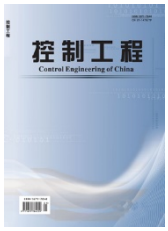

控制工程
Control Engineering of China
ISSN 1671-7848,CN 21-1476/TP

《控制工程》网络首发论文

题目：基于深度学习的 3D 目标检测算法综述
作者：张新宇，徐子贤，闫冬梅，沙晓鹏，顾德英
DOI：10.14107/j.cnki.kzgc.20210180
收稿日期：2021-03-22
网络首发日期：2022-07-28
引用格式：张新宇，徐子贤，闫冬梅，沙晓鹏，顾德英．基于深度学习的 3D 目标检测算法综述[J/OL]．控制工程. https://doi.org/10.14107/j.cnki.kzgc.20210180

控 制 工 程
Control Engineering of China

DOI: 10.14107/j.cnki.kzgc.20210180

基于深度学习的 3D 目标检测算法综述

张新宇[1]，徐子贤[2]，闫冬梅[1]，沙晓鹏[1]，顾德英[1]
(1. 东北大学秦皇岛分校 控制工程学院，河北 秦皇岛 066004；2. 大连民族大学 计算机科学与工程学院，辽宁 大连 116000)

摘 要：随着自动驾驶领域对目标检测的精度和速度需求的提高，目标检测从传统检测算法转向深度学习方向发展。由于 2D 目标检测算法存在小目标丢失等问题，基于深度学习的 3D 目标检测算法以能提供物体的位置，尺寸，方向，等一些空间结构信息的优势，迅速在自动驾驶领域发展起来。该文章简单陈述了二维目标检测算法，将 3D 目标检测算法分成五个类别，分析了各类目标检测算法的优缺点，并详述了最新被提出的基于 GNN（图神经网络）的两种算法。最后对 3D 目标检测所应用的领域和其研究意义进行总结，并对 3D 目标检测今后可能发展的方向作出猜想。

关键词：自动驾驶；深度学习；3D 目标检测；图神经网络
中图分类号：TP391　　**文献标识码：**A

Overview of 3D Object Detection Algorithms Based on Deep Learning

ZHANG Xin-yu[1], *XU Zi-xian*[2], *YAN Dong-mei*[1], *SHA Xiao-peng*[1], *GU De-ying*[1]
(1. School of Control Engineering, Northeastern University at Qinhuangdao, Qinhuangdao 066004, China; 2. School of Computer Science and Engineering, Dalian Minzu University, Dalian 116000, China)

Abstract: With the increase in the accuracy and speed requirements of object detection in the field of automatic driving, target detection has shifted from traditional detection algorithms to deep learning. Due to the 2D object detection algorithm, there are problems such as the loss of small targets, the 3D object detection algorithm based on deep learning has the advantages of providing some spatial structure information such as the position, size, direction, and external shape of the object, and it has developed in the field of autonomous driving rapidly. This article briefly stated the two-dimensional object detection algorithm, divided the 3D object detection algorithm into five categories, analyzed the advantages and disadvantages of various object detection algorithms, and detailed the two newly proposed GNN (graph neural network) based algorithm. Finally, the application field of 3D object detection and its research significance are summarized, and the future direction of 3D object detection may be guessed.

Keywords: Autonomous driving; Deep learning; 3D object detection; Graph neural network

1 引　言

目标检测是计算机视觉领域中必不可少的处理任务，其中，目标检测处于极其重要的位置。基于深度学习的目标检测算法一直是无人驾驶领域研究的热点和难点之一，目标检测算法首先要判断系统摄取的图像中物体是否存在，进行物体的标注，实现物体定位和分类。无人驾驶系统中，起决定性作用的是对外部环境的感知能力，比如路标路牌的识别，障碍物的分类，对行人及车辆的位置精确定位等。

深度学习未兴起时，传统检测算法中的特征提取器和分类器是手动设计的，所有的特征都是人为设定的[1]。在检测过程中，由于目标大小和位置的不确定性，需要大量的候选框遍历整张图片实现定位，这不仅会产生大量多余的候选框极度浪费计算机资源，检测精度也不如人意。

随着 CNN（Convolutional Neural Net-works，卷

收稿日期：2021-03-22；修回日期：2021-12-28
基金项目：河北省高等学校科学研究重点项目（ZD2019305）
作者简介：张新宇（1997-），女，安徽阜阳人，硕士研究生，主要研究方向为深度学习、3D 目标检测等（本文通信作者，Email: 2533166792@qq.com）；闫冬梅（1970-），女，黑龙江齐齐哈尔人，博士，副教授，主要从事智能控制、传感器应用等方面的教学与科研工作。

积神经网络）的发展，研究者凭借其善于处理大规模图像数据的优点逐步用卷积神经网络来替代手工特征提取器进行特征提取，目标检测算法也逐步向基于深度学习的方向发展，检测精度、速度不断提升，基于深度学习的目标检测算法也可以在一定程度上改善由候选区计算冗余的问题。

2D 目标检测不能获取图像深度、尺寸等外部的空间信息，并且当物体距离摄像机较远时，拍摄的物体可能会过小，在复杂的道路环境中，物体之间会存在很多不同程度的遮挡情况，或者是因光线不强、拍摄角度不佳等外在因素造成图像像素不高等问题，这些都会造成图像中目标丢失导致检测精度低。3D 目标检测能提升自动驾驶系统对环境的感知能力和并能得到物体的更为详细的空间三维信息，意在能够获取空间中目标的位置，物体尺寸以及物体姿态。

本文对目标检测算法的发展过程进行了归类总结，根据传感器和图像数据形式的差异，现阶段 3D 目标检测算法大致分为基于单目图像、立体（双目）视觉、点云数据、特征融合网络和图神经网络（GNN）的目标检测算法五大类。表 2 是对部分 3D 目标检测常用的公开数据集的简单介绍。并展望了对未来三维目标检测算法的发展方向。

2 2D 目标检测

2D 目标检测首先对图像中的目标进行检测提取出候选区域，再根据后候选区域进行目标类别分类。过去 2D 目标检测算法主要是朝着两个方面发展，即传统检测算法和基于深度学习的目标检测算法。

2.1 传统的目标检测算法

最早出现的目标检测是 2001 年 Paul Viola 提出的 VJ（Viola-Jones）[2]检测器用于人脸检测，使用多尺度 Harr[3]特征描述人脸特征，利用 AdaBoost 构建的强分类器来进行人脸检测，可以同时提高检测性能和检测效率。较为经典的手工特征提取器还有 SIFT（Scale-invariant Feature Transform，尺度不变特征变换）[4]、HOG（Histogram of Oriented Gradient，方向梯度直方图）[5]、SVM（Support Victor Machine，支持向量机）[6] 等。在传统的目标检测算法中，HOG 常与 SVM 分类器结合使用，例如，Felzenszwalb 2008 年提出的 DPM（Deformable Part Model）[7]。传统的目标检测方法具有非常高的可解释性，但是实现过程中会出现冗余度高，检测时间过长，鲁棒性低等问题。随着各项技术的发展，传统的目标检测渐渐淡出视线，而基于深度学习目标检测算法有了突破性进展。

2.2 基于深度学习的目标检测算法

神经网络之父—— Geoffrey Hinton 等人基于 CNN 提出 AlexNet[8]，并在 2012 年的 ImageNet 图像识别比赛上获得了冠军，证明了深度学习在目标检测领域中的潜力，正式将深度学习引入目标检测领域中。OverFeat[9]是首个应用在基于深度学习目标检测领域的方法，在 2013 年由年纽约大学的 Yann LeCun 团队提出，随着卷积网络的发展，研究者提出很多基于深度学习的目标检测方法，基本分成两类：基于候选区的 Two-stage（两阶段）目标检测算法，基于回归的 One-stage（单阶段）目标检测方法。Two-stage 检测算法首先产生候选区域，再对候选去进行分类。One-stage 是一种端到端的目标检测算法，该算法只需一步就能预测物体类别和位置信息。

2.2.1 Two-stage 目标检测算法

2014 年，Girshick 提出的 R-CNN[10]，该算法分为三步：（1）把选择性搜索算法作为提取图片候选区的主要方法；（2）获取所划分的候选区输入到训练好的卷进神经网络中完成特征提取；（3）用给定的 SVM 和全连接网络对每个候选区的目标进行检测，得到相应目标类别和位置。R-CNN 的提出为目标检测提供了全新的方法，成为了 Two Stage 检测算法的基础，但是大量的候选区域带来了数据冗余的问题，并且增加了大量的计算成本。Two-stage 目标检测算法框图如图 1 所示。

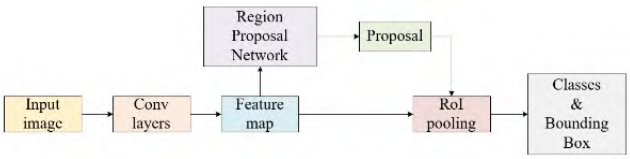

图 1　Two-stage 检测算法架构图

Fig. 1 Two-stage detection algorithm structure diagram

2015 年，何凯明提出 SPP-Net[11]来解决 R-CNN 中的重复卷积的问题，提升了检测速度，但仍然存在训练步骤过多、模型较大等问题。Girshick 又提出了基于 VGG-16[12]的 Fast R- CNN[13]，整合了 R-CNN 和 SPP-Net 各自的优点。Fast-RCNN 将整个图片作为输入，用 RoIPooling 进行特征尺寸变换，并用 Multi-task loss 进行边框回归，减少了训练步骤，Fast- RCNN 的训练速度大约为 SPP-Net 的 3 倍，测试速度较 R-CNN 快 210 倍。但 Fast-RCNN 依然存在不能同时获取候选区域的等问题，所以检测速度上还有提高的空间。2016 年，Girshick 等提出 Faster-

RCNN[14]，将 RPN（Region Proposal Network）替代 Selecive Search 提取候选区域，创造性地将特征提取、边框回归、目标分类聚集到一个网络中，目标检测效率有了很大的提升。

2.2.2 One-stage 目标检测算法

基于 Two-stage 的目标检测算法需要先生成目标候选框后再进行目标分类，会使检测效率降低，为了提升检测速度，研究者提出 One-stage 目标检测算法，目标边框预测和目标分类同时完成，大大提升了检测的速度，One-stage 检测算法结构框图如图 2 所示。常用的目标检测算法框架有 YOLO(You Only Look Once)系列和 SSD 系列等。

2016 年，Redmon 等提出的 YOLO-v1[15]，其只需一个网络便可预估出边界框的位置和物体的类别。YOLO-v1[15]的网络结构图如图 2 所示。该方法速度较 Two-stage 算法有大幅提升，但是划分目标框时只利用固定的 7×7 大小的区域，在实际应用时会出现物体定位错误，在检测密集物体和小物体时出现漏检等问题。

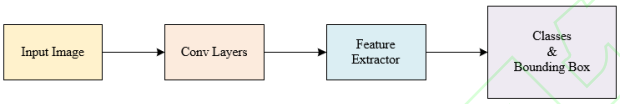

图 2　One-stage 检测算法结构图

Fig. 2 One-stage detection algorithm structure diagram

YOLO-v2[16] 是 Redmon 等基于 YOLO -v1 为基础修改而来的，通过对卷积层后加入 BN（Batch Normalized，批量归一化），使模型不再产生过拟合现象，进一步提高了模型收敛速度，并借鉴 Faster R-CNN 中的 Anchor 来完成检测，由于需要的 Anchor 数量多于 YOLO -v1 中的候选框，导致 mAP 有所下降，但召回率提升了 7%，YOLO-v2 拥有了对小物体检测的能力。为了检测精度和速度能进一步提高，Redmon 等人又提出 YOLO-v3[17]，YOLO-v3 利用了 FPN（Feature Pyramid Network,特征金字塔网络），同时输出 3 种尺度不同的特征图，在保证检测速度较快的前提下，增加了对小目标检测的能力，YOLO-v3 不再使用 Softmax 对每个 Anchor 中的物体进行分类，而是利用了逻辑回归的思想直接输出可能性最高的物体标签。2020 年，Alexey Bochkovskiy 接力 Redmon 提出 YOLO-v4[18]，基于 YOLO-v3 调整，把拼接四张图片作为新的数据增强方法，结合 CmBN(Cross mini-Batch Nor-malization，跨小批量标准化)，以及加入空间注意力机制 SAM(Spatial Attention Module)大大提升了检测性能，将 YOLO-v3 的 AP 和 FPS 分别提高 10%和 12%，但是由于模型尺寸太大导致检测速度降低，YOLO-v5 着重提升了检测速度，主要改变了训练时的损失函数以及预测框筛选机制，检测速度远超 YOLO-v4。由于 YOLO 系列的检测速度快，特征图的表征能力不足，会导致对远处的小目标、存在严重遮挡的目标、或因光亮不足、拍摄角度等不清晰的目标检测精度不佳。

Liu W 等人提出了 SSD[19] 检测算法主要平衡了 R-CNN 系列算法和 YOLO 系列算法的不足，SSD 使用 FPN (Feature Pyramid Networks，多尺度特征图)，对不同大小的卷积层采用不同大小和数量的预设框，分别输出预测的真实边框和类别置信度，最后采用 NMS(Non Maximum Suppression，非极大值抑制法)来输出结果。稍微提高了对图中小目标的检测能力，为了继续优化检测质量，逐步出现了 DSSD[20]，RSSD[21]，ASSD[22]等一系列方法。One-stage 目标检测算法性能对比如表 1 所示。

表 1 One-stage 目标检测算法性能比较

Tab. 1 Performance comparison of one-stage object detection algorithms

Method	Backbone	Dataset	mAP/(%)	FPS (s^{-1})
YOLO-v1	VGG-16	VOC 2007	66.4	45.0
YOLO-v2	Darknet-19	VOC 2007	78.6	40.0
YOLO-v3	Darknet-53	MS COCO	33.0	51.0
YOLO-v4	CSPDarknet53	MS COCO	43.5	23.0
SSD	VGG-16	VOC 2007	77.1	46.0
DSSD	ResNet-101	VOC 2007/2012	78.6/76.3	9.5
RSSD	VGG-16	VOC 2007	80.8	16.6
ASSD	ResNet-101	VOC 2007	83.0	2.7

3 3D 目标检测算法

目前，2D 目标检测只能完成目标平面上的定位，但是自动驾驶车辆所处的环境是立体的， 不仅需要获取目标在 3D 空间中的位置信息，还需要获取尺寸、方向等深度信息，因此基于 2D 图像的目标检测算法已经不能满足无人驾驶的需求。研究者将视线转向空间目标检测，3D 目标检测能得到更为详细的目标物体空间三维信息，提升自动驾驶系统的环境感知能力，三维检测数据集 KITTI[23]、Waymo[24]、nuScenes[25]等也不断发展，也进一步激发了研究者的科研激情。部分 3D 目标检测常用的公开数据集的简单介绍如表 2 所示。

表 2 部分三维目标检测公开数据集对比

Tab. 2 Comparison of some 3d target detection public datasets

Dataset name	Scenes	Classes	Frames	Size(G)	Feature
KITTI	22	8	15k	43.67(train)	7481 张训练/验证集,7518 张测试集;8 类目标 3D 标注
Waymo	1k	4	200k	336.62(train+val)	1000 段驾驶路径, 1150 个场景,每个场景时长 20s
nuScenes	1k	23	40k	75.75(train+val)	1000 个不同场景视频,包含 1400k 张图片;23 类 3D 标注;包含白天、黑夜等不同环境状况
CityScapes	—	8	—	12.7(train+val)	包含 50 个城市的街道场景中的立体视频,双目相机拍摄，共 5000 张精细释，2975 张训练图，500 张验证图和 1525 张测试图
ApolloScape	—	35	140k	206.16(train)	提供了 17062/1973 张训练/测试集图像和相对应的语义标注与深度信息

3.1 基于单目图像的 3D 目标检测算法

单目图像以获取方式简单，成本低等优势受到了研究者青睐，基于单目图像的 3D 目标检测算法以成熟的 2D 目标检测算法为基础，实现了 2D 图像中目标的定位和分类。

Chen 等人在 2016 年提出了 Mono3D[26] 目标检测方法，该方法利用 Faster R-CNN 提取特征，结合上下文信息、位置先验信息和目标形状先验信息等，计算出检测框的总损失函数，来提取精确的目标三维检测框。但是在计算损失函数中存在误差累计的问题，导致 Mono3D 的精度不是非常优越，Mousavian 等人提出了 Deep3Dbbox[27]，以滑动窗口的思想为灵感提出了 Multi-bins （混合离散-连续）回归方法，并直接使用 L2 损失函数估计误差，使用最小二乘法匹配 3D 检测框与 2D 检测框的位置关系，该方法对估计简单目标的位置估计非常适用，但不利于小目标、遮挡目标的检测，相对于 Mono3D 网络架构得到简化，因此提升了检测速度，但检测精度并没有很大的提升。

2019 年，Wang 等人借助成熟的 2D 检测算法提出 GS3D[28]，设计了 2D+O 子网和 3D 子网，它们分别用于生成目标的 2D 框、物体的观察方向和粗略的提取 3D 的框表面特征，并结合经投影产生的 2D 边框输出精准的 3D 边框，这也解决了 Deep3Dbbox 中存在的小目标特征不明显的问题。但是通过多个网络进行语义分割，特征提出，边框回归无疑会减慢检测速度，而 2019 年 Brazil 等人提出的 M3D-RPN[29]使用单一整体的网络进行 3D 目标检测，并设计了深度感知（depth-aware）卷积层来增强对三维场景的理解，学习空间特征，对提升目标检测的精度有很大作用。

2020 年，Li 等人受 CenterNet[30]启发提出 RTM3D[31]，把目标检测看作关键点检测，在空间估计出目标的尺寸、位置和大致方向。考虑到目标过小或太密集时，关键点会出现重叠问题，作者还提出 KFPN（Keypoint Feature Pyramid Network，关键点特征金字塔网络）检测空间中多尺度的关键点应对重叠问题。RTM3D 也是首个仅使用单目图像进行实时检测的算法。Ding 等人提出的 D^{4}LCN[32]基于深度引导卷积的检测算法精度和 RTM3D 相当，并且处理速度略快。

3.2 基于立体视觉的三维目标检测算法

单目图像中的物体尺寸都很小，许多目标被严重的遮挡，因此单目图像能呈现出来的信息是非常有限的。在自动驾驶过程中，立体摄像头能提供比单目摄像头更大的感受野和精准的深度信息，但是获取更多的深度信息会带来更大的计算量，所以如何将左右两个单目相机联合起来成为了双目立体目标检测的热点。

2018 年，Chen 等人提出 3DOP[33]，是一种基于立体视觉的 3D 对象检测方法，结合单目图像和点云数据生成高质量的 3D 边界框，并超越了 2D 边界框的检测精度。利用立体摄像机估计的三维信息，并对 3D 点云数据的特征进行评分，特别的是，作者将评分函数编码了几个有深度信息的特征，如候选框内的点密度、自由空间、可见性以及物体尺寸先验和地面以上的高度，来提高检测精度。

2019 年 Li 等人基于 R-CNN 提出 Stereo R-CNN[34]，高效率利用了立体图像中的稀疏和密集数据、语义信息和几何信息。网络架构如图 3 所示，它直接将双目图像作为网络的输入，并借助 Mask-RCNN 中的区域特征聚集方式(RoI Align)，连接左 RoI 特征和右 RoI 特征进行对象类别的分类，回归出精确的 2D 边框、视点和维度。采用类似于 Mask-RCNN 的结构进行关键点预测，仅使用左边的分支左 RoI 特征来预测目标对象的关键点，为估计 3D

框估计提供更严格的约束。最后连入全连接层和 ReLU 层，用于提取图像语义信息，回归出空间边框、目标类别、尺寸和视点角度。以前的工作侧重于优化视差的估计，但是随着距离的增加，深度估计的误差会增大，这也是影响检测精度的关键因素。2020 年 Peng W 等人提出 IDA-3D[35]，它是一种端到端的学习框架，使用单目生成双目图像作为算法的输入，引入 IDA (Instance Disparity Adaptation，实例深度感知)模块预测三维边框中心的深度，从而提高了检测精度，并且不依赖于训练深度图像，不需要多阶段处理，提升了检测速度。

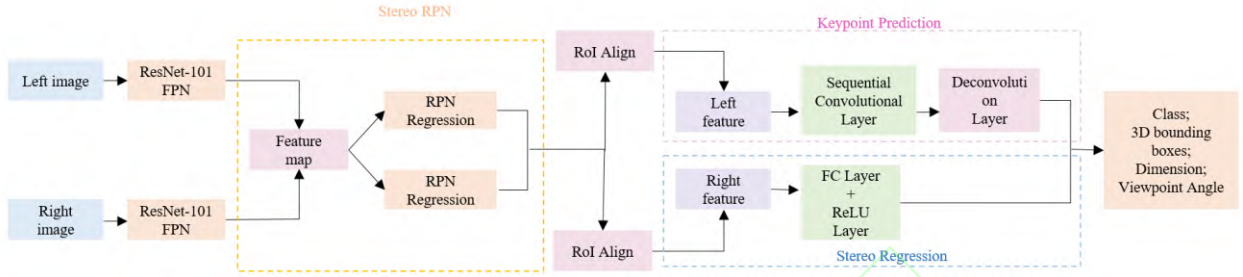

图 3　Stereo RPN 结构图

Fig. 3 Stereo RPN structure diagram

相对于单目视觉而言，立体视觉能提供更多的深度信息，立体视觉最大的难题是立体匹配和视差计算，和单目视觉一样，车身的抖动和车轮胎的变化会影响摄像头外参的变化，所以精准的相机标定也是很难实现的。在自动驾驶领域中，往往还更需要极为详细的空间信息，目前的立体视觉检测算法的精度还不能够达到实时检测的效果。

3.3 基于纯点云数据的三维目标检测算法

与基于图像的检测相比，点云数据提供的更加真实有效的空间结构信息，可以准确定位物体并大致推测出它们的形状。在早期的一些工作中，基于点云数据的三维目标检测方法的流程一般是点云数据预处理（例如，降采样等）、点云分割、点云聚类。例如，Klaas Klasing[36]等使用近邻聚类算法对点云数据聚类从而生成目标候选区域，Jeremie Papon[37]等对点云进行无监督分超割形成多个炒体素，再逐个对超体素进行分类，但可能会存在过度分割或分割不足产生分割错误，造成分类结果不正确，为了改善这一缺点，研究者将深度学习引入到基于点云数据 3D 目标检测中，基于纯点云数据的 3D 检测算法主要分基于体素（voxel-based）和基于点（ point-based）两大类。

2018 年 Zhou 等人提出的 VoxelNet[38]算法是是第一个基于体素（voxel-based）方法的框架，消除手工提取 3D 点云的特征，在同一个阶段上进行特征提取和目标边围框的检测，并设计了一种新的体素编码层(VFE，Voxel feature encoding layer)，VFE 的结构图如图 4 所示，通过对点云进行编码的方式使无序的点云数据具有可描述性，能够有效区分各种几何形状。VoxelNet 的结构主要分为三部分：特征学习网络、卷积中间层和 RPN。特征学习网络把点云数据分成为间距相同的体素网格，并将每个网格内的点转换成可以表示形状信息并且结构相同的向量。卷积中间层将表示体素矢量集合到逐渐扩大的感受野内，来增加更多的物体的外部信息，最后借助 RPN 网络生成 3D 边界检测框和输出物体类别。VoxelNet 的成功也衍生出了一些新的算法，例如 2019 年海康威视出品的 Voxel-FPN[39]，其特征提取的方式和 VoxelNet 中的相同， Voxel-FPN 的精妙之处在于简化了网络结构，但是检测效果依然非常精确。

将点云转换成体素网格会极大的增加需要处理的数据量，使处理速度减慢。基于点（point-based）的方法可以改善这个问题，2017 年，Qi 等提出 PointNet[40]网络，PointNet 对点云数据进行直接处理，并对每个点进行单独处理获取其空间编码，自主学习点云数据中离散状态的关键点，因此 PointNet 也能提高缺失的数据或有波动的数据的稳定性。由于 PointNet 不能获取局部特征，对分析复杂度高的场景能力不足，Qi 等对 PointNet 改进后提出了 PointNet++[41]，运用了 FPS (Farthest point sampling)最远点采样法来对点集进行划分，产生结构相同的区域，引入分层神经网络，以分层的方式处理在空间中采样的一组点，首先从小范围内提取级别低的特征，逐步扩大提取范围，检测过程中感受野会不断增大，实现提取更高级别的图像特征，最终可以获得整个点集的特征，但 PointNet++中处理的是稀疏的点云数据，点云的无序性使其很难根据一个点确定物体位置，2018 年 Qi 等人又提出的 F-PointNet[42]结合图像信息和深度信息并使用 PointNet、PointNet++的结构进行目标检测，首先对 2D 图像进行 2D 检测来完成图中物体目标定位，再借助 2D 检测预估的定位结果，用

其相应的点云数据实现 3D Box 和类别的回归。这种方法实现了目标的高效定位，并对很小的目标也有很高的召回率，即使在强遮挡和点云十分稀疏的情况下，也能够准确地估计出三维边框，在分类、物体检测和语义分割任务中也有较好的表现能力，而且还具有较好的实时性。

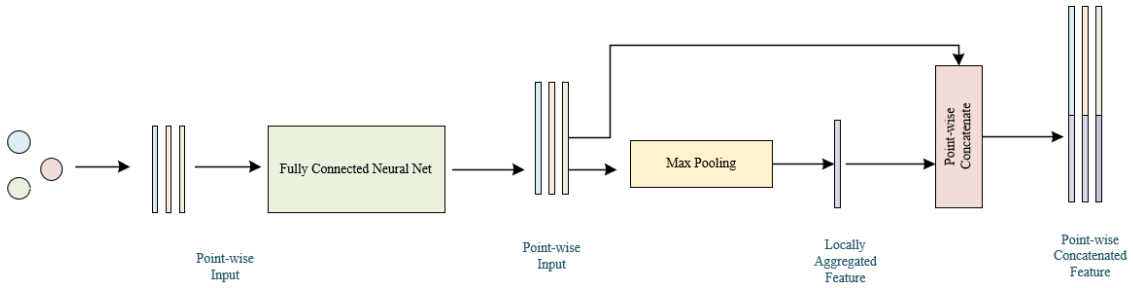

图 4　VFE 结构图

Fig. 4 VFE structure diagram

基于点的方法在于如何应对场景点云数据的无序性问题和如何提高效率，要提高基于点（point-based）方法的效率，就必须改进或者除去费时间的上采样过程。2019 年，Qi 等人提出的 VoteNet[43]，借助霍夫投票的思想，通过直接处理点云数据，避免了量化过程造成的信息损失。VoteNet 是一种基于 point-based 的检测算法，将 PointNet++作为主干网络。投票思想是为稀疏的集合所设计，很适合用于点云数据，VoteNet 只使用图像的几何信息，并不依赖彩色的图像，投票模块是用 MLP (Multi-Layer Perceptron，多层感知器)实现的，使用 ReLU 激活函数和归一化处理，使得投票生成预测的 3D 边界框和对象的语义类别，但缺少上采样过程就会使检测精度降低，2020 年，Yang 等人提出的轻量级的基于点的三维单阶段物体检测框架（3DSSD[44]），实现了检测精度和检测速度之间的均衡。作者大胆的舍弃了基于点的方法中不可缺少的所有上采样层和细化模块，来减少计算成本，为了防止没有上采样带来的性能下降问题，提出基于特征距离的融合采样策略（FS），并使用 MLP 提取特征。作者还设计了一种 3D 中心分配策略，该策略将较高的分类分数分配给更接近实例中心的候选点，以实现更准确的定位检测。在广泛使用的 KITTI 数据集上进行评估，3DSSD 很大程度上优于所有现有的基于体素的方法，并具有与基于点的两阶段方法相当的性能。

3.4 基于融合网络的三维目标检测算法

点云数据具备更加准确的深度信息，而单目图像拥有更加详细的语义信息。结合两者的优点研究者们推出了融合网络检测算法，主要分为两类：多模态信息融合网络、单模态多特征融合网络。融合算法的简要结构框图如图 5、6 所示。

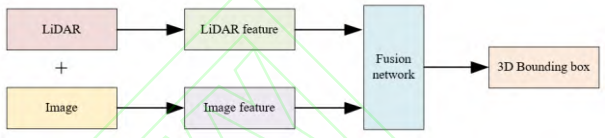

图 5　多模态信息融合网络简图

Fig. 5 Simplified diagram of multimodal information fusion network

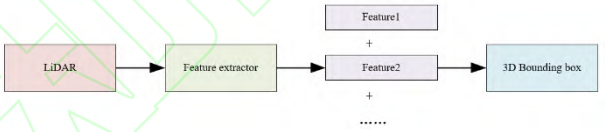

图 6　单模态多特征融合网络简图

Fig. 6 Simplified diagram of single-mode multi-feature fusion network

2017 年 Chen 等提出的 MV3D[45]在多模态信息融合网络中是一项开创性的工作，将不同形式的数据有效融合后再进行特征提取。MV3D 的网络输入是由单目图像数据、点云数据的鸟瞰视图和正前前视图组成的，建议网络用于生成 3D 目标候选区域，将三输入分别处理后得到 3 个不同视图的感兴趣区域，通过融合网络得到较为精确的检测结果。2018 年 Ku 等提出 AVOD[46]，主要是基于 MV3D 网络进行改进，采用全分辨率特征提取器来提取点云数据和彩色图像的特征图，特征提取器由编码器和解码器组成，编码器是对 VGG-16 进行了修改，将其信道数量减半，来减少小物体占据的输出特征图的像素数量，增大感受野；解码器受 FPN 的启示创建了一个自下而上的解码器，使其在保持运行速度的同时，并将特征图上采样还原到原始输入的大小。

基于多模态信息融合的方法还有 2018 年 Xu 等人提出的 PointFusion[47]，把 ResNet[48]和 PointNet 作为主干网络，分别处理图像数据和原始点云数。PointFusion 的结构如图 7 所示，使用 ResNet 从输入 RGB 图像中提取目标的外部形状特征和几何特

征，借助了 PointNet 的方法提取点云数据并对每个点的空间编码进行学习，同时提取聚集的全局点云特征，最后把学习到的特征用于目标分类和语义分割。通过密集融合模型（Dense Fusion）将输入的 3D 点用作密集的空间锚点，预测每一个 3D 点从该点到邻近边框角的位置的空间偏差距离，把图像数据特征和点云特征结合成一个整体，作为全局融合模型（Global Fusion）的输入，直接回归出目标的 3D 边框 8 个顶点的位置，最后得到 3D 目标边界框。

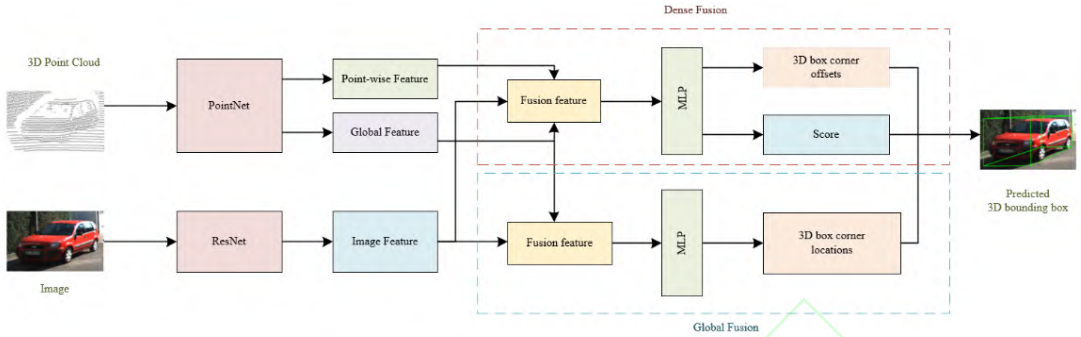

图 7　PointFusion 结构简图

Fig. 7 Schematic diagram of PointFusion structure

基于融合网络最主要的任务就是如何有效的融合单目图像数据和点云数据，单模态多特征融合与多模态信息融合不同的是它只将点云数据作为输入，对其采取不同形式的特征提取的方式，分别得到特征后再经过特征融合模块得到 3D 边界框。2020 年 Shi 等人提出的 PV-RCNN[49]，同样采用的是单模态多特征融合的方式，只将原始的点云数据作为输入，将 3D voxel 卷积神经网络(CNN)和基于 point_based 的网络进行深度的融合，因为经过 voxel_based 网络的处理能够高效的对多尺度特征进行编码，作者采用 voxel-to-keypoint（体素-关键点）编码器，将整个场景的多尺度体素特征编码成一个关键点集合，使用多个感受野从场景中收集更丰富的上下文信息，来进行准确预测 3D 边界框和物体类别。在 2020 年 KITTI 数据集的 3D 检测基准上 PV-RCNN 排名第一。

部分 3D 目标检测算法在 KITTI 数据集上的车辆检测的性能对比如表 3 所示。

表 3 KITTI 数据集上的车辆检测的指标对比

Table 3 Comparison of vehicle detection indicators on KITTI dataset

Method	Type	AP_{2D}(%)			AP_{3D}(IOU=0.5)/%			Runtime/s
		Easy	Moderate	Hard	Easy	Moderate	Hard	
Mono3D	Mono	92.33	88.66	78.96	23.41	15.26	12.80	0.03
Deep3Dbbox	Mono	92.98	89.04	77.17	27.04	20.55	15.88	1.50
GS3D	Mono	86.23	76.35	62.67	4.47	2.90	2.47	2.00
RTM3D	Mono	91.82	86.93	77.41	16.73	11.45	9.92	0.03
M3D-RPN	Mono	89.04	85.08	69.26	49.43	36.18	28.90	0.16
3DOP	Sterero	92.96	89.55	79.38	46.04	34.63	30.09	4.20
Stereo R-CNN	Sterero	93.98	85.98	71.25	85.84	66.28	57.24	0.30
VoxelNet	LiDAR	89.60	84.81	78.57	81.97	65.46	62.85	—
PointNet	LiDAR	95.85	95.17	85.42	82.19	69.79	60.59	0.17
3DSSD	LiDAR	97.69	95.10	92.18	88.36	79.57	74.55	0.04
MV3D	LiDAR＋Mono	96.47	90.83	78.63	74.97	63.63	54.00	0.36
AVOD	LiDAR＋Mono	94.70	88.92	84.13	76.39	66.47	60.23	0.08
PointFusion	LiDAR＋Mono	95.85	95.17	85.42	83.76	70.92	63.65	0.17
Point-GNN	LiDAR	96.58	93.50	88.35	88.33	79.47	72.29	0.60

3.5 基于图神经网络（GNN）的目标检测算法

近几年，随着 GNN[50]（Graph Neural Network，图神经网络）的发展，研究者们尝试将 GNN 引入到 3D 目标检测中去。Shi 等人提出的 Point-GNN[51]，首次将图神经网络应用于点云数据做三维目标检测，预测图像中每个对象的类别和形状。在 Point-

GNN 中，缩小局部的平移方差，作者提出自动配准机制（ auto-registration），利用边界框融合来实现多个顶点检测结果的精准结合。网络整体结构包含三个部分：图形构建、 Point-GNN 网络结构、边界框的融合和置信度。为了减小 Point-GNN 算法的计算负担，作者采用了 voxel 下采样，并将 voxel 编码的信息作为 GNN 的初始特征。因为点云的相对坐标整体具有平移不变性，但对邻近区域的变换较灵敏，所以采用顶点间的相对坐标作为输入，边界框和每个点的所属对象作为网络的输出信息，并且采用 MLP 学习边的特征和顶点的属性。在以 KITTI 数据集为标准的实验中表明，3D 车辆对象检测的平均精度(AP)高达 88.33%，鸟瞰车辆对象检测的平均精度(AP)高达 93.11%。

Najibi 等人也提出了一种基于图卷积网络的 3D 目标检测方法，被称为 DOPS[52]。不仅可以作为室外检测算法，也可以用于室内检测，相比于 Point-GNN 较突出的一点是它能估计物体的形状。DOPS 整体架构主要分为四部分，分别是：逐点特征提取（Per Point 3D Object Prediction）、对象建议合并（Object Proposal Consolidation）、生成 3D 检测框（ Proposing Boxes ）、 物 体 形 状 预 测 （ Shape Prediction）。将 N×I 的点云数据作为输入，使用 SparseConvNet[53]作为主干网络进行特征提取，使用 3D 稀疏卷积得到物体的多个属性，包括中心点坐标、尺寸大小、外部形状等。作者将图卷积应用于对象建议合并部分，图的定义方式是图中的每个点都是点云的点的特征，用图卷积聚合点云数据特征得到 3D 边界框。图 6 为 Shape Prediction 的结构。首先将点云数据进行体素化，通过一个 3D Sparse Encoder 编码成特征，由平均池化层（Avg-pooling）得到了采样空间点的 attention，每个点预测得到该点到物体表面的有效距离，最后使用移动立方体算法（Marching Cubes Algorithm）实现物体的三维重建估计出物体的形状。在 ScanNet 场景中的物体检测中，DOPS 达到了约 5%的结果，在 Waymo 开放的数据集上获得了 3.4%的最高结果，并同时检测到了汽车的形状。

Point-GNN 检测算法只应用点云数据就达到了很高的精度，甚至可以超过基于融合的算法，DOPS 不仅得到了基于点云数据较高的3D目标检测结果，还成功地对目标进行了形状的预测。Point-GNN 和 DOPS 证明了在未来无人驾驶研究领域中，图神经网络(GNN)也可以作为 3D 目标检测的新方法。

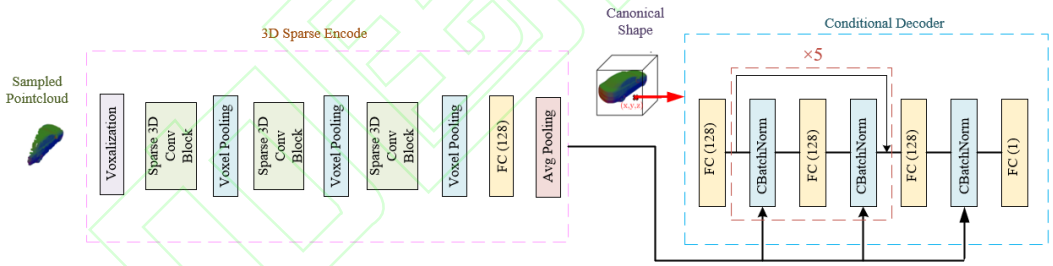

图 8 Shape Prediction 结构图

Fig. 8 Structure diagram of shape prediction

4 结　论

基于深度学习的目标检测已经被广泛应用于无人驾驶等其他计算机视觉领域中，把深度学习技术和机器视觉相结合，能为自动驾驶中的环境感知问题带来强大的解决方案。目标检测是自动驾驶系统中环境感知模块的核心，2D 检测算法已经非常成熟，但 2D 目标检测算法在精度上无法满足无人驾驶及需要目标检测技术的其他领域的需求，因此 3D 目标检测算法将继续是自动驾驶等领域的研究热点和难点。

目前，3D 目标检测算法主要是朝着两个方向发展：（1）基于激光雷达点云数据的目标检测，（2）基于融合网络的目标检测算法。由于脱颖而出的基于图神经网络（GNN）的目标检测算法取得了很好的检测结果，可以推测基于 GNN 的目标检测算法也是未来检测领域一个可观的发展分支。未来目标检测必然会被应用于更多的领域，如医学成像、无人机、军事防御等，在未来的自动驾驶领域中，必定会出现精度更高，实时性更好，综合性更强的目标检测算法。

参考文献(References)

[1] 张鹏, 宋一凡, 宗立波, 等. 3D 目标检测进展综述[J]. 计算机科学, 2020, 47(4): 94-102.
Zhang P, Song Y F, Zong L B, et al. Review of 3D Object Detection [J]. Computer Science 2020, 47(4): 94-102.

[2] Viola P, Jones M J. Robust Real-time Face Detection[J]. International journal of computer vision, 2004, 57(2): 137-154.

[3] Papageorgiou C P, Oren M, Poggio T. A General Framework for Object Detection[C]. Bombay: Sixth International Conference on Computer Vision, 1998.

[4] Lowe D, Distinctive G. Image Features from Scale-invariant Keypoints[J]. International Journal of Computer Vision, 2004, 60(2): 91-110.

[5] Dalal N, Triggs B. Histograms of Oriented Gradients for Human Detection[C]. San Diego: Proceeding of the IEEE Conference on Comuter Vision and Pattern Recognition. 2005.

[6] Platt J C. Fast Training of Support Vector Machines Using Sequential Minimal Optimization, Advances In Kernel Methods[J]. Support Vector Learning, 1999, 10(3): 61-74.

[7] Felzenszwalb, Pedro F, Girshick, et al. Object Detection with Discriminatively Trained Part-based Models[C]. Alaska: the IEEE Conference on Computer Vision and Pattern Recongnition, 2008.

[8] Krizhevsky A, Sutskever I, Hinton G. ImageNet Classification with Deep Convolutional Neural Networks[C]. Lake Tahoe: Advances in Neural Information Processing Systems, 2012.

[9] Sermanet P, Eigen D, Zhang X, et al. OverFeat: Integrated Recognition, Localization and Detection Using Convolutional Networks[J]. arXiv preprint arXiv: 1312.6229, 2013.

[10] Girshick R, Donahue J, Darrell T, et al. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation[C]. Columbus: 2014 IEEE Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014.

[11] He K, Zhang X, Ren S, et al. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(9):1904-1916.

[12] Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-scale Image Recognition[J]. Computer Science, arXiv preprint arXiv:1409.1556, 2014.

[13] R.Girshick. Fast R-CNN[C]. Boston: the IEEE Conference on Computer Vision.

[14] Ren S, He K, Girshick R, et al. Faster R-CNN: Towards Real-time Object Detection with Region Proposal Networks[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017: 39(6): 1137-1149.

[15] Redmon J, Divvala S, Girshick R, et al. You Only Look Once: Unified, Real-Time Object Detection[C]. Las Vegas: the Conference on Computer Vision and Pattern Recongnition, 2016.

[16] J. Redmon, A. Farhadi.YOLO9000: Better,Faster, Stronger[C]. Hawaii: 2017 IEEE Conference on Computer Vision and Pattern Recognition, 2017.

[17] Redmon J, Farhadi A. Yolov3: An incremental improvement[J]. arXiv preprint arXiv: 1804.02767, 2018.

[18] Bochkovskiy A, Wang C Y, Liao H Y M. YOLOv4: Optimal Speed and Accuracy of Object Detection[J]. arXiv: 2004.10934, 2020.

[19] Liu W, Anguelov D, Erhan D, et al. SSD: Single shot multibox detector[C]. Springer: European Conference on Computer Vision, 2016, 9905 LNCS: 21-37.

[20] Fu C Y, Liu W, Ranga A, et al. Dssd: Deconvolutional single shot detector[J]. arXiv preprint arXiv: 1701.06659, 2017.

[21] Jeong J, Park H, Kwak N. Enhancement of SSD by concatenating feature maps for object detection[J]. arXiv preprint arXiv: 1705.09587, 2017.

[22] Yi J, Wu P, Metaxas D N. ASSD: Attentive single shot multibox detector[J]. Computer Vision and Image Understanding, 2019, 189: 102827.

[23] Geiger A, Lenz P, Urtasun R. Are we ready for autonomous driving? The KITTI vision benchmark suite[C]. Providence, RI, USA: 2012 IEEE Conference on Computer Vision and Pattern Recognition, 2012: 3354-3361.

[24] Sun P, Kretzschmar H, Dotiwalla X, et al. Scalability in Perception for Autonomous Driving: Waymo Open Dataset[J]. arXiv: 1912.04838, 2019.

[25] Caesar H, Bankiti V, Lang A H, et al. NuScenes: A mul-timodal dataset for autonomous driving[J]. arXiv preprint arXiv: 1903.11027, 2019.

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

[27] Mousavian A, Anguelov D, Flynn J, et al. 3D Bounding Box Estimation Using Deep Learning and Geometry[C]. Hawaii, USA: 2017 IEEE Conference on Computer Vision and Pattern Recognition, 2017: 7074-7082.

[28] Li B, Ouyang W, Sheng L, et al. GS3D: An Efficient 3D Object Detection Framework for Autonomous Driving[C]. Long Beach, USA: 2019 IEEE Proceedings of the Conference on Computer Vision and Pattern Recongnition, 2019: 1019-1028.

[29] Brazil G, Liu X. M3D-RPN: Monocular 3D Region Proposal Network for Object Detection[C]. South Korea: International Conference on Computer Vision (ICCV), 2019: 9286-9295.

[30] Zhou X, Wang D, Krähenbühl P. Objects as points[J]. arXiv preprint arXiv: 1904.07850, 2019.

[31] Li P, Zhao H, Liu P, et al. RTM3D: Real-Time Monocular 3DDetection from Object Keypoints for Autonomous Driving[C]. Glasgow, UK: European Conference on Computer Vision, 2020: 644-660.

[32] Ng M , Huo Y , Yi H , et al. Learning Depth-Guided Convolutions for Monocular 3D Object Detection[C]. Seattle, WA, USA: 2020 IEEE Conference on Computer Vision and Pattern Recognition, 2020: 11669-11678.

[33] X. Chen, K. Kundu, Y. Zhu, et al. 3D Object Proposals Using Stereo Imagery for Accurate Object Class Detection[C]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018: 1259-1272.

[34] Li P, Chen X, Shen S. Stereo R-CNN Based 3D Object Detection for Autonomous Driving[C]. Los Angeles CA, USA: 2019 IEEE Conference on Computer Vision and Pattern Recognition, 2019: 7636-7644.

[35] Peng W, Pan H, Liu H, et al. IDA-3D: Instance-Depth-Aware 3D Object Detection From Stereo Vision for Autonomous Driving[C]. Seattle, WA, USA: 2020 IEEE Conference on Computer Vision and Pattern Recognition, 2020: 13012-13021.

[36] K. Klasing, D. Wollherr, M. Buss. A clustering method for efficient segmentation of 3D laser data[C]. Pasadena, CA, USA: IEEE International Conference on Robotics and Automation, 2008: 4043-4048.

[37] J.Papon, A.Abramov, M.Schoeler, et al. Voxel Cloud Connectivity Segmentation-Supervoxels for Point Clouds[C]. Portland, OR, USA: 2013 IEEE Conference on Computer Vision and Pattern Recognition, 2013: 2027-2034.

[38] Y. Zhou, O. Tuzel. VoxelNet: End-to-End Learning for Point Cloud Based 3D Object Detection[C]. Salt Lake City, USA: 2018 IEEE Conference on Computer Vision and Pattern Recognition, 2018: 4490-4499.

[39] Wang B, An J, Cao J. Voxel-FPN: multi-scale voxel feature aggregation in 3D object detection from point clouds[C]. Long Beach, USA: 2019 IEEE Computer Vision and Pattern Recognition, 2019, 20(3): 704.

[40] R. Q. Charles, H. Su, et al. PointNet: Deep Learning on Point Sets for

3D Classification and Segmentation[C]. Hawaii, USA: 2017 IEEE Conference on Computer Vision and Pattern Recognition, 2017: 77-85.

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

[42] Qi C R , Wei L , Wu C , et al. Frustum PointNets for 3D Object Detection from RGB-D Data[C]. Salt Lake City, USA: 2018 IEEE Conference on Computer Vision and Pattern Recognition, 2018: 918-927.

[43] Qi C R, Litany O, He K, et al. Deep Hough Voting for 3D Object Detection in Point Clouds[C]. South Korea: IEEE International Conference on Computer Vision (ICCV), 2019: 9277-9286.

[44] Yang Z, Sun Y, Liu S, et al. 3DSSD: Point-based 3D Single Stage Object Detector[C]. Seattle, WA, USA: 2020 IEEE Conference on Computer Vision and Pattern Recognition, 2020: 11040-11048.

[45] Chen X, Ma H, Wan J, et al. Multi-view 3d object detection network for autonomous driving[C]. Hawaii, USA: 2017 IEEE Conference on Computer Vision and Pattern Recognition, 2017: 1907-1915.

[46] Ku J, Mozifian M, Lee J, et al. Joint 3d proposal generation and object detection from view aggregation[C]. Salt Lake City, USA: 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2018: 1-8.

[47] Xu D , Anguelov D, Jain A. PointFusion: Deep Sensor Fusion for 3D Bounding Box Estimation[C]. Salt Lake City, USA: 2018 IEEE Conference on Computer Vision and Pattern Recognition, 2018: 244-253.

[48] He K, Zhang X, Ren S, et al. Deep Residual Learning for Image Recognition[C]. Las Vegas, NV, USA: 2016 IEEE Conference on Computer Vision and Pattern Recognition, 2016: 770-778.

[49] Shi S, Guo C, Jiang L, et al. PV-RCNN: Point-Voxel Feature Set Abstraction for 3D Object Detection[C]. Seattle, WA, USA: 2020 IEEE Conference on Computer Vision and Pattern Recognition, 2020: 10526-10535.

[50] Scarselli F, Gori M, Tsoi A C, et al. The graph neural network model[J]. IEEE transactions on neural networks, 2008, 20(1): 61-80.

[51] Shi W, Rajkumar R. Point-gnn: Graph neural network for 3d object detection in a point cloud[C]. Seattle, WA, USA: 2020 IEEE Conference on Computer Vision and Pattern Recognition, 2020: 1711-1719.

[52] Najibi M, Lai G, Kundu A, et al. DOPS: Learning to Detect 3D Objects and Predict their 3D Shapes[C]. Seattle, WA, USA: 2020 IEEE Conference on Computer Vision and Pattern Recognition, 2020: 11910-11919.

[53] Uhrig J, Schneider N, Schneider L, et al. Sparsity Invariant CNNs[C]. Qingdao, China: 2017 International Conference on 3D Vision (3DV), 2017: 11-20.