智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

融合体素图注意力的三维目标检测算法

鲁斌, 孙洋, 杨振宇

引用本文:
鲁斌, 孙洋, 杨振宇. 融合体素图注意力的三维目标检测算法[J]. 智能系统学报, 2024, 19(3): 598-609.
LU Bin, SUN Yang, YANG Zhenyu. 3D object detection algorithm with voxel graph attention[J]. *CAAI Transactions on Intelligent Systems*, 2024, 19(3): 598-609.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202209008

DOI: 10.11992/tis.202209008
网络出版地址: https://link.cnki.net/urlid/23.1538.TP.20230914.0902.002

融合体素图注意力的三维目标检测算法

鲁斌[1,2], 孙洋[1,2], 杨振宇[1,2]
（1. 华北电力大学 控制与计算机工程学院, 河北 保定 071003; 2. 复杂能源系统智能计算教育部工程研究中心, 河北 保定 071003）

摘　要: 目前基于点云的三维目标检测方法未能充分利用点云局部几何特征, 导致对点云稀疏的目标检测效果不佳。为此, 本文提出基于原始点云体素图注意力的两阶段三维目标检测算法 (voxel graph attention region-CNN, VGT-RCNN)。通过多尺度体素特征插值计算网格中心点特征; 在多尺度非空体素特征上构造局部图; 通过图注意力机制对体素特征进行加权平均, 充分提取并利用目标的局部几何特征完成检测。该算法主要针对当前二阶段算法在进行特征聚合时对不同体素特征的重要性考虑不足进行改进, 引入可学习的权重矩阵, 动态学习体素特性的权重, 提高局部特征表达能力。本文在流行的 KITTI 自动驾驶数据集上进行了充分测试, 取得了具有竞争力的检测效果, 尤其是在对点云稀疏的汽车目标检测上, 准确率有较大提高。本文还对检测效果进行了可视化分析。

关键词: 点云; 三维目标检测; 图注意力; 特征插值; 多尺度特征; 激光雷达; 体素化; 车辆检测
中图分类号: TP391.4　　**文献标志码:** A　　**文章编号:** 1673−4785(2024)03−0598−12

中文引用格式: 鲁斌, 孙洋, 杨振宇. 融合体素图注意力的三维目标检测算法 [J]. 智能系统学报, 2024, 19(3): 598–609.
英文引用格式: LU Bin, SUN Yang, YANG Zhenyu. 3D object detection algorithm with voxel graph attention[J]. CAAI transactions on intelligent systems, 2024, 19(3): 598–609.

3D object detection algorithm with voxel graph attention

LU Bin[1,2], SUN Yang[1,2], YANG Zhenyu[1,2]
(1. School of Control and Computer Engineering, North China Electric Power University, Baoding 071003, China; 2. Engineering Research Center of Intelligent Computing for Complex Energy Systems, Ministry of Education, Baoding 071003, China)

Abstract: Current point cloud-based 3D object detection methods fail to fully use the local geometric features of the point clouds, leading to poor performance in detecting objects of sparse point clouds. To solve this problem, a two-stage 3D object detection algorithm named voxel graph attention region-CNN (VGT-RCNN) is proposed based on the voxel graph attention of raw point clouds. Initially, the grid center point features are calculated by multiscale voxel feature interpolation. Then, a local graph is constructed on the multiscale non-empty voxel features. Finally, a weighted average is conducted for the voxel features by graph attention mechanism, fully extracting and using the local geometric features of the object to complete detection. The algorithm mainly improves the defect of the present two-stage algorithm, which does not sufficiently consider the significance of different voxel features in feature clustering. In addition, a learnable weight matrix is introduced to dynamically learn the weight of the voxel feature and increase the expression ability of local features. The algorithm has been sufficiently tested on the popular KITTI autonomous driving dataset, obtaining competitive detection effects. The accuracy of cars with sparse point clouds has been markedly improved. A visualized analysis is also carried out to determine the detection effect.

Keywords: point cloud; 3D object detection; graph attention; feature interpolation; multiscale features; LiDAR; voxelization; car detection

点云是由激光雷达射出的一系列激光照射到物体后产生的反射点空间坐标构成。激光雷达对光照、距离和复杂环境十分鲁棒, 能够准确便捷地获取环境的三维点云信息, 因此成为三维目标检测领域最流行的传感器, 并在自动驾驶、机器视觉等领域发挥着越来越重要的作用。基于点云的三维目标检测是指以目标所处环境的三维点云为输入, 经算法处理后得到紧凑包围目标的三维边界框, 包括对目标在点云中的位置、体积大小、

收稿日期: 2022−09−06.　**网络出版日期:** 2023−09−14.
基金项目: 国家自然科学基金项目（62371188）; 河北省在读研究生创新能力培养项目（CXZZBS2023153）.
通信作者: 鲁斌. E-mail: lubin@ncepu.edu.cn.

朝向和类别等信息的预测结果。点云具备无序、稀疏和置换不变的特点，使得无法直接通过成熟的二维目标检测算法来对点云中的目标进行检测[1]。在实际场景中，目标所包含点的数量会受到遮挡和距离影响，导致点云对同一目标形状描述存在较大差异。因此，从点云中检测目标仍然是一项充满挑战的工作。

近年来，基于点云的三维目标检测算法发展迅速。Charles 等提出了 PointNet[2] 和 PointNet++[3]，将深度学习思想第一次应用到点云领域。自此以后，基于深度学习的点云目标检测开始蓬勃发展。为了提高检测准确度，针对点云目标尺度不变的特点，一些方法选择通过将被遮挡的点云信息进行补全来实现增加目标点云信息量的目的。例如 VoteNet[4] 通过霍夫投票思想，对目标中心点进行预测。Point-RCNN[5] 通过对点进行监督，使用前景点进行框的回归。3DSSD (3D single stage object detector)[6] 通过特征距离采样，获得更多前景点，然后通过投票机制预测物体的实例质心。SASSD (structure aware single-stage 3D object detection)[7] 通过辅助网络存储点云的空间结构信息来提高检测效果。SESSD (self-ensembling single-stage object detector)[8] 通过辅助网络学习目标特征，然后对感兴趣区域（region of interest, RoI）内的目标点云进行补齐。BtcDet[9] 将点云映射到球坐标下，通过点云补齐恢复遮挡区域的目标。虽然对点云进行补齐会提升检测准确度，但检测效果依赖于补齐效果，实际上无法有效学习到有遮挡或距离较远等点云较为稀疏的目标的特征。

另一类方法选择直接从稀疏点云中学习目标特征，通过对局部、区域或全局特征进行多尺度特征学习，得到能够描述目标的底层特征，以提高检测效果。例如 PV-RCNN[10] 提出点可以保留更多目标空间信息，选择在二阶段聚合多尺度的点和体素的特征，得到了较高的检测效果。Voxel-RCNN[11] 选择对多尺度的体素特征进行聚合，在车辆检测中取得了较高的准确率，但对于行人或自行车的检测效果不佳。Pyramid-RCNN[12] 通过将感兴趣区域划分为多层大小不同的网格来学习多尺度的特征。CT3D[13] 将 Transformer[14] 应用到点云特征编码阶段来对点之间的关系进行建模，对于有遮挡目标的检测效果提升明显。PointGNN[15] 基于图神经网络对局部点云进行特征提取，但对点不够密集的目标检测效果不佳。

当前的主流方法对于局部特征关系的建模基于多尺度的特征聚合或点间的注意力编码完成。基于特征聚合的方法大多使用最大池化函数聚合局部特征，并没有考虑不同点或体素对特征聚合在贡献度上的差异。点间的注意力编码虽然考虑了点间关系，但不可避免地要对点云进行采样，使得点云空间信息丢失严重，且直接对点进行处理效率较低[16]。

由激光雷达原理所致，受到遮挡或距离的影响，物体只在靠近雷达一侧的局部区域存在反射点。因此，即使是采用点云补全思路的检测方法，也需要从局部特征学习目标点云的分布。所以对局部特征的学习对于识别目标非常重要。

为了解决上述问题，本文提出基于体素图注意力思想的 VGT-RCNN (voxel graph attention region-CNN)，创新性地将图注意力思想应用到多尺度体素，提出 Voxel-wise 图注意力机制，以提高对局部几何特征的建模能力。同时，为避免点云稀疏性导致的网格中心点特征计算困难，本文还提出多尺度网格插值池化，用以稳定计算网格中心点特征，保留更多点云空间信息。综上，本文的贡献如下：

1）提出一种多尺度网格中心点插值池化方法，通过计算网格中心点周围最近的 Top-N 个体素的特征来对网格中心点进行特征插值，捕获感兴趣区域的全局特征，有效提升点云稀疏目标的检测效果；

2）提出多尺度体素图注意力机制，对 3D 骨干网络中的每一层非空体素进行多尺度局部图注意力编码，动态学习特征权重，提高算法对局部几何特征的建模能力。

1　相关工作

由于点云的无序、稀疏和置换不变的特点，使用深度学习模型处理点云首先要对点云进行规则化。根据处理点云方式的不同，点云规则化方法主要分为基于点的方法和基于体素的方法。

基于点的方法通常对输入的原始点云进行分层采样和特征聚合。这样做的好处是可以将点云聚合为指定数量的点，以便设计不同的模型来提取点云特征。PointNet 和 PointNet++是这一方向的开创性工作，其将点映射到高维特征空间后，通过分层下采样逐步聚合点云空间信息，实现对点云的特征提取。后续的很多工作都基于此提出了改进。例如 Point-RCNN 采用 PointNet++作为骨干网络，使用最远点采样逐步对点云进行下采样以生成感兴趣区域。3DSSD 改进了点采样方法，结合最远几何距离采样和最远特征距离采样两种采样方式，获得了更多包含目标的前景点。IASSD[17] 学习每个点的语义，以针对不同类别目

标进行采样。一些工作为了提取更丰富的点云特征，融入了注意力机制。例如 PointFormer[18] 使用 Transformer 作为主干网络提取点云特征，通过局部和全局 Transformer 捕捉多尺度点云特征之间的依赖关系。CT3D 将 Transformer 应用到二阶段细化过程中，在感兴趣区域内进行随机采样，使用 Channel-wise 的注意力机制对感兴趣区域进行特征提取。

基于点的方法的难点主要是如何在性能和推理时间之间进行平衡。在进行采样和特征聚合时，往往需要选取合适的点采样方法和球查询半径。最远点采样作为常用的采样方法计算效率较低。较大的球查询半径可以获得更丰富的上下文信息，但会增加模型推理时间和内存消耗。有学者对此提出了改进方法，例如 3DSSD 和 IASSD 选择在不增加计算资源消耗的前提下，增加对前景点的采样数量。随着多线激光雷达技术的不断发展，点云场景中点的数量会不断增加，进一步提升对计算资源的需求。

基于体素的方法首先将点云规则化为均匀的网格，之后使用成熟的 2D 或 3D 卷积进行特征提取。例如 VoxelNet 把点云体素化后使用 3D 卷积进行特征提取，之后再将特征压缩到鸟瞰视角（bird's eye view, BEV）来检测目标。SECOND[19] 对 3D 卷积进行改进，提出了稀疏 3D 卷积，即只在非空体素上进行卷积操作，提高了 3D 卷积效率。PointPillars[20] 将点云转化为柱状结构后对每个柱体提取特征，然后将特征压缩到 BEV 视角，再使用 2D 卷积进行特征提取。此外，还有一系列基于体素的二阶段方法。例如 Voxel-RCNN 使用 SECOND 作为骨干网络，在二阶段聚合了多尺度的体素特征来对预测框进行细化调整。CIASSD[21] 融合多层空间和语义特征，使用交并比（intersection over union, IoU）来监督预测框的置信度。Voxel-FPN[22] 聚合多尺度体素特征，取得比使用单一尺度特征的方法更好的性能。

基于体素的方法的问题主要在于体素化会不可避免地带来点云分辨率的损失。大尺寸体素可以提高卷积效率，但会造成点云空间信息损失较多；小尺寸体素可以避免损失过多点云空间信息，但是会带来更多的计算量。选择合适的体素尺寸和体素内点的编码方式是一个难点。

近年来，由于图结构的强大表现力，基于图的神经网络相关技术发展迅速。图神经网络可以方便地对点类型数据进行建模，对处理点云数据具有独特优势。PointGNN 将图神经网络用于点云目标检测。其使用固定球半径对邻域点云进行有效编码，在 KITTI 数据集上取得了当时最好的检测效果。但由于需要多次迭代建图，使得单阶段的 PointGNN 并没有明显的速度优势。并且由于使用了固定的球查询半径，会出现在指定半径内不存在可用于局部特征聚合的点，导致其对于点云稀疏目标的检测效果一般。最近，李文举等[23] 提出结合图采样和图注意力的三维目标检测算法，即在 PointGNN 中引入图采样和多级注意力机制，提升了对 KITTI 数据集中中等和困难级别目标的检测效果，但检测效果仍然有限。

与 PointGNN 不同，本文提出的 VGT-RCNN 将图注意力思想应用到预测框的二阶段细化，并基于图注意力机制动态学习体素特征的权重，提高模型对局部特征的建模能力。此外，受 Voxel-RCNN 和 Voxel-FPN 启发，我们发现聚合多尺度的体素特征能够有效提高检测效果。因此 VGT-RCNN 选择使用多尺度体素特征插值来计算网格中心点特征，以强化中心点的空间表达能力。对于点云稀疏目标，当前方法在将感兴趣区域网格化后，会出现网格中心点的局部邻域内点云稀少、存在空白区域的问题。例如 PointNet++或 Voxel-RCNN 均采用固定半径的欧氏距离或曼哈顿距离聚合周围体素特征。这会导致无法在包含点较少的感兴趣区域上计算中心点特征。而基于特征插值的方法则通过选取距离网格点最近的 N 个非空体素对网格中心点进行插值，能够有效避免无法聚合特征的问题。

具体来说，VGT-RCNN 首先通过多尺度体素特征插值计算网格中心点特征，并使用插值采样点的坐标与中心点坐标之间的欧氏距离作为中心点特征的补充信息，增强中心点的特征表达能力；然后进行图注意力建模，引入可训练的注意力权重矩阵动态学习中心点邻域内非空体素特征的权重；最后通过加权平均聚合局部邻域特征。实验证明，VGT-RCNN 在 KITTI 数据集上对于简单、中等和困难级别的目标的检测效果相较现有的方法均得到提升，尤其是对于困难级别汽车目标的检测效果提升显著。

2　VGT-RCNN 模型

VGT-RCNN 是两阶段、可端对端训练的三维目标检测算法。VGT-RCNN 使用 SECOND 模型作为第 1 阶段来生成感兴趣区域，包括 3D 骨干网络、2D 骨干网络、RPN 网络和一阶段检测头等结构；在第 2 阶段基于体素图注意力对预测框进行细化，包括多尺度网格点特征插值模块、动态图注意力池化模块和二阶段检测头等结构。具体来说，在二阶段细化中，首先使用距离网格中心

点最近的n个非空体素特征对中心点特征进行插值, 接着动态计算局部体素特征权重, 并在对特征加权后进行特征聚合, 实现对局部特征的充分学习。图 1 为 VGT-RCNN 网络结构。

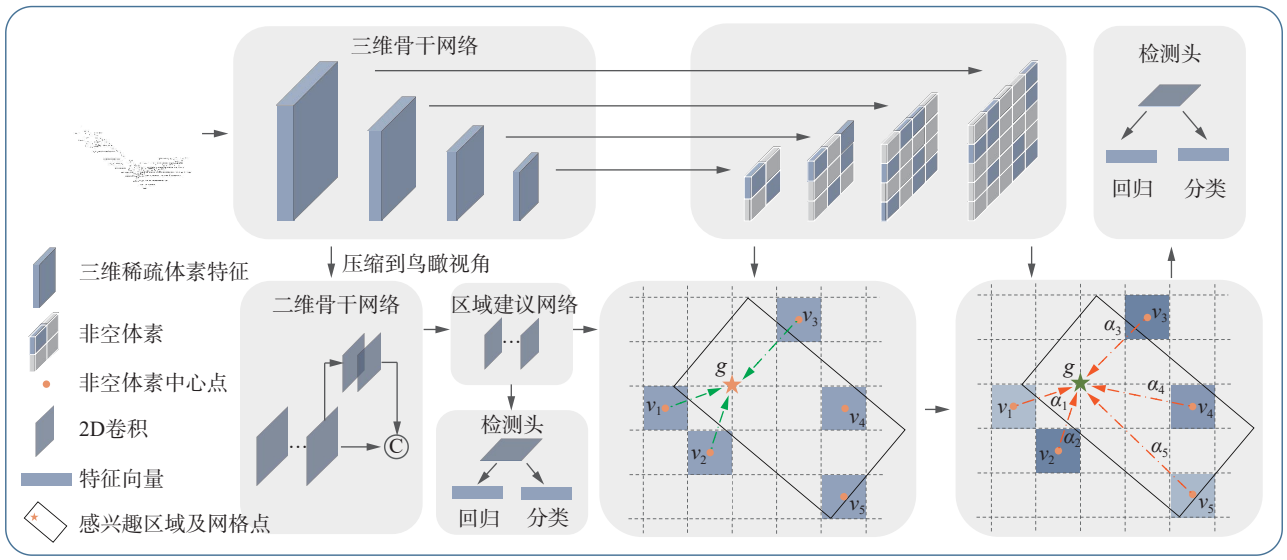


图 1　VGT-RCNN 网络结构

Fig. 1　Network structure of VGT-RCNN

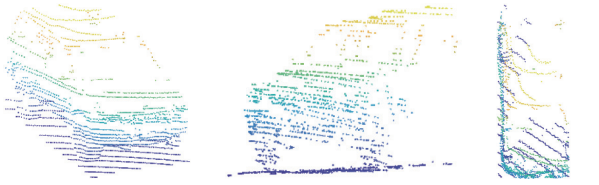
2.1　基于 SECOND 的一阶段骨干网络

作为输入的原始点云集合表示为$P=\{p_1,p_2,\cdots,p_n\}$, 其中$p_i=\{x_i,y_i,z_i,r_i\}$, x_i、y_i、z_i为点的三维坐标, r_i为点的反射率。将整个三维点云空间划分为均匀体素$V\in\mathbf{R}^{L\times W\times H}$, L、W、H为点云空间各方向上体素的数量。接着使用 3D 稀疏卷积对三维体素空间进行特征提取, 经过 4 层稀疏卷积, 依次进行 2 倍下采样操作, 得到不同尺度的体素特征, 各层体素特征表示为$f_v=\{f_v^{1\times},f_v^{2\times},f_v^{4\times},f_v^{8\times}\}$, 其中, 上标$n\times$表示下采样倍数。之后将$f_v^{8\times}$沿$z$轴压缩到 BEV 视角, 得到 BEV 视角下的特征f_{bev}, 再将f_{bev}输入区域建议网络(region proposal network, RPN)得到感兴趣区域。接下来使用本文提出的方法, 对感兴趣区域特征进行细化。

2.2　多尺度网格点特征插值

两阶段目标检测方法通常在第二阶段对感兴趣区域进行池化, 以得到固定尺寸的特征。常用的方法有点采样和网格池化方法。点采样会丢失很多前景点, 不利于学习目标特征, 且采样点数量无法根据目标所包含点云的稀疏程度来设定。如果采样点数量设置较多, 则不利于对含点数量较少的目标进行采样。一旦采样点数量超过其感兴趣区域本身所包含点的数量, 通常会取点坐标平均值来进行补充, 造成点云的空间几何特征的损失。反之, 若采样点数量设置较少, 则目标信息丢失会过于严重, 同样不利于学习目标特征。

图 2 和图 3 形象化地体现了这两种情况。图 2 为 KITTI 数据集中一辆点云密集的车辆的不同视角, 黄色点为车顶, 蓝色点为地面。图 2(a) 为面向雷达侧, 点云较为密集。图 2(b) 为另一侧, 车尾部分由于背对雷达, 没有点云。图 2(c) 为俯视图, 从图中可见, 除了车顶和地面, 车内部不包含点云。因此, 当基于车辆内部网格中心点进行特征聚合时, 会出现无法计算特征的问题。图 3 为两辆点云稀疏的汽车俯视图, 其网格点特征的计算相对更为困难。



(a) 雷达侧视角　(b) 非雷达侧视角　(c) 鸟瞰视角

图 2　汽车点云图

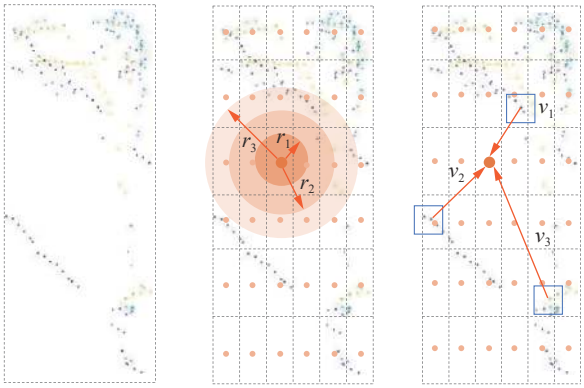
Fig. 2　Point clouds of car

(a) KITTI中等难度汽车　(b) KITTI困难难度汽车

图 3　点云数量少的汽车

Fig. 3　Car with sparse point cloud

网格池化方法在 PV-RCNN 中提出。该方法首先把感兴趣区域划分为固定数量网格，并将特征整合到网格中计算，再基于网格中心点使用球查询方法聚合局部特征，以得到均匀的空间特征。后续算法大多采用这种方法进行二阶段池化。由于点云只存在于物体表面，使用网格池化方法计算物体内部网格特征通常需要较大的球查询半径，而较大的球查询半径会使处在感兴趣区域边缘的网格聚合到过多的背景点，导致提取到的特征包含冗余且无关的上下文信息，影响模型性能。图 4 为鸟瞰视角下感兴趣区域网格中心点特征计算的方法对比。图 4(a) 为感兴趣区域的 BEV 视角，图 4(b) 为通过球查询方式进行局部特征聚合，可以看到在球半径r_1范围内没有点，r_2范围内仅有少量点，直到半径扩大到r_3，才有较多数量的点。这说明通过球半径查询的方式并不能真实反映中心点特征。



(a) 鸟瞰视角感兴趣区域　(b) 球半径查询聚合法　(c) 特征插值法

图 4　鸟瞰视角感兴趣区域网格中心点特征计算

Fig. 4　Point feature calculation of RoI grid centroids

为了解决上述问题，我们提出了 Top-N 特征插值的方法来计算网格中心点特征，图 4(c)为$N=3$时的示意图。在三维空间中，用距离网格中心点最近 3 个非空体素特征对网格中心点进行特征插值，计算公式为

$$f'_{g_i}=\sum_{i=1}^{n}\frac{d_i f_i}{\sum_{k=1}^{n} d_k} \tag{1}$$

式中：f'_{g_i}表示网格中心点特征，n为距离g_i最近的非空体素数量，f_i为第i个非空体素的特征，d_i为第i个体素v_i与网格中心点g_i之间的欧氏距离，$\sum_{k=1}^{n} d_k$为最近的n个非空体素与网格中心点距离之和。在计算距离时，使用体素质心坐标作为体素坐标，令

$$d_i=\sqrt{(x_i-x_g)^2+(y_i-y_g)^2+(z_i-z_g)^2} \tag{2}$$

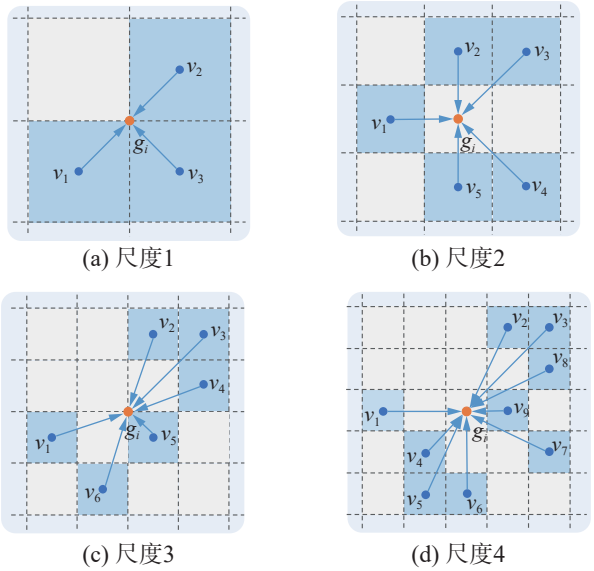
式中：x_i、y_i和z_i为非空体素坐标，x_g、y_g和z_g为网格中心点坐标。将网格点特征的计算方式推广到多尺度体素特征的计算，则网格中心点的多尺度体素特征插值计算公式为

$$f_{g_i}=g\left(\sum_{j=1}^{m}\sum_{i=1}^{n}\frac{d_{i,j}f_{i,j}}{\sum_{k=1}^{n} d_{k,j}}\right) \tag{3}$$

式中：$g(\cdot)$为特征聚合函数，由尺寸为1×1的卷积核、批量归一化函数和LeakyReLU激活函数构成。f_{g_i}为网格中心点多尺度特征，m为 3D 骨干网络中间体素特征的数量，n的定义同前，$d_{i,j}$表示第j个尺度中，用于对网格点g进行特征差值的第i个非空体素与g之间的欧氏距离，$f_{i,j}$为该体素对应的特征。

2.3　动态体素图注意力

传统采样方法中基于多尺度特征聚合的方法无法动态调整特征的权重，使得体素之间相对独立，缺乏对体素间复杂几何关系的建模，限制了模型对局部特征表达能力。为解决上述问题，本文提出了动态体素图注意力方法对局部几何特征进行建模。该方法具有两个特点：第一，可以动态计算中心点附近体素特征的注意力权重，提取更细粒度的局部特征；第二，可以在多尺度体素特征上动态构造局部邻域图，生成融合多尺度的图注意力特征，有效提高特征的表达能力。图 5 为 4 个尺度的体素特征，其中低层特征拥有更高的分辨率，包含更细粒度的空间几何特征，高层特征则包含更丰富的语义特征。



(a) 尺度1　(b) 尺度2　(c) 尺度3　(d) 尺度4

图 5　多尺度体素图的构建

Fig. 5　Multi-scale voxel graphs construction

2.3.1　体素图构建

输入网格中心点g_i和多尺度体素特征f_v，定义

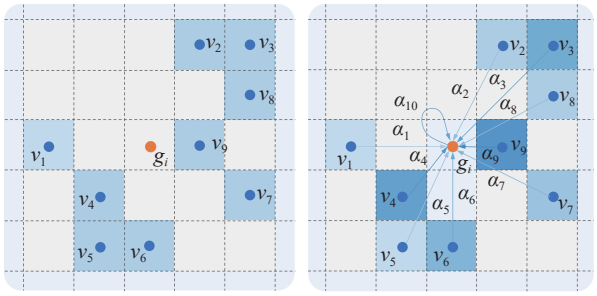
点云局部邻域图为$G=(V,E)$。令$G=\{G_1,G_2,\cdots,G_i\}$, i为 3D 骨干网络中体素尺度的数量。取网格中心点g_i作为图的中心, 距离g_i半径为r的球形邻域范围内的体素为顶点V, 即体素特征集合f_v中的体素v_i, E为v_i到g_i的连线:

$$E=\left\{(g_i,v_{i,n})\left|\sqrt{(g_{i,x}-v_{i,x_n})^2+(g_{i,y}-v_{i,y_n})^2+(g_{i,z}-v_{i,z_n})^2}\right.\right\} \tag{4}$$

式中: $v_{i,n}$表示第i个网格中心点的局部邻域内的体素, n为体素的索引。在每一层体素特征中, 构建具有不同球形半径r的局部邻域图, 表示为$G_i=\{g_{r_1},g_{r_2},\cdots,g_{r_n}\}$。随着$r$的调整, 模型可以学习到不同感受野下的图特征。

2.3.2　网格中心点图注意力特征

定义中心点特征$e_{i,j}=h(v_i,g_j)$, 其中, $h(\cdot)$为特征聚合函数, v_i为网格点g_i邻域中的第i个体素。以 PointNet++为例, 其仅对中心点的局部邻域特征进行聚合, 故$e_{i,j}$只和v_i相关, 未考虑局部点和中心点之间的特征关系, 如图 6(a) 所示, 即$h(v_i,g_j)=h(v_i)$。



(a) 无注图意力的体素特征　(b) 增加注意力后的体素特征

图 6　体素图注意力示意

Fig. 6　Voxel graph attention diagram

为了学习更具表达能力的局部特征, 我们在v_i和g_j之间引入可训练的线性变换矩阵$\boldsymbol{W}$, 以动态更新节点特征的权重, 如图 6(b) 所示, α_i为体素的注意力权重。在进行特征聚合时, 特征结合注意力权重进行加权平均, 得到局部邻域图特征。计算公式为

$$h(v_i,g_j)=\boldsymbol{W}(v_i-g_i)v_i \tag{5}$$

我们将网格中心点图注意力特征的计算分为 3 步: 1)聚合节点信息, 计算节点对于网格中心点的注意力系数; 2)将注意力系数对节点特征进行加权平均, 得到网格中心点的局部图特征; 3)将图节点坐标特征作为补充特征, 与局部图特征进行拼接。

首先, 输入图节点$V=\{v_1,v_2,\cdots,v_n\}$, 其中, $v_i(i\in[1,n])$包括节点的坐标(x_i^v,y_i^v,z_i^v)和节点的特征f_i。使用 Xavier[24] 对$\boldsymbol{W}$进行初始化, 尽可能保证输入和输出服从相同的概率分布, 节点的注意力系数为

$$w_{i,j}=\text{LeakyReLU}(\boldsymbol{W}(\Delta d,\Delta f)) \tag{6}$$

式中: Δd和Δf分别为v_i和g_i之间的欧氏距离和特征距离。使用Softmax函数对同一个子图内的节点权重进行归一化, 得到图的最终注意力权重:

$$\omega_{\text{Attention}}=\text{Softmax}(w_{i,j}) \tag{7}$$

然后, 使用注意力对节点特征进行加权, 得到网格中心点图注意力特征:

$$f_g^{\text{Attention}}=\omega_{\text{Attention}}\cdot f_i \tag{8}$$

式中f_i为体素特征。

最后, 网格中心点的图注意力特征的计算公式为

$$f_g^{\text{Attention}}=\sum_{k\in N}\frac{\exp(\text{LeakyReLU}(\boldsymbol{W}(v_i-g_j)))v_i}{\sum_{m\in N}\exp(\text{LeakyReLU}(\boldsymbol{W}(v_m-g_j)))} \tag{9}$$

式中: k为网格中心点数量, m为网格中心点邻域中体素节点的数量。

本文提出的图注意力的计算方法未使用图节点之间的连接结构, 具有较高的计算效率, 但会带来一定的图信息损失。为了补充图的结构信息, 我们将经过前馈神经网络(feed forward network, FFN)升维后的节点坐标作为图信息的补充特征, 拼接到节点的图注意力特征上, 以提高节点特征的表达能力, 计算方式如下:

$$f^{\text{pos}}=\text{FFN}(x_i,y_i,z_i,r_i) \tag{10}$$

$$f_g=\{f_g^{\text{Attention}},f^{\text{pos}}\} \tag{11}$$

式中: $f_g^{\text{Attention}}$表示图注意力特征, f^{pos}为体素的空间坐标特征。然后, 计算 3D 骨干网络中所有尺度体素的图注意力特征, 再将各尺度特征进行拼接, 得到感兴趣区域的多尺度体素图注意力特征:

$$f_g^{\text{multi}}=\{f_g^{1\times},f_g^{2\times},f_g^{4\times},f_g^{8\times}\} \tag{12}$$

最后, 将f_g^{multi}输入检测头进行候选框的分类和位置回归。

2.4　检测头与损失

网络最后输出包含 7 个自由度的三维预测框, 表示为(x,y,z,w,l,h,θ), 其中x、y、z为预测框中心点, w、l、h表示以中心点为参考的预测框的长、宽、高, θ表示预测框的朝向角度。在进行检测时, 数据集中真实框的标签称为真实值, 经过网络计算得到的框称为预测框。

模型初始化时, 在鸟瞰视角特征图的每个位置上预设两种不同方向的候选框, 然后根据框的置信度进行排序, 取前 512 个候选框进行非极大值抑制 (non-maximum suppression, NMS), 接着计算预测框和真实框的交并比, 以 0.55 为阈值进行正负样本划分, 最后进行框的分类和回归。

预测框和真实框之间的误差表示为$(\Delta x,\Delta y,\Delta z,$

$\Delta w,\Delta l,\Delta h,\Delta\theta$), 每个自由度的计算公式为

$$\begin{aligned}\Delta x=\frac{x_{\rm gt}-x_{\rm reg}}{d},\ \Delta y=\frac{y_{\rm gt}-y_{\rm reg}}{d},\ \Delta z=\frac{z_{\rm gt}-z_{\rm reg}}{h_{\rm reg}},\\ \Delta w=\log\frac{w_{\rm gt}}{w_{\rm reg}},\ \Delta l=\log\frac{l_{\rm gt}}{l_{\rm reg}},\ \Delta h=\log\frac{h_{\rm gt}}{h_{\rm reg}},\\ \Delta\theta=\theta_{\rm gt}-\theta_{\rm reg}\end{aligned}\tag{13}$$

式中：$d=\sqrt{w_{\rm reg}^2+l_{\rm reg}^2}$，下标$gt$表示真实框参数，reg表示预测框参数，$\Delta x$、$\Delta y$和$\Delta z$分别表示预测框和真实框中心点之间在3个维度上的误差值，Δw、Δl和Δh分别表示预测框和真实框之间在宽、长和高方向上的误差值。

网络的损失分为一阶段区域建议网络检测头损失$L_{\rm rpn}$和二阶段感兴趣区域检测头损失$L_{\rm rcnn}$，总损失计算公式为

$$L_{\rm total}=L_{\rm rpn}+L_{\rm rcnn}\tag{14}$$

检测头由带有非线性激活函数的前馈神经网络构成，预测框的位置和置信度分由两个独立的网络进行计算。

$L_{\rm rpn}$分为预测框置信度损失$L_{\rm cls}$和回归损失$L_{\rm reg}$，计算公式为

$$L_{\rm rpn}=L_{\rm cls}(c_g,c_g^*)+L_{\rm reg}(r_g,r_g^*)\tag{15}$$

式中：c_g和r_g为真实框类别和位置标签，c_g^*和r_g^*为对应的预测值。$L_{\rm cls}$的计算和SECOND保持一致，使用Focal Loss[25]平衡正负样本的损失：

$$L_{\rm cls}=-c_g\log(c_g^*)-(1-c_g)\log(1-c_g^*)\tag{16}$$

$L_{\rm reg}$使用Smooth-L1损失函数，仅使用正样本进行计算：

$$L_{\rm reg}=(c_g\geqslant 1)\sum_i L_{\rm Smooth-L1}(r_g,r_g^*)\tag{17}$$

式中$c_g\geqslant 1$表示正样本。

最后得到总的$L_{\rm rpn}$损失：

$$\begin{aligned}L_{\rm rpn}=\frac{1}{N_{\rm fg}}(-c_g\log(c_g^*)-(1-c_g)\log(1-c_g^*)+\\ (c_g\geqslant 1)\sum_i L_{\rm Smooth-L1}(r_g,r_g^*)\end{aligned}\tag{18}$$

式中$N_{\rm fg}$为包含被检测目标的前景框数量。二阶段检测头损失$L_{\rm rcnn}$的计算和$L_{\rm rpn}$的计算是相同的。

3　实验分析

我们在KITTI数据集上对VGT-RCNN进行训练和测试。KITTI[26]自动驾驶场景数据集包含7 481个训练样本和7 518个的测试样本。作为一种普遍使用的方法，训练样本被分为一个包含3 712个样本的训练集和一个包含3 769个样本的验证集[27]。根据目标所包含点的数量和被遮挡情况，KITTI将目标检测难度分为简单、中等和困难3个等级。本节给出了模型在测试集和验证集上的性能。

3.1　实验环境及参数

实验环境与参数如表1所示。

表 1　实验环境与参数

Table 1　Environment configuration for experiment

环境	参数
CPU	AMD EPYC 7543 32-Core Processor 15核
GPU	NVIDIA RTX 3 090
显存	24 GB
操作系统	Ubuntu 20.04
Python版本	3.8
深度学习框架	PyTorch 1.8.1
CUDA版本	11.1
cuDNN版本	8.0
代码编辑环境	Visual Studio Code 1.66.1

3.2　实验设置

1）点云预处理。对每帧点云在x、y、z 3个方向上分别限定大小为[0,70.4]m、[−40.0,40.0] m、[−3.0,1.0] m，舍弃场景外的点云。为防止点数量太少导致无法有效提取空间特征的问题，包含点数量少于5的目标会被过滤掉。最后，使用和SECOND一样数据增强方式，以保证实验结果的公平比较：在每帧点云中增加真实目标的数量，包括目标点云和标签；对点进行随机缩放和旋转，缩放倍数范围为[0.95,1.05]，旋转角度范围为[−π/4,+π/4]；对所有真实目标模拟转向，转向角度范围为[−π/2,+π/2]；沿x轴对点云进行随机翻转。

2）点云体素化。设定体素在x、y、z 3个方向上的大小为(0.05,0.05,0.1) m，则对应体素化后的三维体素空间大小为[1 408,1 600,40]。在体素编码阶段，三维体素空间以1×、2×、4×、8×倍数进行下采样，各层特征尺寸为[16,32,64,64]。

3）二阶段检测参数。在训练阶段，感兴趣区域数量设置为128。网格数量设置为[6×6×6]，使用距离网格点最近的3个非空体素进行插值。使用多尺度体素特征构造图，各层体素特征图构建半径分别为[0.4,0.4,0.8,1.6]，对应每一层每个顶点限制最大输入边的数量为[16,16,16,16]。在网格点特征插值时，将各层体素特征尺寸统一调整为32维。经过4层图注意力特征计算后，感兴趣区域的细化特征尺寸为128维。最后，将细化特征输入检测头进行预测框的分类和位置回归。

4）训练参数。对于KITTI数据集，训练80 epoch，batch size设置为8，使用Adam_onecycle优化器，学习率最高设置为0.01。

5）推理参数。感兴趣区域数量为100。在后处理过程中，NMS阈值为0.1，置信度阈值为0.7。

6）损失参数权重。VGT-RCNN的损失分为

预测框置信度、位置和朝向损失 3 种, 3 种损失权重分别为 1、2 和 0.2。

3.3　实验结果与分析

我们在 KITTI 验证集和测试集上对模型进行了实验, 并将测试集实验结果提交到官方平台与其他 SOTA 方法进行对比, 如表 2 所示, 取得了具有竞争力的效果。实验采用 KITTI 官方提供的评价指标衡量模型性能, 其中 IoU 阈值为 0.7, 准确率(average precision, AP)以 40 个召回位置计算。

我们将 VGT-RCNN 与其他 SOTA 方法进行了对比。在提交到官方平台测试时, 使用 80% 训练样本对模型进行训练, 置信度阈值设置为 0.7。实验结果如表 2 所示。在竞争激烈的 KITTI 测试集上, VGT-RCNN 以 14 f/s 的推理速度获得了具有竞争力的检测效果, 在简单、中等和困难类别检测中分别达到 91.24%、82.36% 和 79.34% 的准确率, 性能提升明显。实验结果已在 KITTI 官网公开。

表 2　KITTI 测试集上汽车检测性能比较

Table 2　Comparison of car detection performance on the KITTI test set

方法		推理速度/(f/s)	3D检测准确率/%		
			简单	中等	困难
点云+图像的方法	MV3D[28]	3.0	74.97	63.63	54.00
	AVOD-FPN[29]	10.0	83.07	71.76	65.73
	F-PointNet[30]	5.9	82.19	69.79	60.59
	ContFuse[31]	16.7	83.68	68.78	61.67
	PointSIFT+SENet[32]	—	85.99	72.72	64.58
	UberATG-MMF[33]	12.5	88.40	77.43	70.22
	3D-CVF[34]	13.3	89.20	80.05	73.11
基于点的方法	Point-RCNN[15]	10.0	86.96	75.64	70.70
	STD[35]	12.5	87.95	79.71	75.09
	Patches[36]	—	88.67	77.20	71.82
	3DSSD[6]	26.3	88.36	79.57	74.55
	PV-RCNN[10]	8.9	90.25	81.43	76.82
基于体素的方法	VoxelNet[3]	4.4	77.47	65.11	57.73
	SECOND[19]	20.0	83.34	72.55	65.82
	PointPillars[20]	42.0	82.58	74.31	68.99
	Voxel R-CNN[10]	25.2	90.90	81.62	77.06
	CT3D[12]	—	87.83	81.77	77.16
	SASSD[18]	25.0	88.75	79.79	74.16
	VGT-RCNN(本文)	14.0	91.24	82.36	79.34

我们与同类基于图的方法 PointGNN 进行了对比, 如表 3 所示。从表中数据可见, VGT-RCNN 的检测准确率具有明显优势。在 KITTI 数据集, 对简单、中等和困难汽车类别检测 3D 准确率分别领先 Point-GNN 约 2.91%、2.89% 和 7.05%。同时, VGT-RCNN 在推理速度方面优势明显, 以 14 f/s 的推理速度领先, 约为 PointGNN 的 8.9 倍。这是由于 PointGNN 对点云空间进行全局建图, 而 VGT-RCNN 仅在感兴趣区域进行建图, 减少了对背景信息的无效建图, 且只考虑网格点和体素之间的连接, 未考虑体素与体素之间的关系, 使建图效率得到极大提升。

表 3　VGT-RCNN 与 PointGNN 在 KITTI 测试集的对比

Table 3　Comparison of VGT-RCNN and PointGNN on the KITTI test set

方法	推理速度/(f/s)	3D检测准确率/%			BEV检测准确率/%		
		简单	中等	困难	简单	中等	困难
PointGNN[15]	1.56	88.33	79.47	72.29	93.11	89.17	83.90
VGT-RCNN	14.00	91.24	82.36	79.34	94.59	90.89	86.36
提升	12.44	2.91	2.89	7.05	1.48	1.72	2.46

我们还在验证集上与其他 SOTA 方法进行对比，如表 4 所示。在中等难度类别检测效果上达到最好，在困难难度类别检测中，仅落后 SA-SSD 约，进一步表明本文提出的方法的有效性。我们分析认为是基于插值的网格点特征池化对有遮挡或含点云数量少的目标具有特征强化的作用，可以提取到更有效的网格点特征，进而提升了模型性能。

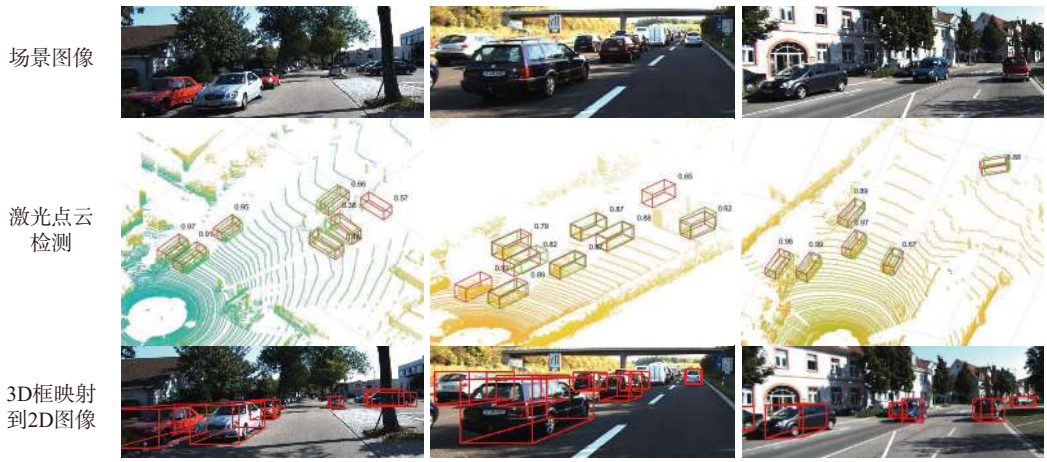
表 4　KITTI 验证集上汽车检测性能比较

Table 4　Comparison of car detection performance on the KITTI validation set

方法		推理速度/(f/s)	3D检测准确率/%		
			简单	中等	困难
基于点的方法	Point-RCNN9[5]	10.0	88.88	78.63	77.38
	STD[35]	12.5	89.70	79.80	79.30
	3DSSD[6]	26.3	89.71	79.45	78.67
	PV-RCNN[10]	8.9	89.35	83.69	78.70
基于体素的方法	VoxelNet[3]	4.4	81.97	65.46	62.85
	SECOND[19]	20.0	88.61	78.62	77.22
	PointPillars[20]	42.0	86.62	76.06	68.91
	SASSD[7]	25.0	90.15	79.91	78.78
	VGT-RCNN(本文)	14.0	89.61	84.01	78.76

3.4　可视化分析

我们将 VGT-RCNN 的检测结果在 KITTI 数据集的多个实际场景中进行了可视化展示，包括乡村道路、城市街道、高速公路，如图 7 所示。其中红色框为预测框，绿色为真实框。可以看到 VGT-RCNN 预测的红色框与真实框重合度较高，说明预测框位置检测精确。图中还给出了框的置信度分数，置信度越接近 1 表示类别预测越准确。整体上看，模型对置信度的预测准确度较高，能够准确检测到场景中的汽车，在图 7(a) 和 7(c) 中，模型检测到数据集中未被标记的汽车，证明模型具有较好的泛化能力。



(a) 场景一　(b) 场景二　(c) 场景三

图 7　VGT-RCNN 可视化示意

Fig. 7　VGT-RCNN visualization diagram

3.5　消融实验

对所提出的模型进行消融实验，以分析每个组件的性能。所有实验均在 KITTI 训练集和验证集上进行，训练集和验证集的划分遵循文献 [27] 提出的方法。准确率以 40 个召回位置进行计算。

3.5.1　模块有效性

表 5 是各模块有效性实验，其中 V.G.A 表示体素图注意力模块，M.F.I 表示多尺度体素特征插值模块。当仅使用特征插值时，虽然对中等和困难级别目标检测效果一般，但准确率依然有所提升。这是由于特征插值仅计算目标全局特征，没有聚合多尺度感受野的局部特征。多尺度体素图注意力特征的计算依赖于网格中心点特征，我们将插值特征替换为点的坐标特征进行对比实验，

结果表明多尺度体素图注意力能够提高对局部几何结构的建模能力, 可以有效提高检测准确度。将网格中心点特征换成多尺度插值特征后, 检测性能得到进一步提高, 这说明尽管中等和困难级别目标点云稀疏, 特征插值仍可以有效保留点云空间信息, 从而提高检测准确率。

表 5　多尺度特征插值和体素图注意力在 KITTI 验证集上的消融实验

Table 5　Multi-scale feature interpolation and voxel graph attention ablation experiment on the KITTI validation set　%

V.G.A	M.F.I	3D检测准确率		
		简单	中等	困难
√	×	89.52	79.35	78.68
×	√	89.54	79.41	78.63
√	√	89.61	84.01	78.76

3.5.2　多尺度体素特征有效性

在进行特征插值和体素图注意力计算时, 模型使用了 3D 骨干网络中的多层体素特征, 我们验证了多层体素特征对模型性能的影响, 从 $f_v^{1\times}$ 至 $f_v^{8\times}$ 逐步增加体素特征数量, 每组实验使用相同的超参数, 实验结果如表 6 所示。随着加入多尺度特征, 模型性能逐渐提高。这是因为不同尺度的体素特征可以提供不同感受野的特征, 对局部特征的学习具有补充作用。这也证明了本文提出的多尺度特征插值方法和体素图注意力机制的有效性。

表 6　多尺度体素特征有效性在 KITTI 验证集上的消融实验

Table 6　Multi-scale voxel feature effective ablation experiment on the KITTI validation set　%

$f_v^{1\times}$	$f_v^{2\times}$	$f_v^{4\times}$	$f_v^{8\times}$	3D检测准确率		
				简单	中等	困难
√	×	×	×	88.87	78.41	77.13
√	√	×	×	89.17	82.74	78.45
√	√	√	×	89.22	83.25	78.54
√	√	√	√	89.61	84.01	78.76

4　结束语

本文研究了当前点云目标检测方法对局部几何特征提取能力不足的问题, 提出了基于体素图注意力的两阶段三维目标检测方法 VGT-RCNN。针对感兴趣区域网格池化过程中对点云稀疏目标网格中心点特征计算不稳定的问题, 提出了多尺度体素特征插值方法, 提升了网格中心点特征计算的稳健性。提出了体素图注意力机制, 可以聚合局部几何特征并自适应计算局部体素特征权重, 使模型的局部特征表达能力更加鲁棒。我们在 KITTI 数据集上进行了大量的实验并与现有的 SOTA 方法进行了对比, 取得了具有竞争力的检测效果和结果。

参考文献:

[1] 郭毅锋, 吴帝浩, 魏青民. 基于深度学习的点云三维目标检测方法综述 [J]. 计算机应用研究, 2023, 40(1): 20–27.
GUO Yifeng, WU Dihao, WEI Qingmin. Overview of single-sensor and multi-sensor point cloud 3D target detection methods[J]. Application research of computers, 2023, 40(1): 20–27.

[2] CHARLES R Q, HAO Su, MO Kaichun, et al. PointNet: deep learning on point sets for 3D classification and segmentation[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017: 77–85.

[3] CHARLES R Q, YI Li, SU Hao, 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: ACM, 2017: 5105–5114.

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

[5] SHI Shaoshuai, WANG Xiaogang, LI Hongsheng. PointRCNN: 3D object proposal generation and detection from point cloud[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 770–779.

[6] YANG Zetong, SUN Yanan, LIU Shu, et al. 3DSSD: point-based 3D single stage object detector[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020: 11037–11045.

[7] HE Chenhang, ZENG Hui, HUANG Jianqiang, et al. Structure aware single-stage 3D object detection from point cloud[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020: 11870–11879.

[8] ZHENG Wu, TANG Weiliang, JIANG Li, et al. SE-SSD: self-ensembling single-stage object detector from point cloud[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021: 14489–4498.

[9] XU Qiangeng, ZHONG Yiqi, NEUMANN U. Behind the curtain: learning occluded shapes for 3D object detection[J]. Proceedings of the AAAI conference on artificial intelligence, 2022, 36(3): 2893–2901.

[10] SHI Shaoshuai, GUO Chaoxu, JIANG Li, et al. PV-RCNN: point-voxel feature set abstraction for 3D object detection[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020: 10526–10535.

[11] DENG Jiajun, SHI Shaoshuai, LI Peiwei, et al. Voxel R-CNN: towards high performance voxel-based 3D object detection[J]. Proceedings of the AAAI conference on artificial intelligence, 2021, 35(2): 1201–1209.

[12] MAO Jiageng, NIU Minzhe, BAI Haoyue, et al. Pyramid R-CNN: towards better performance and adaptability for 3D object detection[C]//2021 IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021: 2703–2712.

[13] SHENGA Hualian, CAI Sijia, LIU Yuan, et al. Improving 3D object detection with channel-wise transformer[C]//2021 IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021: 2723–2732.

[14] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[EB/OL]. (2017–06–12)[2022–09–06]. http://arxiv.org/abs/1706.03762.

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

[16] 王亚东, 田永林, 李国强, 等. 基于卷积神经网络的三维目标检测研究综述 [J]. 模式识别与人工智能, 2021, 34(12): 1103–1119.
WANG Yadong, TIAN Yonglin, LI Guoqiang, et al. 3D object detection based on convolutional neural networks: a survey[J]. Pattern recognition and artificial intelligence, 2021, 34(12): 1103–1119.

[17] ZHANG Yifan, HU Qingyong, XU Guoquan, et al. Not all points are equal: learning highly efficient point-based detectors for 3D LiDAR point clouds[EB/OL]. (2022–03–21) [2022–09–06]. http://arxiv.org/abs/2203.11139.

[18] PAN Xuran, XIA Zhuofan, SONG Shiji, et al. 3D object detection with pointformer[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021: 7459–7468.

[19] YAN Yan, MAO Yuxing, LI Bo. SECOND: sparsely embedded convolutional detection[J]. Sensors, 2018, 18(10): 3337.

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

[21] ZHENG Wu, TANG Weiliang, CHEN Sijin, et al. CIA-SSD: confident IoU-aware single-stage object detector from point cloud[J]. Proceedings of the AAAI conference on artificial intelligence, 2021, 35(4): 3555–3562.

[22] KUANG Hongwu, WANG Bei, AN Jianping, et al. Voxel-FPN: multi-scale voxel feature aggregation for 3D object detection from LIDAR point clouds[J]. Sensors, 2020, 20(3): 704.

[23] 李文举, 储王慧, 崔柳, 等. 结合图采样和图注意力的 3D 目标检测方法 [J]. 计算机工程与应用, 2023, 59(9): 237–244.
LI Wenju, CHU Wanghui, CUI Liu, et al. 3D object detection method combining on graph sampling and graph attention[J]. Computer engineering and applications, 2023, 59(9): 237–244.

[24] GLOROT X, BENGIO Y. Understanding the difficulty of training deep feedforward neural networks[C]//Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics. Sardinia: JMLR Workshop and Conference Proceedings, 2010: 249–256.

[25] LIN T Y, GOYAL P, GIRSHICK R, et al. Focal loss for dense object detection[C]//2017 IEEE International Conference on Computer Vision. Venice: IEEE, 2017: 2999–3007.

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

[27] CHEN Xiaozhi, KUNDU K, ZHU Yukun, et al. 3D object proposals for accurate object class detection[C]//Proceedings of the 28th International Conference on Neural Information Processing Systems-Volume 1. Montreal: ACM, 2015: 424–432.

[28] CHEN Xiaozhi, MA Huimin, WAN Ji, et al. Multi-view 3D object detection network for autonomous driving[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017: 6526–6534.

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

[30] QI C R, LIU Wei, WU Chenxia, et al. Frustum PointNets for 3D object detection from RGB-D data[C]//2018

IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 918–927.

[31] LIANG Ming, YANG Bin, WANG Shenlong, et al. Deep continuous fusion for multi-sensor 3D object detection[C]// European Conference on Computer Vision. Cham: Springer, 2018: 663–678.

[32] ZHAO Xin, LIU Zhe, HU Ruolan, et al. 3D object detection using scale invariant and feature reweighting networks[J]. Proceedings of the AAAI conference on artificial intelligence, 2019, 33(1): 9267–9274.

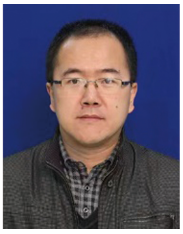
[33] LIANG Ming, YANG Bin, CHEN Yun, et al. Multi-task multi-sensor fusion for 3D object detection[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 7337–7345.

[34] 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]//European Conference on Computer Vision. Cham: Springer, 2020: 720–736.

[35] YANG Zetong, SUN Yanan, LIU Shu, et al. STD: sparse-to-dense 3D object detector for point cloud[C]//2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019: 1951–1960.

[36] LEHNER J, MITTERECKER A, ADLER T, et al. Patch refinement: localized 3D object detection[EB/OL]. (2019–10–09)[2022–09–06]. http://arxiv.org/abs/1910.04093.

作者简介:



鲁斌, 教授, 博士生导师, 博士, CCF 高级会员, 主要研究方向为智能计算与计算机视觉, 综合能源系统与大数据分析。主持、参与国家、省部级科技项目 7 项, 主持企事业委托项目 18 项, 作为第一完成人获全国商业科技进步二等奖 1 项, 作为校内第一完成人获河北省科技进步奖 3 项、市级科技进步奖 4 项, 获专利授权 10 项, 发表学术论文 68 篇, 出版专著 3 部。 E-mail: lubin@ncepu.edu.cn。

孙洋, 博士研究生, 主要研究方向为机器学习、计算机视觉。E-mail: bless2016@163.com。

杨振宇, 博士研究生, 主要研究方向为机器学习、计算机视觉。E-mail: yangzhenyu536@163.com。