义的2D-3D边界框匹配锚点，通过首先利用成熟的2D检测器在单目图像上检测出目标物体的2D边界框，再利用2D-3D锚点匹配的方法推断出物体的3D边界框。在此基础上，Groomed-NMS模型[10]提出了一种新型的物体候选框筛选方法GrooMeD-NMS，通过将传统的物体候选框筛选方法——非最大抑制（non-maximum suppression，NMS）转变成一种可微分的数学矩阵运算，解决了模型训练与推断过程中物体候选框参数不匹配的问题，从而优化了模型的输出。M3DSSD模型[11]在物体2D-3D边界框锚点匹配过程中提出了基于特征对齐的优化策略，并且在物体深度信息估计过程中加入了基于多尺度特征采样的非对称注意力模块，优化了3D检测模型对目标物体的深度信息估计过程。

ROI-10D模型[12]首次提出了二阶式的单目3D检测模式，该模型首先使用二阶单目检测模型Faster RCNN[13]在图像平面上生成出目标物体的2D感兴趣区域，然后再以此区域预测目标物体的2D中心位置(u, v)、观测角α，以及相机坐标系下物体的相对深度δ_d和相对三维尺度$(\delta_x, \delta_y, \delta_z)$，最后利用相机的$KT$矩阵还原出世界坐标系下目标物体的3D边界框表示。GS3D模型[14]在单目图像特征提取过程中，除了对目标物体结构信息的提取外，还加入对物体视觉表面信息的提取，这种信息的加入可以帮助模型消除物体3D边界框边缘模糊的问题。此外，MonoDIS[15]，MonoGRNet[16]，MonoRCNN[17]等模型也都对基于二阶检测方法的单目3D检测模式作了相应改进，其中：MonoDIS提出了一种新的模型训练损失函数；MonoGRNet提出了一种半监督式的物体中心点深度估计与定位方法；MonoRCNN提出了一种基于“几何距离因子分解”的物体深度估计方法。

3.2 基于先验信息指导的单目3D目标检测

基于先验信息指导的单目3D目标检测指的是借助单目图像的某些特定先验知识，来弥补单目图像在3D检测过程中无法提供场景深度信息的弊端，从而优化模型的3D检测效果。

Mono3D模型[18]首先提出了基于物体形状、路面假设、物体语义信息等先验信息为指导的单目3D检测模式，该模型通过结合上述先验信息和某些特定类别物体的3D候选框在图像平面的2D投影与物体真实2D边界框之间的差异，从而找出该物体的最优3D候选框。在此基础上，DeepMENTA[19]，Mono3D++[20]，3D-RCNN[21]等模型也先后提出了以物体轮廓信息为指导的单目3D检测模式，这类方法首先将物体的3D轮廓表示成一系列连接线组成的网格，且网格的顶点个数固定不变，3D检测模型可以通过预测物体3D轮廓上网格顶点的位置，重构出目标物体的3D边界框。

Deep3DBox模型[22]提出了基于物体2D-3D边界框几何连续性限制为指导的单目3D检测模式，该方法假设物体的2D边界框是其3D边界框在前向位置的投影，那么2D-3D边界框之间的这种几何限制就可以帮助网络模型更好地完成3D检测任务。基于此模式，Shift R-CNN[23]，RTM3D[24]，MonoPair[25]等模型也相继提出了基于物体几何连续性限制为指导的单目3D检测方法，其中：Shift R-CNN在基于IoU的3D候选框筛选模块中提出了一种新型的模型训练损失函数，以此来优化物体3D边界框的选择过程；RTM3D模型以物体2D-3D边界框上透视关键点的联系性代替了物体2D-3D边界框连接边之间的几何限制，从而提高了模型的推断速度；MonoPair模型通过考虑同一场景下不同物体之间的配对关系，以此来解决单目图像中被遮挡物体难以被检测的问题。

MultiFusion模型[26]首次提出了基于场景深度信息为指导的单目3D检测模式，该模型通过引入一个辅助性的单目图像深度估计网络来生成场景的深度信息，然后将其以辅助特征通道的形式并入到单目图像的数据特征图中，从而使基于单目图像的3D检测过程也能具备目标物体的深度信息。在此之后，PCT模型[27]提出了一种轻量级的基于场景深度信息为指导的3D检测方法，通过借助单目深度估计网络生成的图像深度图，并以连续串行的图像特征学习方式，由粗至细循序渐进地优化目标物体的3D定位。

Pseudo-LiDAR模型[28]首次提出了基于伪点云

的 3D 检测模式，所谓伪点云指的是一系列由深度图转化生成的具有深度信息的特殊像素点：对于单目图像中的任一像素点(u, v)及其由深度估计网络生成的深度信息 d，如果知道相机的光学焦距(f_u, f_v)和成像基准点(c_u, c_v)，那么就可以通过 $x=(u-c_u)z/f_u$，$y=(v-c_v)z/f_v$，$z=d$的方式将像素点 (u, v) 转换成其伪点云的格式，继而可以使用基于点云的 3D 检测模型完成 3D 目标检测任务。在 Pseudo-LiDAR 基础上，AM3D 模型[29] 加入了单目图像的 RGB 颜色信息作为辅助、PS-LiDAR 模型[30] 加入了图像实例分割后物体的语义信息作为辅助，这些辅助信息的加入均帮助基于伪点云的 3D 检测模型实现了更准确的预测。近来，Ma 等[31] 对基于伪点云的 3D 检测模式进行了深入分析，提出了性能更优的 PatchNet 模型，该模型认为伪点云的 3D 表达能力主要源于其坐标转换过程，而非数据本身。因此，伪点云数据可进一步转换成具备深度信息的 2D 图像格式，继而可以使用成熟的 2D CNN 模型来处理。这不但能加快模型的运行速度，还能提高模型的 3D 检测精度。

3.3　基于双目立体图像的 3D 目标检测

相比于单目图像，双目图像可以通过立体匹配的方式获取较为准确的场景深度信息，因此，基于双目立体图像的 3D 目标检测算法研究也成为自动驾驶场景感知技术中的一大热点。

3DOP 模型[32] 首次提出了基于全场景双目视差图的 3D 检测模式，该模型首先对同一时刻产生的左右两帧图像 I_l 和 I_r 进行立体匹配，那么左图 I_l 上任一像素点(u, v)的深度值 d 就可以表示为 $d=(fb)/p$。其中：p 为该像素点通过立体匹配获得的视差值；f 表示左相机的焦距；b 表示左右相机基线之间的距离。在通过上述方法遍历获取图像中每个像素点的深度值之后，基于双目图像的 3D 目标检测迎刃而解。YOLOStereo3D 模型[33] 在 3DOP 模型基础上提出了更轻量级的双目立体匹配模块，并且通过使用多尺度特征融合策略和一阶式的目标检测策略，使得 YOLOStereo3D 模型在取得良好检测结果的同时，其推断速度达到了 12.5 帧/s。

Disp R-CNN 模型[34] 提出了基于双目实例化视差图的 3D 检测模式，与全场景的双目视差图生成过程不同，实例级的图像视差图生成方式更有利于图像中目标物体的 3D 回归。因为在成熟的图像实例化分割网络帮助下，3D 检测过程中与目标物体 3D 边界框回归无关的背景像素点会被提前过滤掉，Disp R-CNN 模型实现了比 3DOP 模型更好的检测效果。在此基础上，DSGN 模型[35] 提出了基于图像实例化分割的一阶双目 3D 检测策略、RT3D-Stereo 模型[36] 提出了基于 2D 目标检测和 2D 实例化分割融合的双目 3D 检测策略等。通过不同形式的改进，基于双目图像实例化视差图的 3D 检测模型在检测精度或运行速度上都得到了相应提高。

Stereo R-CNN 模型[37] 提出了一种双网络互联式的双目 3D 检测模式，该模型同时使用两个 2D 检测器对同一时刻生成的左右两帧图像分别进行特征提取，并同时生成目标物体的左右两个感兴趣区域，然后通过对两个感兴趣区域进行对齐，并根据 2D-3D 边界框之间的关键点约束，从而实现对目标物体 3D 边界框的精准回归。除此之外，TL-Net 模型[38] 也采用了双网络互联式的 3D 检测方式，并且 TL-Net 还在左右两个感兴趣区域对齐阶段引入一种三角定位法，进一步优化了目标物体的 3D 定位。

由于双目图像立体匹配形成的视差图可以在相机内参的帮助下转换成场景深度图，进而可以转化为伪点云数据，因此，可以使用基于伪点云的 3D 检测方法完成对双目立体图像的 3D 检测。Pseudo-LiDAR++模型[39] 遵循了上述思路，提出了基于伪点云的双目 3D 检测模式，并且优化了双目图像生成视差图的过程，使模型在“远处物体难以被检测”的问题上得到相应改善。Pseudo-Stereo 模型[40] 受伪点云的启发，提出了基于“伪立体图像”的双目 3D 检测模式，该模型首先提出了一种“左图生成右图”从而得到双目立体图像的方法，然后再利用生成的双目立体图像完成 3D 目标检测任务。CG-Stereo 模型[41] 提出了一种基于全场景伪点云的双目 3D 检测方法，并且在双目视差图转换为伪点云的过程中，CG-Stereo 模型还会预测生成一个“目标物体置信度分布图”，并以此图来指导基于伪点云的 3D 检测过程，进一步优化模型的 3D 检测性能。

3.4　模型性能及优缺点说明

从性能角度来看，基于双目立体图像的 3D 检测方法其检测结果整体上最优，其次是基于先验信息指导的单目 3D 检测方法，最后是基于单目图

像的3D检测方法。从模型优缺点角度来看，由于单目图像只能依赖相机的内外参估计图像中目标物体的深度信息，精度有限，导致基于单目图像的3D检测性能整体上与其他类型的方法有一定的差距。在图像中某些特定先验信息的指导下，单目3D检测模型的表现确实可以得到进一步优化，但是这些先验信息的获取途径往往需要额外的计算代价，因此，这类3D检测方法的泛化性能往往不尽如人意。基于双目立体图像的3D检测方法通常要比前两种方法具有更好的检测效果，但是双目视差图的产生过程也会给3D检测模型带来更高的计算压力，并且双目视差图对场景深度信息的恢复能力有限，无法保证模型对远距离物体的精准3D识别。

4 基于激光雷达点云的3D目标检测

激光雷达点云以激光光束反射点的形式描述场景，当光束照射到物体上时会被反射回来，通过计算光束发射–接收之间的时间差，即可精确地获取目标物体的三维空间位置信息。然而，激光雷达点云数据相对稀疏，且格式无序、不规则，常规的CNN模型无法直接对其进行处理，建模过程往往需要一些特殊的操作。当前，对激光雷达点云常用的处理手段包括：投影法(projection)、体素化(voxelization)、点簇化(point-based)、柱状化(pillar-based)，以及图结构化(graph-based)等。

4.1 基于点云投影的3D目标检测

投影法指的是将三维激光雷达点云按照某种投影方法投影成2D图像的形式，然后再利用常规的2D CNN模型来处理。当前，对于激光雷达点云的投影方法主要有两种：前向视场(range-view, RV)投影和鸟瞰图(bird-eye-view, BEV)投影。

VeloFCN模型[42]首次提出了基于激光雷达点云RV投影的3D检测模式，该模型首先将三维的激光雷达点云前向投影成一种特殊格式的2D Point-map。在此Point-map中，每个点云数据点(x, y, z)的2D位置可以表示为横向$\theta=\arctan(y/x)$、竖向$\varphi=\arcsin\left[z/(x^2+y^2+z^2)\right]$的形式。此外，对于每个$(\theta,\varphi)$数据点，其数据特征由深度值$d=\mathrm{sqrt}(x^2+y^2)$和高度值$z$构成。在遍历构建完此2D Point-map之后，就可以利用全卷积神经网络对其进行2D检测，检测得出的物体2D边界框中也就相应包含了其三维形状还原所需的全部信息。在此之后，LaserNet模型[43]提出了速度超快的基于激光雷达点云RV投影的3D检测方式。该模型首先使用全卷积神经网络在点云RV投影图上检测出目标物体的2D边界框，然后再对2D边界框中的每个“像素”点都预测一个3D边界框，最后通过统计这些3D边界框在俯视角度上的概率分布，让模型学习一个概率分布模型来指导多个3D边界框进行聚类，从而得出最终的物体3D边界框预测结果。

BirdNet模型[44]首次提出了基于激光雷达点云BEV投影的3D检测模式。该模型首先将激光雷达点云俯向投影成包含数据点高度、反射强度、密度等信息在内的2D BEV-map，然后再使用Faster RCNN 2D检测器在此BEV-map上进行2D目标检测，最后再结合路面信息和物体距离地面的高度信息，共同还原出目标物体的3D边界框。此后，BirdNet+模型[45]在BirdNet基础上作了部分改进，将BirdNet中的3D边界框回归方式变成了一种端到端的处理方式，取消了费时的3D边界框后处理操作，使得模型的检测准确度和推断速度都得到一定程度的提高。PIXOR模型[46]摒弃了BirdNet和BirdNet+模型中的物体3D边界框二阶回归方式，提出了基于一阶检测模式的点云BEV投影3D检测模式，使得模型的3D检测速度得到进一步提升。

HDNet模型[47]认为高清地图(HD Maps)提供的先验信息可以提高3D检测器的性能，因此提出了基于高清地图与激光雷达点云BEV投影融合的3D检测模式，并且对于无法提供高清地图的场景，HDNet模型还提出了一个基于激光雷达点云的高清地图估计模块，以此来补全模型中多模态数据的融合过程。在此之后，MVF模型[48]提出了基于激光雷达点云BEV投影和RV投影融合的3D检测方法，通过利用不同视角下的点云数据，并配合提出的动态体素化点云特征提取策略，使得MVF模型的3D检测性能在众多基于激光雷达点云投影的3D目标检测算法中脱颖而出。

4.2 基于点云体素化的3D目标检测

体素化是将点云划分为空间连续、格式规则的三维网格格式，每个网格区域的局部特征由该网格内所有数据点特征的平均值或最大值来表示，然后再利用3D CNN对其进行特征提取和建模处理。

VoxelNet 模型[49] 首次提出了基于激光雷达点云体素化的 3D 检测模式，该模型提出了一种通用的点云体素化特征表达方式——体素特征编码层（voxel feature encode, VFE）。在 VFE 层中，对于每一个点云体素块，首先求得其质心$(\bar{v}_x,\bar{v}_y,\bar{v}_z)$，然后借此质心将该体素块内所有数据点的特征 $[x_i, y_i, z_i, r_i]$ 增强为 $\boldsymbol{p}_i=[x_i,y_i,z_i,r_i, x_i-\bar{v}_x,y_i-\bar{v}_y,z_i-\bar{v}_z]^{\mathrm{T}}$，之后再利用 MLP 模块和最大池化操作即可得出该体素块的局部特征表示。经过 VFE 层的堆叠，并以 3D CNN 模块来提取各体素块内的数据特征，即可完成 3D 目标检测任务。在此基础上，SA-SSD 模型[50] 对点云体素块的划分过程作了改进，由于原始的点云体素化过程涉及大量的数据遍历计算，耗时巨大且工作效率低下，因此，SA-SSD 提出了一种新型的基于张量计算的点云体素块划分方式，大大提高了模型的计算效率。HVNet 模型[51] 对 VFE 层进行了部分改进，提出了多尺度混合体素特征编码层 HVFE，与原始的 VFE 层相比，HVFE 层可以同时提取多层不同尺度的点云体素块特征，帮助 3D 检测模型实现更精确的物体 3D 边界框预测。

由于在基于点云体素化的 3D 目标检测模型中，3D CNN 模块通常必不可少但计算量巨大，因此对 3D 卷积操作进行改进也是点云体素化方法研究中的热点。VoxelRCNN 模型[52] 提出在激光雷达点云体素化处理之后，再对其进行 BEV 投影，然后使用 2D CNN 模型对物体候选框进行估计，从而减小模型的计算量。SECOND 模型[53]、Vote3Deep 模型[54]，以及 Focals-Conv 模型[55] 分别提出了基于哈希运算、基于 Voting 机制，以及基于动态自适应机制的 3D 稀疏卷积模块，通过对原始耗时的 3D 卷积模块进行不同方式的稀疏化改进，大大降低了激光雷达点云体素化过程中 3D CNN 模块的计算量，进而加速模型的运算。

PV-RCNN 模型[56] 提出了一种基于激光雷达点云体素化和点簇化联合的 3D 检测模式，与以 VFE 层提取点云数据特征的方式不同，PV-RCNN 提出使用点簇化方法来提取各体素块内点云的数据特征，从而避免了 VFE 层中因池化操作带来的数据特征细节丢失问题。之后，PV-RCNN++ 模型[57] 在 PV-RCNN 基础上作了部分改进，提出了效率和拟物化程度更高的体素块数据点采样方法，并且对 PV-RCNN 模型中基于点簇化的体素块数据特征聚合过程作了改进，进一步提高了 PV-RCNN++模型的运行速度。

TED 模型[58] 提出了基于等变变换（transformation-equivariant, TE）的激光雷达点云体素化 3D 检测模式，该模型认为在基于 3D 稀疏卷积的点云体素块特征提取过程中，卷积模块只能保证点云的平移不变性特点，不能保证其旋转和反转不变性特点，因此 TED 提出了一种基于特征通道共享变换的等变稀疏卷积模块来改善这一弊端。此外，为了消除特征通道共享后的冗余，并优化物体候选框的筛选过程，TED 还提出了一种等变体素池化操作，以此帮助模型实现更快更准的 3D 检测。PVT-SSD 模型[59] 提出了基于 Transformer 的激光雷达点云体素化 3D 检测模式，该模型首先利用体素化方式对点云进行规则化处理，以及对各体素块内的局部数据特征进行提取，然后将每个体素块视为 Transformer 模型中的一个 Token 单元，使用 Transformer 结构对各体素块之间的全局信息进行提取，以此获取激光雷达点云的局部和全局数据信息，并用于 3D 目标检测任务。

4.3　基于点云点簇化的 3D 目标检测

PointNet[60] 和 PointNet++模型[61] 的提出为基于点云点簇化处理的 3D 目标检测算法研究提供了理论基础。在这类方法中，点云中的每个数据点都被视为一个独立的操作单元，通过相应的数据点分簇处理和多层结构共享 MLP 层对簇内数据点的特征进行提取，使得特征相似的数据点趋向一致、特征不同的数据点差异性被放大，通过判别这种数据差异性，即可将点云中不同的物体区分开来。

PointRCNN 模型[62] 率先提出了基于激光雷达点云点簇化处理的 3D 检测模式，该模型首先使用 PointNet++对激光雷达点云进行特征提取，并将点云中属于目标物体的数据点分割出来，然后依据这些数据点，采用二阶回归的方式对物体 3D 边界框进行预测，即：先生成物体 proposal、再对 proposal 进行细化。与之类似，STD 模型[63] 也采用了二阶式的激光雷达点云点簇化 3D 检测方式，但与 PointRCNN 不同的是，在物体 proposal 细化阶段，STD 采用了体素化的方式对 proposal 中的点云数据进行特征提取，这样可以将点云的稀疏特征稠密化，从而节省计算空间。LiDAR-RCNN 模型[64] 经过对上述方法进行分析，发现基于点云点簇化和二阶回归方式的 3D 检测模型都忽视了物

体 proposal 中的尺度模糊问题：由于点云的不规则性和局部密度不统一性，导致由 PointNet 或 PointNet++生成的物体 proposal 中可能存在少量偏移、斜切等尺度模糊问题，进而影响物体 proposal 的细化工作，因此，LiDAR-RCNN 模型针对这一问题进行了补救，进一步优化了模型的 3D 检测结果。

3DSSD 模型[65]提出了一阶式的激光雷达点云点簇化 3D 检测模式，移除了二阶检测方法中先生成物体 proposal 再对 proposal 进行细化的环节，从而有效地提升了模型的运算速度。此外，3DSSD 模型还提出了一种新型的点云融合采样策略，相较于原始 PointNet++结构中的最远点采样方法，新提出的采样方法可以更有效地保留点云数据点的特征信息，进而更利于物体 3D 边界框的精确回归。除此以外，IA-SSD 模型[66]也提出了一阶式的点云点簇化 3D 检测方式，通过利用点云语义分割的思想，使用 PointNet++对激光雷达点云进行实例级大尺度下采样，在将点云的数据规模降低到非常小的同时，还保证了点云中目标物体的语义信息大致得以保留，从而在不影响模型检测精度的情况下，大大提高了模型的运行速度。

PointFormer 模型[67]提出了基于 Transformer 的点云点簇化 3D 检测模式，该模型对点云的特征提取方式与点簇化方式类似，也是先对点云进行分簇处理，然后再对各簇内的数据点进行特征提取，即 PointNet++结构中 Set Abstraction（SA）层的作用。但与原始 SA 层不同的是，PointFormer 中使用的是基于注意力机制的 Transformer 层对点簇内数据点的特征信息进行提取，原始 SA 层中使用的是多层共享的 MLP 网络。此外，PointFormer 模型还提出了 3 种不同尺度的 Transformer 模块：以点簇内各数据点为基本单元的局部 Transformer 模块、以各点簇为基本单元的全局 Transformer 模块，以及用于多尺度特征融合的局部–全局融合 Transformer 模块，通过借助 Transformer 强大的数据特征提取能力，以及多尺度的数据特征融合策略，使得 PointFormer 可以替换作为多款成熟 3D 目标检测模型的主干网络，以此优化模型的 3D 检测能力。

4.4 基于点云柱状化的 3D 目标检测

柱状化是将不规则的三维点云用规则的三维柱形（pillar）区域划分表示，与体素化中三维网格的划分形式不同，柱状化划分过程对点云在高度方向上不作区分，即在相同鸟瞰区域内的所有点云数据点都将被划分到同一个柱形区域中，不管其高度为多少。之后，同一柱形区域内的所有数据点在经过同等特征变换之后，都沿着高度方向作特征压缩，从而将三维点云特征转化为 2D BEV 形式，继而可以使用 2D CNN 模型处理，不再需要耗时的 3D 卷积操作。

PointPillar 模型[68]首先提出了基于激光雷达点云柱状化处理的 3D 检测模式，该模型首先使用柱状化方式对点云进行划分，对于每一个柱状区域中的点云数据，PointPillar 模型首先使用 PointNet 对其进行特征提取，然后沿着高度方向对所有柱状区域内的点云数据特征作 BEV 投影，并以投影之后的 2D BEV 特征图作为 3D 检测模型的数据输入。由于 BEV 投影已经将三维的点云特征转换为二维，因此对于激光雷达点云 2D BEV 特征图的处理只需要常规的 2D CNN 模型，无需 3D 卷积操作。在此基础上，Pillar-OD 模型[69]对 PointPillar 模型作了部分改进：在 Pillar-OD 中，激光雷达点云除了作 Pillar 形式的 BEV 投影外，还作了前向 RV 投影，两种不同视角的投影特征图会被拼接在一起，共同完成 3D 目标检测任务。

SST 模型[70]提出了基于 Transformer 的点云柱状化 3D 检测模式，该模型认为在基于 PointNet 的点云柱状区域特征提取过程中，数据下采样过程会造成部分点云特征信息丢失，不利于小目标物体的精准检测，因此，SST 模型提出使用 Transformer 模块对柱状区域内的点云数据点进行特征提取，且因 Transformer 中对数据划分 Token 的操作不涉及下采样操作，可以帮助模型实现更好的 3D 检测效果，尤其是对小目标物体的检测。

PillarNeXt 模型[71]从点云特征聚合和网络模块优化的角度出发，对点云的柱状化处理过程进行了深入分析：首先分别测试了体素化和柱状化等点云处理方式在点云特征提取和聚合过程中对计算资源的消耗情况以及对模型性能的影响，证明了基于点云柱状化处理的 3D 检测器可以实现优于或等同于基于点云体素化 3D 检测器的检测效果；其次，PillarNeXt 借鉴了 2D 检测器的成功经验，增大了柱状化点云特征提取过程中的网络模块感受野，并且引入了多尺度特征融合策略来丰富柱状区域内的点云数据特征，从而优化了 3D 检测模型的检测效果。

4.5 基于点云图结构化的 3D 目标检测

图结构化指的是将不规则的点云依据数据点

之间的拓扑连接关系，构建成图结构型数据，然后再利用图数据的处理方式来建模。与点簇化处理方式类似，图结构化的点云特征提取方法可以直接将网络模型应用在不规则的激光雷达点云上，无需任何预处理操作。

Point-GNN 模型[72]首次提出了基于激光雷达点云图结构化处理的 3D 检测模式，在此模型中，点云各局部范围内的数据点首先会被相互连接，构建成图结构型数据，然后再利用图卷积神经网络来逐层提取不同尺度范围内的点云数据特征，并以此完成 3D 目标检测任务。与基于点云点簇化的 3D 检测方法相比，基于点云图结构化的 3D 检测模型不仅能够提取激光雷达点云中各数据点的特征信息，还能提取不同数据点之间的拓扑连接关系，帮助 3D 检测模型更好地完成 3D 检测任务。

SVGA-Net 模型[73]和 STA-GCN 模型[74]先后将视觉注意力机制引入到基于点云图结构化的 3D 目标检测中，在以图结构表示的点云特征提取过程中，通过使用局部注意力机制给图结构中的每一个节点分配不同的权重，可以区分出点云中不同数据点在卷积过程中的不同重要性。通过使用全局注意力机制为不同的子图分配权重，可以区分出点云图结构中各个局部区域在物体识别过程中的不同重要性。通过使用不同层次、不同属性的注意力机制，可以帮助模型更有效地提取点云特征，从而优化模型的 3D 检测性能。

4.6　模型性能及优缺点说明

从性能角度来看，基于激光雷达点云不同处理方法的各类 3D 检测模型性能各异，但整体水平差异性不大，各类模型的检测效果各有优劣。从模型优缺点角度来看，基于投影的点云处理方法虽然可以将三维的激光雷达点云投影到二维，然后利用成熟的 2D CNN 模型来处理，但是数据降维过程难免会造成部分点云信息丢失，从而影响模型的 3D 检测性能。体素化可以更好地保留点云的原始数据信息，更有利于目标物体的精准回归，然而高负载的 3D 卷积模块在基于点云体素化的 3D 检测模型中很难完全被取代，因此计算效率问题成为该类方法中的一大痛点。点云点簇化处理的优点在于可以直接将网络模型应用到不规则的激光雷达点云上，无需任何预处理操作，但是点簇化过程中的簇划分半径、簇内数据点的个数都不易确定，可能对基于点云体素化 3D 检测模型的运行效率会有影响。在基于激光雷达点云柱状化的 3D 检测模型中，由于耗时的 3D 卷积模块被去除，因此，该类方法的执行效率要高于其他方法。然而，点云柱状化处理过程还是存在划分区域难以定量的问题。柱状区域的面积越小，则越容易保留点云数据的细化特征，但是会加剧模型的计算量；柱状区域的面积增大，则会减少模型计算过程中的内存占用，并会加速模型的运行，但是也会相应地丢失更多点云数据特征。点云的图结构化表示可以使 3D 检测模型获取更丰富的数据特征信息，从而帮助模型实现更精确的 3D 目标检测，然而点云的图结构化过程，以及图神经网络对点云特征的提取过程都涉及了大量的数据遍历计算，这种高负荷的计算代价也限制了基于点云图结构化的 3D 目标检测模型在大规模激光雷达点云中的应用。

5　基于图像–点云融合的 3D 目标检测

RGB 图像和激光雷达点云的数据模态不同，且有着各自的成像优势和数据缺点，通过将两种数据进行融合，借助多源数据之间的互补优势，从而可以获取更好的 3D 检测效果。当前，RGB 图像与激光雷达点云的融合方式主要包含 3 种：前期融合（early-fusion）、中期融合（middle-fusion）、后期融合（late-fusion）。

5.1　基于图像–点云前期融合的 3D 目标检测

图像–点云的前期融合又称数据级融合，是将两种模态的数据在原始数据层级进行融合，产生包含更多特征信息（通道）的新数据集合，然后以此为模型输入来完成自动驾驶 3D 目标检测任务。前期融合可以更好地保留 RGB 图像与激光雷达点云的原始数据信息，从而确保 3D 检测模型能够获取更多的多模态数据细节，更有利于其对目标障碍物的精准回归。

F-PointNet 模型[75]首先提出了基于 2D 目标检测的图像–点云前期融合 3D 检测模式，该方法首先利用一个 2D 检测器在 RGB 图像上将目标物体的 2D 边界框检测出来，然后将其投影到激光雷达点云上，并筛选出位于 2D 边界框之内的点云数据点，然后将这些数据点的三维坐标信息与图像上物体 2D 边界框内的像素信息拼接在一起，从而形成一个包含多模态数据信息在内的目标物体 3D 点云视锥空间。在此之后，F-PointNet 使用了两个 PointNet 结构，一个用于将物体 3D 点云视锥空间

内属于目标物体的数据点分割出来，另一个利用分割出的数据点回归出目标物体的3D边界框。在F-PointNet基础上，F-ConvNet模型[76]提出将图像–点云融合形成的3D点云视锥空间进行分段处理，然后每个分段分别使用单独的PointNet结构进行特征提取，最后综合所有分段的特征共同预测出目标物体的3D边界框信息。F-PointPillar模型[77]在F-PointNet基础上作了两处改进：一个是将基于PointNet的3D点云视锥空间分割过程改为基于高斯分布统计的数据点分割；另一个是将基于PointNet的3D边界框回归过程改为基于Pillar的3D边界框回归。

RoarNet模型[78]提出了基于2D-3D边界框几何连续性限制的图像–点云前期融合3D检测模式。与视锥方式类似，RoarNet首先使用一个2D检测模块在RGB图像上将目标物体的2D边界框检测出来；然后将其投影到对应的激光雷达点云上，并依据2D边界框在点云投影方向上的几何位置，生成一系列可能的物体3D候选框；最后再对其进行筛选，从而选择出一个最优答案。与之类似，General-Fusion模型[79]提出了使用激光雷达点云BEV投影与图像2D检测结果进行融合，然后再根据2D-3D边界框之间的几何限制，从一系列3D候选框中筛选出目标物体的真实3D边界框。MVP模型[80]提出借助相机的内参矩阵，可以将基于RGB图像的2D检测结果转化成一系列的3D视觉虚拟点(伪点云)，然后再利用这些虚拟点来补全和增强稀疏的激光雷达点云数据，从而优化基于激光雷达点云的3D检测模型的检测结果。

PointPainting模型[81]和IPOD模型[82]都提出了基于2D语义分割的图像–点云前期融合3D检测模式，这类方法首先使用2D图像语义分割模型将RGB图像上目标物体的语义像素点分割出来，然后再将这些像素点投影到激光雷达点云上，最后这些被像素语义信息“染色”过的激光雷达点云会被送入基于激光雷达点云的3D检测器中来完成3D目标检测任务。在此基础上，Complexer-YOLO[83]提出了一阶式的图像语义分割与激光雷达点云融合3D检测方法，通过提出一种新的3D候选框筛选指标SRTs(Scale-Rotation-Translation score)来代替费时的3D IoU筛选方法，从而加速了模型的训练与推断。

5.2 基于图像–点云中期融合的3D目标检测

RGB图像与激光雷达点云的中期融合指的是两种模态的数据在经过各自独立的特征提取之后，两种不同形式的数据特征feature或者两种数据各自生成的目标物体proposal被融合在一起，共同完成目标物体的3D检测任务。

MV3D模型[84]首先提出了基于图像–点云proposal融合的3D检测模式，该模型同时以激光雷达点云的FV投影、BEV投影，以及RGB图像等多视角数据为输入，各种数据首先各自生成不同视角下目标物体的proposal区域，然后各proposal区域经过裁剪、对齐与池化等操作被拼接融合在一起，共同完成目标物体的3D边界框回归任务。在此基础上，PI-RCNN模型[85]提出将图像2D语义分割生成的物体2D proposal与基于PointNet++生成的激光雷达点云3D proposal进行融合、SCANet模型[86]提出使用空间注意力机制对RGB图像和激光雷达点云BEV投影图生成的2D proposal与BEV proposal进行融合、FusionPainting模型[87]提出将图像2D语义分割生成的物体2D proposal与激光雷达点云3D语义分割生成的3D proposal进行融合、Graph-RCNN模型[88]提出将图像2D目标检测生成的2D proposal与基于体素化生成的激光雷达点云3D proposal进行融合。此外，Fusion-RCNN模型[89]和TransFusion模型[90]都提出基于Transformer的图像–点云proposal融合方案，即两种数据在单独生成各自模态的proposal区域之后，再利用基于注意力机制的Transformer模块将两种proposal融合在一起。

PointFusion模型[91]提出了基于图像–点云feature融合的3D检测模式，该模型首先同时使用PointNet和ResNet分别对激光雷达点云和RGB图像进行特征提取，然后将两种不同模态的数据特征级联拼接在一起，共同送入到一个MLP结构进行目标物体的3D边界框回归。在此之后，MVX-Net模型[92]提出使用Faster-RCNN和VoxelNet分别对图像和激光雷达点云进行特征提取，然后将两种数据特征拼接在一起；LaserNet++模型[93]提出使用ResNet对图像和激光雷达点云FV投影图同时进行特征提取，然后再把两者融合拼接在一起；3D-CVF模型[94]使用VoxelNet和ResNet对激光雷达点云和6方位视角的RGB图像同时进行特征提取，然后再将多种特征图对齐融合在一起；SimpleFusion[95]使用ResNet和PointPillar分别对图像和激光雷达点云进行特征提取，并在特征融合过程中提出了一个“动态融合模块”来优化图像

–点云的特征融合过程。除上述模型外，VF-Fusion 模型[96] 还提出了一种新的基于“点–射线”形式的图像–点云特征拼接方式，相比于之前基于“点–点”形式的特征拼接方式，VF-Fusion 能够实现更合理的图像–点云特征融合效果；DeepFusion 模型[97] 提出了一种基于注意力机制的图像–点云特征对齐方法，以此也能优化 RGB 图像和激光雷达点云的特征融合过程。

EPNet 模型[98] 提出了基于图像–点云多尺度 feature 融合的 3D 检测模式，与之前的 feature 级融合方法相比，EPNet 中图像与点云的数据特征不仅在特征提取网络的最后进行融合，而且在两个特征提取网络的每一层都进行融合，这种层次更丰富的图像–点云融合特征可以帮助 3D 检测模型实现更精准的目标检测。在此基础上，MSMDFusion 模型[99] 提出了基于多视角图像–点云多尺度 feature 融合的 3D 检测方法、DAMFusion 模型[100] 提出了基于注意力机制的图像–点云多尺度 feature 融合方法。近来，CAT-Det 模型[101] 提出了基于 Transformer 的图像–点云多尺度 feature 融合方法，该模型分别使用 Pointformer 和 Imageformer 分别对激光雷达点云和 RGB 图像同时进行特征提取，并且在每一层提取层都使用跨模态的 Transformer 模块对两种特征进行融合，最后再以此多尺度、多模态的数据特征进行目标物体的 3D 回归。

AVOD 模型[102] 提出了基于图像–点云 feature-proposal 双层级中期融合的 3D 检测模式，该模型的激光雷达点云和 RGB 图像特征融合过程，不仅包含 feature 级的融合，还包含了 proposal 级的融合。这种 feature-proposal 双层级的融合方式虽然在一定程度上可能会增加模型的计算代价，但是对模型 3D 检测性能的提高也有一定的帮助作用。此外，MMF 模型[103] 也采用了这种 feature-proposal 双层级特征融合方式，并且在 feature 级特征融合过程中还加入了多尺度特征融合，以此获取更丰富的多模态数据特征，帮助了 MMF 模型在顺利完成 3D 检测任务的同时，还具备了 2D 检测、定位，以及深度补全等功能。

5.3　基于图像–点云后期融合的 3D 目标检测

图像–点云的后期融合指的是利用不同的网络模块在不同模态的数据上分别作决策判断，然后再将各种决策结果依据某种规则整合起来，生成一个综合的、效果更优的检测结果。

CLOCs 模型[104] 首先提出了基于图像–点云后期融合的 3D 检测模式，该模型首先使用独立的 2D 检测器和 3D 检测器分别生成目标物体的 2D 候选框和 3D 候选框，然后根据同一物体 2D-3D 边界框之间的几何限制将两种候选框融合在一起从而产生更准确的模型输出。在此之后，Fast-CLOCs 模型[105] 又通过改进 CLOCs 模型中的 2D 检测器和 3D 检测器，提出了效率更高、更轻量的图像–点云后期融合 3D 检测模式，进一步提高了 CLOCs 模型的检测精度和运算速度。

5.4　模型性能及优缺点说明

从性能角度来看，当前基于图像–点云中期融合的 3D 检测方法普遍研究较多，其 3D 检测效果相较于其他两类方法也总体较优。从模型优缺点角度来看，基于图像–点云前期融合的 3D 检测方式是将 RGB 图像与激光雷达点云的原始信息叠加作为 3D 检测模型的输入，因此能够更多地保留图像和点云的原始数据信息。然而，前期融合方法一般需要引入独立且完整的 2D 检测器或 2D 语义分割模型，因此可能会增加 3D 检测模型的整体计算代价。基于图像–点云中期融合的 3D 检测方法可以获取更丰富的多模态数据特征，更有利于目标物体 3D 边界框的精确回归，因此，该类方法的 3D 检测效果整体上要优于图像–点云前期融合模型和后期融合模型。然而，多模态数据的特征级融合要比数据级融合复杂得多，无论是融合过程中不同数据特征之间的对齐，还是网络模型的构建及训练，都会给基于图像–点云中期融合的 3D 检测模型带来挑战。基于图像–点云后期融合的 3D 目标检测主要聚焦于不同模态数据的决策性输出，这种方式不用考虑两种模态数据在数据格式层面和特征差异层面的融合难点，融合过程相较于前两种方式更直接，模型也更容易部署。然而，图像–点云后期融合方式没有充分利用 RGB 图像和激光雷达点云之间的数据互补优势，导致这类方法的检测鲁棒性要略逊于前两种方法。

6　总结与展望

虽然当前基于深度学习的自动驾驶 3D 目标检测技术研究火热，但仍然存在一些问题没有得到有效解决，具体如下：

a. 基于图像的 3D 检测方法借助了深度学习技

术在图像研究领域的多年积累，然而，由于图像对于目标深度感知的先天缺陷，基于图像方法的检测性能明显不如基于点云的方法。如何有效利用图像成本低、计算快这些有利条件的同时，提高图像的深度信息捕获能力，仍然需要更多的探索。

b. 激光雷达探测距离远，对目标的深度信息感知准确，能形成3D目标检测的可靠数据源。但是激光雷达硬件成本高，计算代价大，并且点云是三维的，具有稀疏、无序和不规则性，属于非欧式数据，使得基于点云的3D目标检测方法比基于图像的方法配备成本更高，算法更复杂。如何降低激光雷达的成本，对点云数据直接建模、更加高效地提取不规则点云的信息特征都是该类方法需要进一步完善的地方。

c. 基于图像–点云融合的3D目标检测方法利用多源数据的互补优势，弥补各自的缺陷，能够更好地实现3D目标检测，成为目前的主流研究方向。但是此类方法计算成本大，网络构架复杂，在多模态数据的一体化建模、融合算法的网络化、不同层级融合的网络构架设置，以及如何开展训练等多方面存在问题。

针对上述问题，笔者认为基于深度学习的自动驾驶3D目标检测未来的研究趋势和发展方向集中在以下几个方面：

a. 构建高性能的视差检测深度学习模型，建立具有高精度视差标注信息的双目或多目图像数据集。

现有的基于双目或多目图像立体匹配获取的场景深度信息精度有限，如何通过深度学习方法获得高精度的视差图以精确还原场景的深度信息是计算机视觉的基础问题[106]，也将极大地推动基于图像的3D目标检测。为了达成这一目的，需要建立具有高精度视差标注信息的双目或多目图像数据集，以及探索构建更合理有效的图像视差值检测模型。

b. 探索适用于非欧式数据特征提取的卷积算子。

卷积算子是各类神经网络得以高效运行的核心，但是在当前成熟的深度学习技术中，核心的卷积算子都仅适用于规则化的欧式数据，未见适用于非欧数据特征提取的核心卷积算子，因此，研究适用非欧式点云数据特征提取的卷积算子（比如当前热门的PointFormer[107]，GraphFormer[108]等），继而提高对点云的直接处理能力是提升3D目标检测的重要途径。

c. 多模态数据融合网络构架的研究。

基于图像–点云融合的3D目标检测是目前的主流研究方向，但深度学习的融合网络构架复杂，融合方式多种多样，采用何种构架与方式将两种数据进行有效融合以获取最优的融合效果，以及如何训练都是关键的问题，亟待系统性的研究比较。

d. 多任务一体化建模方法的研究。

目标检测只是智能汽车环境感知任务中的一项，其他任务还包括对道路、交通标志，以及周边环境的检测与辨识。现有的基于深度学习的环境感知方法通常将这些任务分为检测、分类和场景分割问题，采用多个不同的网络来实现这些任务，这对车载计算设备的性能和成本提出了极大的挑战。如何利用多源传感器的融合优势，采用一个统一的网络，实现在一次前向传播中同时完成多个任务，这也是目前亟待解决的问题之一。

参考文献：

[1] YURTSEVER E, LAMBERT J, CARBALLO A, et al. A survey of autonomous driving: common practices and emerging technologies[J]. IEEE Access, 2020, 8: 58443–58469.

[2] ARNOLD E, AL-JARRAH O Y, DIANATI M, et al. A survey on 3D object detection methods for autonomous driving applications[J]. IEEE Transactions on Intelligent Transportation Systems, 2019, 20(10): 3782–3795.

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

[4] CAESAR H, BANKITI V, LANG A H, et al. nuScenes: A multimodal dataset for autonomous driving[C]// Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020: 11618–11628.

[5] SUN P, KRETZSCHMAR H, DOTIWALLA X, et al. Scalability in perception for autonomous driving: Waymo open dataset[C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020: 2443–2451.

[6] ZHOU X, WANG D, KRÄHENBÜHL P. Objects as points [EB/OL]. (2019-04-16). https://arxiv.org/abs/

1904.07850.

[7] LIU Z C, WU Z Z, TÓTH R. SMOKE: single-stage monocular 3D object detection via keypoint estimation[C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Seattle: IEEE, 2020: 4289–4298.

[8] ZHANG Y P, LU J W, ZHOU J. Objects are different: flexible monocular 3D object detection[C]//Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021: 3288–3297.

[9] BRAZIL G, LIU X M. M3D-RPN: monocular 3D region proposal network for object detection[C]//Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019: 9286–9295.

[10] KUMAR A, BRAZIL G, LIU X M. GrooMeD-NMS: grouped mathematically differentiable NMS for monocular 3D object detection[C]//Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021: 8969–8979.

[11] LUO S J, DAI H, SHAO L, et al. M3DSSD: monocular 3D single stage object detector[C]//Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021: 6141–6150.

[12] MANHARDT F, KEHL W, GAIDON A. ROI-10D: monocular lifting of 2D detection to 6D pose and metric shape[C]//Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 2064–2073.

[13] REN S Q, HE K M, GIRSHICK R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[C]//Proceedings of the 28th International Conference on Neural Information Processing Systems. Montreal: MIT Press, 2015. 91–99.

[14] LI B Y, OUYANG W L, SHENG L, et al. GS3D: an efficient 3D object detection framework for autonomous driving[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 1019–1028.

[15] SIMONELLI A, BULÒ S R, PORZI L, et al. Disentangling monocular 3D object detection[C]//Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019: 1991–1999.

[16] QIN Z Y, WANG J L, LU Y. MonoGRNet: a geometric reasoning network for monocular 3D object localization[C]//Proceedings of the 33rd AAAI Conference on Artificial Intelligence. Honolulu: AAAI, 2019: 8851–8858.

[17] SHI X P, YE Q, CHEN X Z, et al. Geometry-based distance decomposition for monocular 3D object detection[C]//Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021: 15152–15161.

[18] CHEN X Z, KUNDU K, ZHANG Z Y, et al. Monocular 3D object detection for autonomous driving[C]//Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016: 2147–2156.

[19] CHABOT F, CHAOUCH M, RABARISOA J, et al. Deep MANTA: a coarse-to-fine many-task network for joint 2D and 3D vehicle analysis from monocular image[C]//Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017: 1827–1836.

[20] HE T, SOATTO S. Mono3D++: monocular 3D vehicle detection with two-scale 3D hypotheses and task priors[C]//Proceedings of the 33rd AAAI Conference on Artificial Intelligence. Honolulu: AAAI, 2019: 8409–8416.

[21] KUNDU A, LI Y, REHG J M. 3D-RCNN: instance-level 3D object reconstruction via render-and-compare[C]//Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 3559–3568.

[22] MOUSAVIAN A, ANGUELOV D, FLYNN J, et al. 3D bounding box estimation using deep learning and geometry[C]//Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017: 5632–5640.

[23] 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 2019 IEEE International Conference on Image Processing. Taipei: IEEE, 2019: 61–65.

[24] LI P X, ZHAO H C, LIU P F, et al. RTM3D: real-time monocular 3D detection from object keypoints for autonomous driving[C]//Proceedings of the 16th European Conference on Computer Vision. Glasgow: Springer, 2020: 644–660.

[25] CHEN Y J, TAI L, SUN K, et al. MonoPair: monocular 3D object detection using pairwise spatial relationships[C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020: 12090–12099.

[26] XU B, CHEN Z Z. Multi-level fusion based 3D object detection from monocular images[C]//Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 2345–2353.

[27] WANG L, ZHANG L, ZHU Y, et al. Progressive

coordinate transforms for monocular 3d object detection[J]. Advances in Neural Information Processing Systems, 2021, 34: 13364-13377.

[28] 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 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 8437–8445.

[29] MA X Z, WANG Z H, LI H J, et al. Accurate monocular 3D object detection via color-embedded 3D reconstruction for autonomous driving[C]//Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019: 6850–6859.

[30] WENG X S, KITANI K. Monocular 3D object detection with pseudo-LiDAR point cloud[C]//Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision Workshop. Seoul: IEEE, 2019: 857–866.

[31] MA X Z, LIU S N, XIA Z Y, et al. Rethinking pseudo-LiDAR representation[C]//Proceedings of the 16th European Conference on Computer Vision. Glasgow: Springer, 2020: 311–327.

[32] CHEN X Z, KUNDU K, ZHU Y K, et al. 3D object proposals using stereo imagery for accurate object class detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(5): 1259–1272.

[33] LIU Y X, WANG L J, LIU M. YOLOStereo3D: A step back to 2D for efficient stereo 3D detection[C]// Proceedings of the 2021 IEEE International Conference on Robotics and Automation. Xi'an: IEEE, 2021: 13018–13024.

[34] SUN J M, CHEN L H, XIE Y M, et al. Disp R-CNN: stereo 3D object detection via shape prior guided instance disparity estimation[C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020: 10545–10554.

[35] CHEN Y L, LIU S, SHEN X Y, et al. DSGN: Deep stereo geometry network for 3D object detection[C]// Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020: 12533–12542.

[36] KöNIGSHOF H, SALSCHEIDER N O, STILLER C. Realtime 3D object detection for automated driving using stereo vision and semantic information[C]//Proceedings of the 2019 IEEE Intelligent Transportation Systems Conference. Auckland: IEEE, 2019: 1405–1410.

[37] LI P L, CHEN X Z, SHEN S J. Stereo R-CNN based 3D object detection for autonomous driving[C]//Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 7636–7644.

[38] QIN Z Y, WANG J L, LU Y. Triangulation learning network: from monocular to stereo 3D object detection[C]//Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 7607–7615.

[39] YOU Y R, WANG Y, CHAO W L, et al. Pseudo-LiDAR++: accurate depth for 3D object detection in autonomous driving[C]//Proceedings of the 8th International Conference on Learning Representations. Addis Ababa: ICLR, 2020.

[40] CHEN Y N, DAI H, DING Y. Pseudo-stereo for monocular 3D object detection in autonomous driving[C]//Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022: 877–887.

[41] LI C Y, KU J, WASLANDER S L. Confidence guided stereo 3D object detection with split depth estimation[C]//Proceedings of the 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems. Las Vegas: IEEE, 2020: 5776–5783.

[42] LI B, ZHANG T, XIA T. Vehicle detection from 3d lidar using fully convolutional network [EB/OL]. (2016-08-29).https://arxiv.org/abs/1608.07916.

[43] MEYER G P, LADDHA A, KEE E, et al. LaserNet: an efficient probabilistic 3D object detector for autonomous driving[C]//Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 12669–12678.

[44] BELTRÁN J, GUINDEL C, MORENO F M, et al. BirdNet: a 3D object detection framework from LiDAR information[C]//Proceedings of the 2018 21st International Conference on Intelligent Transportation Systems. Maui: IEEE Press, 2018: 3517–3523.

[45] BARRERA A, GUINDEL C, BELTRÁN J, et al. BirdNet+: end-to-end 3D object detection in LiDAR bird' s eye view[C]//Proceedings of the 2020 IEEE 23rd International Conference on Intelligent Transportation Systems. Rhodes, 2020: 1–6.

[46] YANG B, LUO W J, URTASUN R. PIXOR: real-time 3D object detection from point clouds[C]//Proceedings of the 2018 IEEE/CVF conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 7652–7660.

[47] YANG B, LIANG M, URTASUN R. HDNET: exploiting HD maps for 3D object detection[C]//Proceedings of the 2nd Annual Conference on Robot Learning. Zürich: CoRL, 2018: 146–155.

[48] ZHOU Y, SUN P, ZHANG Y, et al. End-to-end multi-view fusion for 3D object detection in LiDAR point

clouds[C]//Proceedings of the 3rd Annual Conference on Robot Learning. Osaka: CoRL, 2020: 923–932.

[49] ZHOU Y, TUZEL O. VoxelNet: end-to-end learning for point cloud based 3D object detection[C]//Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 4490–4499.

[50] HE C H, ZENG H, HUANG J Q, et al. Structure aware single-stage 3D object detection from point cloud[C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020: 11870–11879.

[51] YE M S, XU S J, CAO T Y. HVNet: hybrid voxel network for LiDAR based 3D object detection[C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020: 1628–1637.

[52] DENG J, SHI S, LI P, et al. Voxel r-cnn: towards high performance voxel-based 3d object detection[C]// Proceedings of the 35th AAAI Conference on Artificial Intelligence. British Columbia, 2021, 35(2): 1201-1209.

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

[54] ENGELCKE M, RAO D, WANG D Z, et al. Vote3Deep: fast object detection in 3D point clouds using efficient convolutional neural networks[C]// Proceedings of the 2017 IEEE International Conference on Robotics and Automation. Singapore: IEEE, 2017: 1355–1361.

[55] CHEN Y K, LI Y W, ZHANG X Y, et al. Focal sparse convolutional networks for 3D object detection[C]// Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022: 5418–5427.

[56] SHI S S, GUO C X, JIANG L, et al. PV-RCNN: point-voxel feature set abstraction for 3D object detection[C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020: 10526–10535.

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

[58] WU H, WEN C, LI W, et al. Transformation-equivariant 3D object detection for autonomous driving[C]// Proceedings of the 37th AAAI Conference on Artificial Intelligence. Washington DC,2023, 37(3): 2795-2802.

[59] YANG H H, WANG W X, CHEN M H, et al. PVT-SSD: single-stage 3D object detector with point-voxel transformer[C]//Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023: 13476–13487.

[60] CHARLES R Q, SU H, MO K C, et al. PointNet: deep learning on point sets for 3D classification and segmentation[C]//Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017: 77–85.

[61] 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. Long Beach: Curran Associates Inc. , 2017: 5105–5114.

[62] SHI S S, WANG X G, LI H S. PointRCNN: 3D object proposal generation and detection from point cloud[C]//Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 770–779.

[63] YANG Z T, SUN Y N, LIU S, et al. STD: sparse-to-dense 3D object detector for point cloud[C]//Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019: 1951–1960.

[64] LI Z C, WANG F, WANG N Y. LiDAR R-CNN: an efficient and universal 3D object detector[C]//Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021: 7542–7551.

[65] YANG Z T, SUN Y N, LIU S, et al. 3DSSD: point-based 3D single stage object detector[C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020: 11037–11045.

[66] ZHANG Y F, HU Q Y, XU G Q, et al. Not all points are equal: learning highly efficient point-based detectors for 3D LiDAR point clouds[C]//Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022: 18931–18940.

[67] PAN X R, XIA Z F, SONG S J, et al. 3D object detection with pointformer[C]//Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021: 7459–7468.

[68] LANG A H, VORA S, CAESAR H, et al. PointPillars: fast encoders for object detection from point clouds[C]//Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 12689–12697.

[69] WANG Y, FATHI A, KUNDU A, et al. Pillar-based object detection for autonomous driving[C]//Proceedings of the 16th European Conference on Computer Vision. Glasgow: Springer, 2020: 18–34.

[70] FAN L, PANG Z Q, ZHANG T Y, et al. Embracing single stride 3D object detector with sparse transformer[C]//Proceedings of the 2022 IEEE/CVF

Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022: 8448–8458.

[71] LI J Y, LUO C X, YANG X D. PillarNeXt: rethinking network designs for 3D object detection in LiDAR point clouds[C]//Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023: 17567–17576.

[72] SHI W J, RAJKUMAR R. Point-GNN: graph neural network for 3D object detection in a point cloud[C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020: 1708–1716.

[73] HE Q D, WANG Z N, ZENG H, et al. SVGA-Net: sparse voxel-graph attention network for 3D object detection from point clouds[C]//Proceedings of the 36th AAAI Conference on Artificial Intelligence. Vancouver, 2022: 870–878.

[74] WANG L, SONG Z Y, ZHANG X Y, et al. SAT-GCN: self-attention graph convolutional network-based 3D object detection for autonomous driving[J]. Knowledge-Based Systems, 2023, 259: 110080.

[75] QI C R, LIU W, WU C X, et al. Frustum PointNets for 3D object detection from RGB-D data[C]//Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 918–927.

[76] WANG Z X, JIA K. Frustum ConvNet: sliding frustums to aggregate local point-wise features for amodal 3D object detection[C]//Proceedings of the 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems. Macau: IEEE, 2019: 1742–1749.

[77] PAIGWAR A, SIERRA-GONZALEZ D, ERKENT Ö, et al. Frustum-PointPillars: a multi-stage approach for 3D object detection using RGB camera and LiDAR[C]//Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021: 2926–2933.

[78] SHIN K, KWON Y P, TOMIZUKA M. RoarNet: a robust 3D object detection based on RegiOn approximation refinement[C]//Proceedings of the 2019 IEEE Intelligent Vehicles Symposium. Paris: IEEE, 2019: 2510–2515.

[79] DU X X, ANG M H, KARAMAN S, et al. A general pipeline for 3D detection of vehicles[C]//Proceedings of the 2018 IEEE International Conference on Robotics and Automation. Brisbane: IEEE, 2018: 3194–3200.

[80] YIN T, ZHOU X, KRÄHENBÜHL P. Multimodal virtual point 3d detection[J]. Advances in Neural Information Processing Systems, 2021, 34: 16494-16507.

[81] VORA S, LANG A H, HELOU B, et al. PointPainting: sequential fusion for 3D object detection[C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020: 4603–4611.

[82] YANG Z, SUN Y, LIU S, et al. Ipod: intensive point-based object detector for point cloud [EB/OL]. (2018-12-13). https://arxiv.org/abs/1812.05276.

[83] SIMON M, AMENDE K, KRAUS A, et al. Complexer-YOLO: real-time 3D object detection and tracking on semantic point clouds[C]//Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Long Beach: IEEE, 2019: 1190–1199.

[84] CHEN X Z, MA H M, WAN J, et al. Multi-view 3D object detection network for autonomous driving[C]//Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017: 6526–6534.

[85] XIE L, XIANG C, YU Z X, et al. PI-RCNN: an efficient multi-sensor 3D object detector with point-based attentive cont-conv fusion module[C]//Proceedings of the 34th AAAI Conference on Artificial Intelligence. New York: AAAI, 2020: 12460–12467.

[86] LU H H, CHEN X S, ZHANG G Y, et al. SCANet: spatial-channel attention network for 3D object detection[C]//Proceedings of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing. Brighton: IEEE, 2019: 1992–1996.

[87] XU S Q, ZHOU D F, FANG J, et al. FusionPainting: multimodal fusion with adaptive attention for 3D object detection[C]//Proceedings of the 2021 IEEE International Intelligent Transportation Systems Conference. Indianapolis: IEEE, 2021: 3047–3054.

[88] YANG H H, LIU Z L, WU X P, et al. Graph R-CNN: towards accurate 3D object detection with semantic-decorated local graph[C]//Proceedings of the 17th European Conference on Computer Vision. Tel Aviv: Springer, 2022: 662–679.

[89] XU X L, DONG S C, XU T F, et al. FusionRCNN: LiDAR-camera fusion for two-stage 3D Object detection[J]. Remote Sensing, 2023, 15(7): 1839.

[90] BAI X Y, HU Z Y, ZHU X G, et al. TransFusion: robust LiDAR-camera fusion for 3D object detection with transformers[C]//Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022: 1080–1089.

[91] XU D F, ANGUELOV D, JAIN A. PointFusion: deep sensor fusion for 3D bounding box estimation[C]//Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 244–253.

[92] SINDAGI V A, ZHOU Y, TUZEL O. MVX-Net:

multimodal VoxelNet for 3D object detection[C]//Proceedings of the 2019 International Conference on Robotics and Automation. Montreal: IEEE, 2019: 7276–7282.

[93] MEYER G P, CHARLAND J, HEGDE D, et al. Sensor fusion for joint 3D object detection and semantic segmentation[C]//Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Long Beach: IEEE, 2019: 1230–1237.

[94] 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.

[95] ZHANG Y C, GUO Y, NIU H B, et al. SimpleFusion: 3D object detection by fusing RGB images and point clouds[C]//Proceedings of the 2023 International Conference on Pattern Recognition, Machine Vision and Intelligent Algorithms. Beihai: IEEE, 2023: 47–51.

[96] LI Y W, QI X J, CHEN Y K, et al. Voxel field fusion for 3D object detection[C]//Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022: 1110–1119.

[97] LI Y W, YU A W, MENG T J, et al. DeepFusion: Lidar-camera deep fusion for multi-modal 3D object detection[C]//Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022: 17161–17170.

[98] HUANG T T, LIU Z, CHEN X W, 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.

[99] JIAO Y, JIE Z Q, CHEN S X, et al. MSMDFusion: fusing LiDAR and camera at multiple scales with multi-depth seeds for 3D object detection[C]//Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023: 21643–21652.

[100] LIU Z, CHENG J, FAN J, et al. Multi-modal fusion based on depth adaptive mechanism for 3D object detection[J]. IEEE Transactions on Multimedia, 2023:1-11.

[101] ZHANG Y N, CHEN J X, HUANG D. CAT-DET: contrastively augmented transformer for multimodal 3D object detection[C]//Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022: 898–907.

[102] KU J, MOZIFIAN M, LEE J, et al. Joint 3D proposal generation and object detection from view aggregation[C]//Proceedings of the 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems. Madrid: IEEE, 2018: 1–8.

[103] LIANG M, YANG B, CHEN Y, et al. Multi-task multi-sensor fusion for 3D object detection[C]//Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 7337–7345.

[104] PANG S, MORRIS D, RADHA H. CLOCs: camera-LiDAR object candidates fusion for 3D object detection[C]//Proceedings of the 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems. Las Vegas: IEEE, 2020: 10386–10393.

[105] PANG S, MORRIS D, RADHA H. Fast-CLOCs: fast camera-LiDAR object candidates fusion for 3D object detection[C]//Proceedings of the 2022 IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa: IEEE, 2022: 3747–3756.

[106] LAGA H, JOSPIN L V, BOUSSAID F, et al. A survey on deep learning techniques for stereo-based depth estimation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 44(4): 1738–1764.

[107] CHEN Y J, YANG Z L, ZHENG X W, et al. PointFormer: a dual perception attention-based network for point cloud classification[C]//Proceedings of the 16th Asian Conference on Computer Vision. Macao: Springer, 2022: 432–449.

[108] CAI D, LAM W. Graph transformer for graph-to-sequence learning[C]//Proceedings of the 34th AAAI Conference on Artificial Intelligence. New York: AAAI, 2020: 7464–7471.

（编辑：丁红艺）