LiDar点云指导下特征分布趋同与语义关联的3D目标检测

郑 锦[1,2],蒋博韬[1],彭 微[1],王 森[1]
(1. 北京航空航天大学计算机学院,北京 100191;2. 虚拟现实技术与系统全国重点实验室,北京 100191)

摘 要: 针对现有基于伪点云的3D目标检测算法精度远低于基于真实激光雷达(Light Detection and ranging,LiDar)点云的3D目标检测,本文研究伪点云重构,并提出适合伪点云的3D目标检测网络.考虑到由图像深度转换得到的伪点云稠密且随深度增大逐渐稀疏,本文提出深度相关伪点云稀疏化方法,在减少后续计算量的同时保留中远距离更多的有效伪点云,实现伪点云重构.本文提出LiDar点云指导下特征分布趋同与语义关联的3D目标检测网络,在网络训练时引入LiDar点云分支来指导伪点云目标特征的生成,使生成的伪点云特征分布趋同于LiDar点云特征分布,从而降低数据源不一致造成的检测性能损失;针对RPN(Region Proposal Network)网络获取的3D候选框内的伪点云间语义关联不足的问题,设计注意力感知模块,在伪点云特征表示中通过注意力机制嵌入点间的语义关联关系,提升3D目标检测精度.在KITTI 3D目标检测数据集上的实验结果表明:现有的3D目标检测网络采用重构后的伪点云,检测精度提升了2.61%;提出的特征分布趋同与语义关联的3D目标检测网络,将基于伪点云的3D目标检测精度再提升0.57%,相比其他优秀的3D目标检测方法在检测精度上也有提升.

关键词: 3D目标检测;伪点云;语义关联;分布趋同;注意力感知
基金项目: 国家自然科学基金(No.61876014)
中图分类号: TP391.4 **文献标识码:** A **文章编号:** 0372-2112(2024)05-1700-16
电子学报URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20221141

3D Object Detection Based on Feature Distribution Convergence Guided by LiDar Point Cloud and Semantic Association

ZHENG Jin[1,2], JIANG Bo-tao[1], PENG Wei[1], WANG Sen[1]
(1. *School of Computer Science and Engineering*, *Beihang University*, *Beijing* 100191, *China*;
2. *State Key Laboratory of Virtual Reality Technology and Systems*, *Beijing* 100191, *China*)

Abstract: In view of the accuracy of existing 3D object detection algorithms based on Pseudo-LiDar is far lower than that based on real LiDAR (Light Detection and ranging), this paper studies the reconstruction of Pseudo-LiDar and proposes a 3D object detection algorithm suitable for Pseudo-LiDar. Considering that the Pseudo-LiDAR obtained by image depth is dense and gradually sparse along the increase of depth, a depth related Pseudo-LiDAR sparsification method is proposed to reduce the subsequent calculation amount while retaining more useful Pseudo-LiDAR in the middle and long distance, so as to realize the reconstruction of Pseudo-LiDAR. Furthermore, a 3D object detection algorithm based on object feature distribution convergence under the guidance of LiDar point cloud and semantic association is proposed. During network training, a laser point cloud branch is introduced to guide the generation of Pseudo-LiDAR object features, so that the generated Pseudo-LiDar object feature distribution converges to the feature distribution of laser point cloud object, thereby correcting the detection error caused by the difference between the two data sources. Aiming at the insufficient semantic association between Pseudo-LiDar in the 3D candidate bounding-box obtained by RPN (Region Proposal Network) network, an attention perception module is designed to embed the semantic association between points through the attention mechanism in the feature representation of Pseudo-LiDar, so as to improve the accuracy of 3D object detection. The experimental results on KITTI 3D object detection dataset show when the existing 3D object detection network adopts the reconstructed Pseudo-LiDar, the detection accuracy is improved by 2.61%. Furthermore, the proposed 3D object detection network with the fea-

收稿日期:2022-10-11;修回日期:2022-12-23;责任编辑:朱梅玉

ture distribution convergence and semantic association improves the accuracy by 0.57%. Compared with other excellent methods, it also improves the detection accuracy.

Key words: 3D object detection; Pseudo-LiDar; semantic association; distribution convergence; attention perception
Foundation Item(s): National Natural Science Foundation of China (No.61876014)

1 引言

区别于2D目标检测只能得到目标在图像平面的位置信息,不能提供目标的空间位置和结构信息,3D目标检测在感知目标的同时,能进一步预测物体在三维空间中的坐标位置、三维尺寸、偏航角度等信息,而这些信息是路径规划、避碰等驾驶任务的重要依据. 随着自动驾驶等应用的普及,3D目标检测正发挥着不可或缺的作用. 在自动驾驶中一般通过相机、LiDar(Light Detection and ranging)等感知设备获取周围场景及目标的信息. 其中,相机采集周围环境的2D图像,包含丰富的RGB纹理及语义;LiDar采集周围环境的点云数据,包含精确的位置和丰富的3D几何结构. 因此,基于点云来实现3D空间感知更直接,基于LiDar点云的3D目标检测能够获得更高精度. 但LiDar造价高,采用成本更低的相机作为替代方案、基于图像进行3D目标检测是研究热点.

基于图像的3D目标检测方法一般分为两类. 一类是直接基于图像的3D目标检测,主要采用CNN(Convolutional Neural Network)提取RGB图像特征并得到目标候选框,进而结合先验知识、几何约束、模板匹配、语义信息等得到精确的3D检测框. 另一类是基于深度图像转伪点云的3D目标检测,根据相机参数将场景深度投影回3D空间得到伪点云,然后基于伪点云完成3D目标检测. 由于基于LiDar点云的3D目标检测算法发展迅速,模型的检测精度也较高,因此基于伪点云的3D目标检测方法应用较广,这类方法在获取伪点云后,通常采用现有的、比较成熟的LiDar点云3D目标检测网络.

本文针对现有基于伪点云的3D目标检测算法精度远低于基于真实LiDar点云的3D目标检测算法的问题,分析了伪点云和LiDar点云在数量、分布等方面的差异,以及伪点云3D目标检测并不能有效利用LiDar点云潜在信息的深层次原因,提出伪点云重构方法,以及适合伪点云的3D目标检测网络. 本文主要贡献如下.

(1)伪点云重构. 由于伪点云与真实LiDar点云存在数量差异,伪点云稠密且冗余,且不同远近距离下数量差异还存在差别,设计深度相关的伪点云稀疏化处理方法,在去除冗余的同时保留中远距离下有限的目标伪点云,为中远距离目标的3D检测奠定基础.

(2)基于伪点云的3D目标检测网络设计. 由于伪点云与真实LiDar点云存在分布差异,现有的基于LiDar点云的3D目标检测网络并不完全适用于伪点云. 为此,本文提出LiDar点云指导下特征分布趋同与语义关联的3D目标检测网络. 一方面针对伪点云与LiDar点云分布不一致的特点,在网络训练时引入LiDar点云指导分支,采用LiDar点云指导伪点云目标特征的生成,使生成的伪点云目标特征的分布趋同于LiDar点云目标特征的分布,降低数据源不一致造成的检测性能损失;另一方面,引入注意力感知模块,在伪点云特征中嵌入语义关联关系,增强伪点云特征表示,有效提升基于伪点云的3D目标检测精度.

(3)在KITTI 3D目标检测验证集上的实验结果表明,现有的3D目标检测网络采用重构后的伪点云,检测精度提升了2.61%,提出的特征分布趋同与语义关联的3D目标检测网络将伪点云上的检测精度再提升0.57%. 这两项工作可应用于现有基于伪点云的3D目标检测网络,提升检测精度.

本文方法针对实际应用时只有图像数据、没有LiDar数据提出,这意味着当实际应用时,本文提出的网络只需要输入图像数据转换得到的伪点云,无需真实的LiDar点云数据. 而在训练阶段,由于采用的KITTI 3D目标检测数据集本身自带左视图、右视图、激光LiDar点云、GT真实标注框、相机标定参数,可以根据左视图、右视图、相机标定参数生成场景深度,进而生成伪点云. 因此,在训练阶段可以同时利用LiDar点云、伪点云数据,利用LiDar点云指导伪点云目标特征的生成.

2 相关工作

目前,主流的3D目标检测算法分为基于LiDar点云的3D目标检测和基于图像的3D目标检测,本文研究基于图像的3D目标检测,其又分为两类:直接基于图像的方法、基于图像深度转伪点云的方法. 此外,考虑到上下文语义关联能够增强3D目标的空间特征,并能对3D目标检测提供隐性的指导,本文还分析了现有语义关联的相关方法.

2.1 直接基于图像的3D目标检测

直接基于图像的方法包括基于双目图像和单目图像两类. 基于双目图像的代表性算法有3DOP(3D Object Proposals)[1],TLNet(Triangulation Learning Network)[2],Stereo-RCNN[3]等. 3DOP[1]是基于双目图像进行3D目标检测的开创性工作,它根据双目图像来估计3D信

息,将深度相关特征(如点密度、可见性、目标大小先验等)编码成能量函数,并最小化能量函数来生成3D目标候选区域,进而采用Fast-RCNN来联合回归精细目标框. TLNet[2]利用3D锚来显式构建双目图像中感兴趣区域(Region Of Interest,ROI)之间的实例对应关系以实现3D目标检测,引入信道重加权策略增强表示特征并减弱噪声信号以促进学习过程,提高检测精度. Stereo-RCNN[3]扩展了Faster-RCNN,将网络架构由单路变成双路,利用双目图像稠密的语义信息和物体本身的几何信息,同时检测和关联双目图像中的对象,并利用关键点和2D几何约束来估计3D目标位置. Stereo-RCNN中的对齐策略可显著提高深度估计的准确性,但严重依赖两个边界关键点的对应区域. 若目标存在遮挡,该目标会被丢弃而无法被匹配. 此外,近年提出的Disp R-CNN[4]和ZoomNet[5]结合额外的实例分割掩码和部件位置图来提高检测精度. 然而,它们对2D检测算法的依赖限制了最终3D检测的性能. 此外,3D几何信息也未得到充分利用. DSGN(Deep Stereo Geometry Network)[6]直接采用CNN网络提取特征,并构建可微体结构来编码场景中的隐式3D几何结构,从而实现端到端双目3D目标检测. 但是,DSGN需要从双目图像对创建一个中间的平面扫描体,以学习相机的立体约束,然后将其转换为3D空间中的3D体以提取3D几何信息,计算量大、模型耗时. 最近,PLUME[7]直接在3D空间构建几何体,这种几何体在深度估计和3D检测网络之间共享,从而加速计算.

基于单目图像的3D目标检测包括直接回归策略、基于模板匹配、基于几何约束策略等方法. 例如,直接回归策略方法中,Mono3D(Monolithic 3-D)[8]直接生成稠密的3D候选区域,随后通过手工设计的特征对候选区域打分,最后经过NMS(Non Maximum Suppression)得到检测结果. SS3D(Single Shot 3D)[9]采用了类似CenterNet的检测框架,首先得到中心位置,然后回归2D和3D的检测框顶点坐标得到3D框. 基于模板匹配的方法中,Deep MANTA[10]根据模板相似度搜索车辆模板库中最佳匹配的CAD模型,接着,将2D点及CAD模型的3D点通过EPnP[11]算法解算出3D框. Mono3D++[12]通过预测14个2D关键点得到可变形线框模型,结合先验信息联合能量最小化重投影误差得到最终的车辆形状和姿态. 基于几何约束策略的方法中,Deep3DBox[13]通过假定图像中2D框紧密包围目标3D框在图像中的投影来得到2D/3D紧密性约束,然后根据该约束构建方程,求解中心点位置得到3D框. Fang等人[14]使用中心投影结合相似三角形约束,通过闭合形式的解获得初始3D框,并使用视角分类确定最终位置. Ground-aware[15]预测与高度相关的地面中心点投影偏差来提取地面特征,最后融合目标特征回归得到深度残差以实现3D检测.

总的来看,直接基于图像的方法往往依赖CNN网络提取图像特征以实现2D目标检测,并通过构建几何约束等模型得到目标的深度信息,计算复杂且准确性不足. 同时,3D检测结果极大程度上依赖2D检测的精度. 因此,现有的直接基于图像的方法整体检测精度不高,尤其是基于单目图像的方法. 针对车辆目标的"Moderate"类,目前SOTA方法3D AP精度只有17%左右.

2.2 基于图像深度转伪点云的3D目标检测

基于图像的3D目标检测方法相比基于LiDar点云的3D目标检测,检测精度存在明显差距. 大部分学者将这种差距归因于基于图像的目标深度预测不够准确. Wang等人[16]在CVPR 2019上首次提出了伪点云(Pseudo-LiDar)数据表示形式,认为造成检测性能差距的最大原因不是数据的质量,而是其表示形式. 目前,大部分网络直接将深度图作为RGB图像的附加通道,这种表示方式经过CNN处理之后无法准确表达图像的深度信息,会造成一定程度的深度失真,尤其是对于远距离目标,失真更加严重. 作者建议将基于图像的深度图转换为Pseudo-LiDar表示形式,从而模拟LiDar信号. 点云的形式能更真实地表达目标在3D空间中的几何结构. 而且通过这种表示方式,可以应用已有的基于LiDar的检测算法完成3D目标的检测.

Pseudo-LiDar的提出为基于图像深度的3D目标检测提供了新思路,也在一定程度上缩小了基于图像的3D目标检测与基于LiDar点云的3D目标检测算法的差距,但是该方法也存在一些不足. 例如,基于LiDar的3D检测网络很大程度上依赖点云位置的准确性,而伪点云位置通过深度估计结果转换而来,现有深度估计网络对远距离的物体以及物体的边缘估计准确性仍然不高,越远的物体深度估计误差越大,从而导致伪点云的位置发生偏移,影响了后续3D检测网络的精度. 为此,You等人[17]改进了Pseudo-LiDar并提出了Pseudo-LiDar++,利用成本较低但稀疏的LiDar传感器采集少量的点云数据来监督深度估计网络,从而消除深度估计偏差. Pseudo-LiDar++同时调整了双目深度估计网络架构及损失函数,使距离遥远的物体深度估计更加精确,得到位置更精确的伪点云数据.

另外,深度转化而来的伪点云过于稠密,为了使伪点云和真实的LiDar点云差异更小,需要进行伪点云的稀疏化,通常有两种方式:一是按照空间稀疏,将三维空间看作一个立方体,然后将伪点云根据坐标放入这个三维空间中,若多个点落入同一位置,则任意取其中的某个点,稀疏后的点云为该三维立体空间中的点云;

二是按照角度稀疏,即模仿64线LiDar的点云采集方法,对采集的范围按角度进行划分,竖直方向上划分为64个角度,水平方向上划分为512个角度,将落入每个位置的伪点云保留下来,同理,多个点落入同一位置则任选一点. 现有方法一般采用第二种稀疏化方式. RefinedMPL(Refined Monocular PseudoLiDAR)[18]认为从图像转化而来的Pseudo-LiDar包含了太多背景点信息,因此通过前景分离、距离分层采样等方式进行伪点云数据的稀疏化.

现有基于图像深度转伪点云的3D目标检测首先根据图像深度图生成伪点云,对伪点云进行稀疏化等处理,然后采用比较成熟的基于LiDar点云的3D目标检测算法,将网络的输入由LiDar点云替换成伪点云,最后基于伪点云完成3D目标检测. 然而,直接采用LiDar点云3D目标检测网络,这种设计没有充分考虑伪点云和真实LiDar点云在数量、分布上的差异,没有有效利用伪点云中的目标结构特点,存在一定的局限性.

2.3　上下文语义关联

现有的3D目标检测会利用一些辅助任务来增强3D目标的空间特征,从而为3D目标检测提供一些隐性指导. 常用的辅助任务包括语义分割、IoU(Intersection over Union)分支预测、目标形状补全、部件识别等. 其中,语义分割利用获取的上下文语义关联关系获取额外的信息,例如:前景分割可以提供目标的隐含位置;语义上下文形成物体所处语境的上下文关联关系,结合类似于“飞机更有可能出现在天空中,而非水面上[19]”这样的“目标-背景”共存关系增强目标的空间特征表示;语义分割可以作为预处理方法,过滤背景样本,提升3D目标检测效率.

利用语义信息指导目标检测时可以将语义分割网络作为一个特定的特征提取器,将其加入目标检测框架中,如加入语义分割感知和编码的CNN模型[20]和用语义分割网络增强Faster R-CNN模型[21]. 另一种方案是类似Stuffnet[22]引入额外的语义分割分支,并将语义分割训练损失结合到目标检测框架中,利用多任务损失函数进行辅助学习.

在3D目标检测涉及的立体匹配中,当前基于CNN的立体匹配方法的关键之一就是如何有效利用上下文,一些研究试图合并语义信息来细化成本量或视差图. 例如,Displets[23]利用三维车辆建模获得的对象信息来解决立体匹配中的歧义,从而得到更精确的视差图. 基于图像的3D目标检测方法中,由于RGB图像无法像点云那样提供准确的三维空间信息,因此一些方法深入探究了立体图像的上下文关联关系. 例如,3DOP[1]提出结合上下文并使用多任务损失来联合回归边界框坐标和目标方向,进行3D目标检测;AVOD[24]模型输入RGB图像和鸟瞰图,采用特征金字塔、编解码器,在保证最终的特征图相对于输入的全解析的同时,结合底层和高层的语义信息,提高小目标的检测性能;PointPainting[25]利用语义分割网络对图像中各像素进行归类,然后根据图像与点云之间的变换关系,将语义分割结果投影到点云上,以投影后的点云作为原始数据在各种3D目标检测网络上进行物体识别. 语义关联关系的引入已经被证明能够提升3D目标检测的精度.

在引入上下文关联关系时,注意力机制旨在关注与目标更相关的特征来模仿人类的视觉系统,而不引入与目标无关的上下文. 注意力机制通常估计注意力权值图,进而对原始的特征图进行重新加权. 在计算机视觉任务中注意力机制包括通道注意力[26]、空间注意力[27]和两者的混合注意力,用来获取特征在通道维度和空间维度上的相关性. 经过注意力机制得到的特征具有更大的感受野,也包含丰富的上下文特征关联关系. 此外,点云网络倾向于利用自注意力[28]结构. 它可以估计长距离依赖关系,而不考虑两者之间的特定顺序,从而计算点间的关系[29]或者通道间的关系[30].

3　本文方法

点云的表示形式能够很好地表征三维空间物体的结构,因此将图像根据其深度图转换到三维空间得到伪点云,再采用基于LiDar点云的3D目标检测网络进行基于伪点云的3D目标检测是目前常用的思路,其检测精度相比直接基于图像的3D检测也有一定的优势. 然而,生成的伪点云和真实LiDar点云在数量、分布上存在明显差异,而基于LiDar点云的3D目标检测网络也并不完全适用于伪点云. 本文方法的初衷是解决自动驾驶中低成本3D目标的准确检测问题,思路来源就是要缩小伪点云和真实LiDar点云的差异,以及设计适合伪点云的3D目标检测网络. 因此,本文对生成的伪点云进行特性分析,根据伪点云的特点设计重构方法,缩小伪点云与真实LiDar点云的差异,然后将处理后的伪点云作为3D目标检测网络的输入,同时引入LiDar点云分支指导伪点云目标特征的生成,使伪点云和真实LiDar点云目标特征分布趋同,并在伪点云特征表示中嵌入语义关联关系,提升3D目标检测网络在伪点云上的检测精度.

3.1　伪点云重构

3.1.1　伪点云的特性分析

Wang等人[16]提出的Pseudo-LiDar伪点云数据表示形式,显式地表达了空间位置信息. 其中,视差图结合相机参数转换为深度图,进而投影回3D空间得到(x,y,z)点的坐标表示形式. 将伪点云的反射率值都设为1,变成和LiDar点云维度一致,转换后的伪点云数据表示形

式为(x,y,z,l).

根据上述方法生成的伪点云,虽然在数据表示形式上和LiDar点云一致,但由于数据来源不同,伪点云由像素根据深度信息投影回3D空间得到,而LiDar点云由LiDar扫描得到,两者在数量、分布等方面仍存在极大差异. 图1展示了KITTI 3D数据集下一组双目图像的左视图、右视图以及对应的激光点云、伪点云的可视化结果. 深度图转换生成的伪点云和真实LiDar点云在数量、分布上存在明显差异. 本文对伪点云和LiDar点云的特性进行分析比较,找出二者差异并重构伪点云,使伪点云在保留自身优势的基础上更趋于真实的LiDar点云,从而更加适用于基于LiDar点云的3D目标检测网络.

首先,从数量上分析. 在KITTI数据集[31]采集平台的64线LiDar装置下,获取的LiDar点云大约为12万个,考虑到LiDar点云呈现为360°环境信息,而相机采集前视图且3D检测任务中只检测前视图中的目标,因此只保留投影位置在RGB图像内的LiDar点云,数量大约为2万个. 而伪点云与RGB图像中的像素一一对应,每个像素都能投射为3D空间中的一个伪点云,以KITTI数据集中尺寸为1 242 × 375像素的图像为例,对应生成的伪点云个数为465 750个,只保留正常视野范围内的点后剩余的伪点云大约为32万个. 因此,伪点

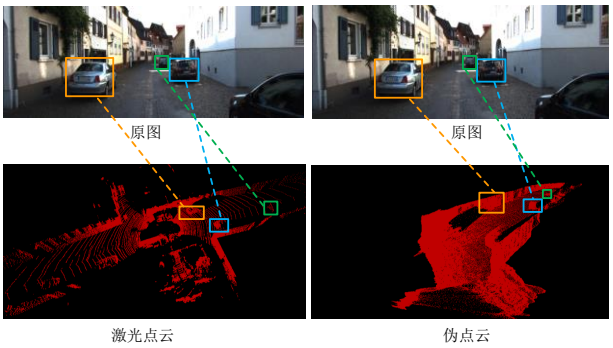


图1 同一场景下真实LiDar点云和生成的伪点云

云个数是LiDar点云的16倍,伪点云的数量远远多于LiDar点云,因此伪点云比LiDar点云更加稠密. 为了使伪点云数量与LiDar点云更吻合,需要对伪点云进行稀疏化处理.

其次,从伪点云的分布分析. 统计发现KITTI数据集中95%的伪点云对应的是背景点,如天空、路面、树木、房子等,前景目标点占比只有5%,这也说明用于3D目标检测的伪点云可以稀疏化. 进而,本文统计了KITTI 3D数据集中伪点云与LiDar点云数量随着深度变化的分布情况,将0~80 m的深度范围均匀划分为8个区间,统计每一个深度区间内伪点云以及激光点云的数目,从而得到两者在不同深度区间的数量分布情况,统计结果如图2所示.

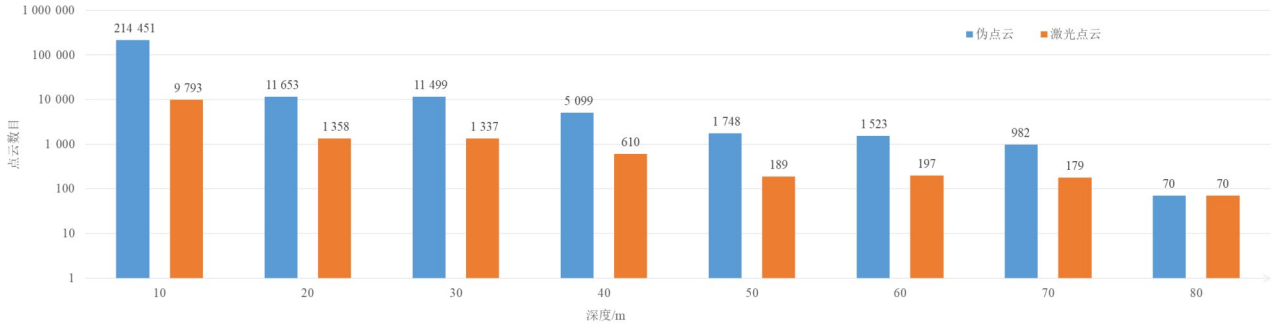


图2 不同深度点云数目统计图

由以上统计结果可以看出,伪点云总体而言是稠密的,存在大量的冗余信息. 不同深度下伪点云和LiDar点云的数量变化有差异,即随着距离由近及远,两者的数量都逐渐减少,但不同深度下减少比例有所不同:在20 m以内,伪点云数目是激光点云数目的10倍以上;在20~70 m内,伪点云数目是激光点云数目的5~10倍;在70 m以外,二者数目相近. 同时,可观察出伪点云大部分集中在40 m以内,占总数目的99%左右,冗余严重. 我们也注意到,LiDar点云在40 m外的点很少,并不能很好地支持中远距离的目标检测. 相比而言,伪点云能保留部分远距离目标点,可以更好地表征目标并提高3D目标检测网络的性能. 因此,稀疏化应更多针对近距离的伪点云,同时为了提升中远距离目标的检测性能,需要尽量保留中远距离目标对应的伪点云,有效利用伪点云的优势.

3.1.2 深度相关的伪点云稀疏化

现有直接采用空间稀疏和角度稀疏对伪点云稀疏化的方法,对不同距离的伪点云,稀疏程度是一致的,无论是近距离还是远距离的目标都会被稀疏,从而导致远距离目标剩余点云数量过少而出现漏检.

本文提出深度相关的伪点云稀疏化方法,根据KITTI 3D数据集中伪点云与LiDar点云数量随着深度

变化的分布情况统计,设置0~20 m,20~40 m,40~70 m,70 m外4个区间,分别对应近距离、中距离、中远距离、远距离,进行随深度增大而稀疏程度递减的伪点云稀疏化,对应的视场角划分如式(1)所示:

$$\mathrm{Pse}_{\mathrm{sparify}}=\begin{cases}D=D_1,\ W=W_1, & d\leqslant 20\\ D=D_2,\ W=W_2, & 20<d\leqslant 40\\ D=D_3,\ W=W_3, & 40<d\leqslant 70\end{cases}\quad(1)$$

其中,类比旋转扫描的多线激光雷达,D表示垂直视场角的划分数目,等同激光雷达的线数;W表示旋转形成的水平视场角的划分数目,等同激光雷达旋转扫描时进行收发探测的次数. 当$d\leqslant 70$时,根据距离远近设置不同的稀疏化水平,距离越远,划分的视场角数目越多,稀疏化程度越低;当$d>70$时不再进行稀疏化,尽量保留目标点云. 这一设置是根据KITTI 3D数据集得到的,KiTTI 3D数据集包含了大量训练样本,因此在类似KITTI这样的街景数据集中具有普适性. D_1、D_2、D_3,W_1、W_2、W_3的具体取值在实验部分进行说明.

3.2　LiDar点云指导的3D目标检测

将现有基于LiDar点云的3D目标检测网络直接用于伪点云时,检测性能急剧下降,原因主要有两个:数据本身的问题、网络学习能力不够. 伪点云重构从数据上缩小了伪点云与LiDar点云在数量、空间远近距离分布上的差异,但是针对两者学习到的目标表征依然存在差异. 本文研究基于伪点云的3D目标检测网络,利用LiDar点云指导挖掘伪点云的潜在信息,提升网络的学习能力,从而有效提高3D目标检测的精度.

3.2.1　网络整体结构

以现有基于LiDar点云的CT3D[32]为基础设计本文的3D目标检测网络. CT3D主要包括以下模块:RPN网络、Transformer结构、检测头. 其中,RPN网络提取3D候选框,Transformer结构编解码候选框内的点云,检测头回归置信度和3D框. 当CT3D用于伪点云3D目标检测时,将LiDar点云输入替换为伪点云.

考虑到伪点云与LiDar点云二者的差异,本文在训练阶段引入LiDar点云指导伪点云的目标特征生成,使二者特征分布趋于一致. 此外,为了融合更丰富的语义关联关系,引入注意力感知模块,从而提升网络在伪点云上的检测性能. 改进后的网络结构如图3所示,在原有CT3D的基础上增加了一个LiDar点云指导分支以及一个注意力感知模块,并引入特征分布一致性损耗函数KLD Loss来监督伪点云目标特征生成. 在训练阶段,网络的输入既有LiDar点云,也有伪点云;在验证阶段,网络的输入只有伪点云,去除了LiDar点云分支. 现有LiDar点云、伪点云融合的方法,往往利用伪点云得到3D候选框后再转换到三维视锥中回归3D检测框,而本文方法利用LiDar分支指导伪点云的目标特征分布,因此可以在验证阶段完全去除LiDar点云分支,仅需要伪点云作为输入,对数据要求更低.

针对伪点云分支,首先伪点云经过RPN网络提取3D候选框,CT3D中直接采用SECOND(Sparsely Embedded CONvolutional Detection)[33]网络架构作为RPN网络模块,在提取3D候选框的同时也能得到分类信息. 根据RPN网络得到的3D候选框,可分别获取其框内对应目标的伪点云和LiDar点云. 伪点云分支中目标对应的伪点云先经过注意力感知模块,从而在伪点云特征表示中嵌入语义关联关系,再经过伪点云目标特征编码器生成伪点云目标特征. 伪点云目标特征编码器分为编码器和解码器两部分,其主要功能是对由RPN网络得到的proposal候选框以及候选框内的点云数据进行编解码,每个候选框对应得到一个固定维度为$1\times D$的特征向量,当3D候选框的数目为N时,得到$N\times D$维的特征向量$\boldsymbol{f}_{\mathrm{pse}}$来表征整幅图的目标点云.

针对LiDar点云指导分支,根据RPN网络得到的3D候选框,可获取其框内对应目标的LiDar点云,这些LiDar点云经过LiDar点云目标特征编码器后被编码为一个$1\times D$维的向量,N个候选框得到$N\times D$维的特征向量$\boldsymbol{f}_{\mathrm{lidar}}$,用来指导伪点云目标特征$\boldsymbol{f}_{\mathrm{pse}}$的生成.

在LiDar点云指导下生成的伪点云目标特征$\boldsymbol{f}_{\mathrm{pse}}$被送入到两个FFN(Feed-Forward Network)层,分别得到置信度预测分数c和预测框的回归结果(x,y,z,w,l,h,θ).

3.2.2　LiDar点云指导分支

引入的LiDar点云指导分支生成LiDar点云目标的特征表示,然后用该特征去指导伪点云目标特征的生成,使伪点云目标特征表示的分布趋同于LiDar点云目标特征表示的分布,从而降低数据源不一致造成的检测性能损失. LiDar点云指导分支子网络的结构如图3所示. 图3中的LiDar点云目标特征编码器与伪点云目标特征编码器在结构上一致,只不过前者是对候选目标框内的LiDar点云进行编码,而后者对候选目标框内的伪点云进行编码.

在该分支中,首先根据RPN网络得到的3D候选框,将其映射到原始LiDar点云数据中获取其内部的点云数据. 考虑到候选框可能存在位置偏差,对候选框区域进行一定比例的扩大,由立方体扩大为圆柱体,并在高度上不予限制,底面圆的半径为

$$\mathrm{radius}=\alpha\sqrt{\left(\frac{w}{2}\right)^2+\left(\frac{l}{2}\right)^2}$$

其中,α为一个超参数;w,l分别表示候选框的宽度和长度,从而包含更多目标点云,完成候选框内LiDar点云的选取工作. 随后将选取的LiDar点云送入LiDar点云目标特征编码器,其主要包含编码器和解码器两部分. 编码器对Lidar点云进行点编码和自注意力编码,点编

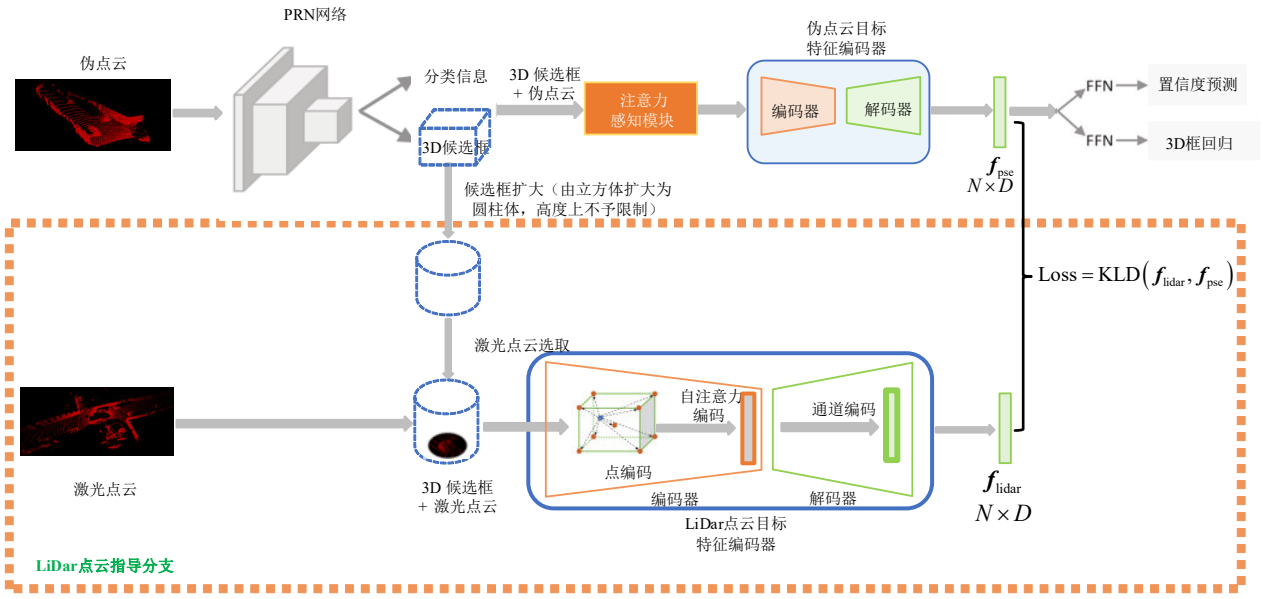


图3 特征分布趋同与语义关联的3D目标检测网络结构

码采用每个采样点相对候选框8个角点的相对坐标来生成点特征，自注意力编码中采用多头自注意力机制对每个候选框的特征 $\boldsymbol{X}=\left[\boldsymbol{f}_1^{\mathrm{T}}\boldsymbol{f}_2^{\mathrm{T}}\cdots\boldsymbol{f}_N^{\mathrm{T}}\right]\in\mathbb{R}^{N\times D}$ 编码丰富的上下文关系；解码器对编码后的特征进行通道解码，针对点云中不同通道之间存在很强的结合关系，采用通道权重解码器，最后得到LiDar点云目标特征向量 $\boldsymbol{f}_{\mathrm{lidar}}$.

LiDar点云指导分支和伪点云分支中两个编码器输出的特征表示尺寸一致，均为 $N\times D$，在二者之间采用损失函数进行约束，利用LiDar点云生成的目标特征 $\boldsymbol{f}_{\mathrm{lidar}}=[y_1\,y_2\,\cdots\,y_n]$ 来指导伪点云的目标特征 $\boldsymbol{f}_{\mathrm{pse}}=[\hat{y}_1\,\hat{y}_2\,\cdots\,\hat{y}_n]$ 学习．本文选取的损失函数为KLD Loss，计算公式如下：

$$L_{\mathrm{KLD}}=\mathrm{KLD}\left(\boldsymbol{f}_{\mathrm{lidar}},\boldsymbol{f}_{\mathrm{pse}}\right)=\sum_{j=1}^{n}\hat{y}_j\log\frac{\hat{y}_j}{y_j}\tag{2}$$

针对整个网络的损失函数，除了上述特征一致性损失KLD Loss之外，还包括RPN网络损失、置信度预测损失、检测框的回归损失．因此，网络总损失函数 L_{total} 由RPN网络损失 L_{RPN}、置信度预测损失 L_{conf}、检测框回归损失 L_{reg}、特征一致性损失 L_{KLD} 这4部分构成，其公式表示如下：

$$L_{\mathrm{total}}=\alpha_1L_{\mathrm{RPN}}+\alpha_2L_{\mathrm{conf}}+\alpha_3L_{\mathrm{reg}}+\alpha_4L_{\mathrm{KLD}}\tag{3}$$

其中，$\alpha_1,\alpha_2,\alpha_3,\alpha_4$ 分别为各项损失的权重系数．

采用交叉熵[34]来计算置信度预测损失 L_{conf}，其公式如下：

$$L_{\mathrm{conf}}=-c^t\log(c)-\left(1-c^t\right)\log\left(1-c\right)\tag{4}$$

$$c^t=\min\left(1,\max\left(0,\frac{\mathrm{IoU}-\alpha_B}{\alpha_F-\alpha_B}\right)\right)\tag{5}$$

其中，c 为预测框的置信度分数；c^t 表示预测框的置信度真值；IoU表示3D预测框与标注框的交并比；α_F 和 α_B 分别表示区分前景和背景的IoU阈值．实验中，α_F=0.75，α_B=0.25，与本文的基线方法CT3D保持一致．

对于检测框的回归损失 L_{reg}，实际上是计算回归的偏移量与偏移量真值之间的距离．对于某个3D候选框的表示 $(x^a,y^a,z^a,w^a,l^a,h^a,\theta^a)$，其对应的ground truth标注框为 $(x^g,y^g,z^g,w^g,l^g,h^g,\theta^g)$．其中 x,y,z 表示在三维空间中的中心点坐标；w,l,h 表示宽、长和高；θ 表示Anchor的朝向角，即与相机坐标系 Z 轴方向的夹角．偏移量真值的计算为

$$x^t=\frac{x^g-x^a}{d^a},y^t=\frac{y^g-y^a}{d^a},z^t=\frac{x=z^g-z^a}{d^a}\tag{6}$$

$$w^t=\log\left(\frac{w^g}{w^a}\right),l^t=\log\left(\frac{l^g}{l^a}\right),h^t=\log\left(\frac{h^g}{h^a}\right)\tag{7}$$

$$\theta^t=\theta^g-\theta^a\tag{8}$$

其中，d^a 为3D候选框底部的对角线长度，计算公式如下：

$$d^a=\sqrt{(w^a)^2+(l^a)^2}\tag{9}$$

检测框中心点位置以及长宽高尺寸的预测损失计算公式如下：

$$L_{\mathrm{reg-loc+dim}}=\left(\mathrm{box}_{\mathrm{prediction}}-\mathrm{box}_t\right)^2\tag{10}$$

其中，$\mathrm{box}_{\mathrm{prediction}}$ 为整个网络输出的偏移量预测结果 $(x^p,y^p,z^p,w^p,l^p,h^p,\theta^p)$，它是回归的偏移量；$\mathrm{box}_t$ 为前述计算得到的偏移量真值 $(x^t,y^t,z^t,w^t,l^t,h^t,\theta^t)$.

为解决朝向相反的候选框被判定为相同的问题，对朝向角的预测主要包括方向预测和角度预测两个部

分,其中,方向预测部分可转化为判断正向和负向的二分类问题,如果绕Z轴的偏转真值(即θ)大于零,则朝向为正,否则为负[33]. 采用交叉熵计算方向损失L_{dir}:

$$L_{\mathrm{dir}}=-u^t\log(u)-(1-u^t)\log(1-u) \tag{11}$$

其中,u为预测方向;u^t表示朝向真值.

角度预测部分角度损失计算公式如下:

$$L_{\mathrm{reg}\text{-}\theta}=\mathrm{SmoothL1}\left(\sin\left(\theta^p-\theta^t\right)\right) \tag{12}$$

因此有

$$L_{\mathrm{reg}}=\beta_1 L_{\mathrm{reg}\text{-}\mathrm{loc}+\mathrm{dim}}+\beta_2 L_{\mathrm{reg}\text{-}\theta}+\beta_3 L_{\mathrm{dir}} \tag{13}$$

其中,β_1,β_2,β_3分别为各项损失的权重系数.

RPN网络部分的损失L_{RPN}包含分类损失L_{cls}以及候选框回归损失L'_{reg}两部分,即

$$L_{\mathrm{RPN}}=\gamma_1 L_{\mathrm{cls}}+\gamma_2 L'_{\mathrm{reg}} \tag{14}$$

其中,γ_1和γ_2分别为各项损失的权重系数.

采用Focal loss[35]计算分类损失L_{cls},计算公式如下:

$$L_{\mathrm{cls}}=-\alpha(1-p_t)^{\gamma}\log p_t \tag{15}$$

其中,α和γ为两个超参数;p_t表示预测分数. 实验中$\alpha=0.25,\gamma=2$.

对于候选框回归损失L'_{reg},其计算方式与L_{reg}一致,但RPN网络输出的3D候选框预测偏移量$(x^c,y^c,z^c,w^c,l^c,h^c,\theta^c)$是预测的候选框相对Anchor的偏移量,其偏移真值为标注框相对Anchor的偏移量.

3.2.3　注意力感知模块

在伪点云的生成过程中,为了使伪点云的表示形式与LiDar点云(x,y,z,r)保持一致,第4个维度采用了数字1填充的方式,所有点的反射率系数r均设置为1. 本质上该操作只是为了在数据格式上保持一致,对算法本身并没有提供额外的有效信息. 另外,伪点云只含有坐标位置信息,而图像有丰富语义,语义信息对3D目标检测任务来说又非常重要,能够帮助网络辨别哪些点属于同一类甚至同一个目标. 但图像和伪点云属于不同模态数据,直接融合较为复杂,需要考虑特征对齐. 研究如何有效地将图像暗含的语义信息嵌入到伪点云中非常有价值.

为了有效融合伪点云位置信息以及图像语义信息,本文设计注意力感知模块,在网络训练过程中对相同类别的目标伪点云进行关联并在伪点云特征表示中嵌入语义关联关系,有效提升3D目标检测的精度.

首先将伪点云的表示形式由原来的$(x,y,z,1)$变为(x,y,z,cls),第四个维度由原先的固定值反射率系数变成当前RGB图像语义分类结果,从而将语义标签嵌入到伪点云的数据表示形式中. 进而,针对RPN网络获取的3D候选框内的目标伪点云,其已经包含了语义类别信息cls,设计注意力感知模块,结合语义信息来构建伪点云之间的位置关联关系和语义关联关系,从而在伪点云特征表示中嵌入语义关联,并将该特征输入后续网络,进而得到3D目标检测的结果.

注意力感知模块结构如图4所示. 输入为3D候选框内的伪点云,特征维度为$B\times M\times D$. 其中B表示目标个数;M表示每个目标对应伪点云的数目;D表示每个伪点云的特征维度,本文中$D=4$,即(x,y,z,cls). 该模块主要包含点云的两种注意力机制,分别为点间注意力以及通道间注意力.

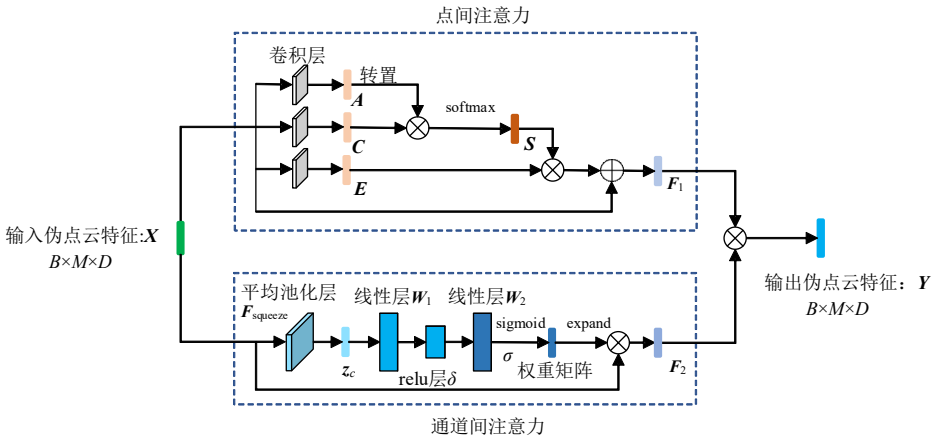


图4　注意力感知模块

针对3D候选框内的伪点云,大部分点属于目标点,具有相同类别,但是由于深度估计不够准确或者3D候选框位置不够准确,候选框内会混杂一些背景点. 由于每个点内部嵌入了语义标签并包含位置坐标,因此可通过点间注意力机制对相同类别的目标点进行特征关联,有效地区分目标伪点云,并得到点与点之间的位置坐标关系,获取其他点的上下文信息,构建伪点云之间的位置分布关系. 点间注意力机制针对输入的伪点

云特征$\boldsymbol{X}$,分别经过不同卷积层得到特征$\boldsymbol{A}$,$\boldsymbol{C}$,$\boldsymbol{E}$. 对$\boldsymbol{A}$,$\boldsymbol{C}$进行以下操作:

$$\boldsymbol{S}_{ij}=\mathrm{softmax}\left(\frac{\exp\left(\boldsymbol{A}_i\cdot\boldsymbol{C}_j\right)}{\sum_{i=1}^{M}\exp\left(\boldsymbol{A}_i\cdot\boldsymbol{C}_j\right)}\right) \tag{16}$$

结合$\boldsymbol{E}$,输出特征$\boldsymbol{F}_1$,计算公式如下:

$$\boldsymbol{F}_1=\boldsymbol{S}\cdot\boldsymbol{E}+\boldsymbol{X} \tag{17}$$

除了点间的位置分布关系,还需要关注伪点云通道之间的关联关系. 伪点云的不同通道既提供了其在3D空间中的位置信息,还额外提供了该点的语义类别,因此采用通道注意力机制来构建伪点云通道内部的关联关系. 对于输入的伪点云特征$\boldsymbol{X}$,首先经过平均池化层对其进行空间上的压缩,得到z_c:

$$z_c=F_{\mathrm{squeeze}}\left(\boldsymbol{X}_{:,c}\right)=\frac{1}{M}\sum_{i=1}^{M}\boldsymbol{X}_{i,c} \tag{18}$$

其中,$c\in(x,y,z,\mathrm{cls})$,z_c再经过一个线性层$\boldsymbol{W}_1$、一个Relu层、一个线性层$\boldsymbol{W}_2$,最后经过一个sigmoid激活得到权重矩阵$\boldsymbol{s}$,将权重矩阵的每个通道以相同数值扩充得到$M\times D$的权重矩阵后,与原始特征相乘得到输出特征$\boldsymbol{F}_2$,其计算公式如下:

$$\boldsymbol{s}=\sigma\left(\boldsymbol{W}_2\delta\left(\boldsymbol{W}_1z_c\right)\right) \tag{19}$$

$$\boldsymbol{F}_2=\mathrm{expand}(\boldsymbol{s})\cdot\boldsymbol{X} \tag{20}$$

最后,将$\boldsymbol{F}_1$和$\boldsymbol{F}_2$进行点乘操作,得到输出伪点云特征$\boldsymbol{Y}$,其计算公式如下:

$$\boldsymbol{Y}=\boldsymbol{F}_1\cdot\boldsymbol{F}_2 \tag{21}$$

输出伪点云特征的尺寸与输入伪点云特征的尺寸保持一致. 针对注意力感知模块,输入的原始伪点云,每个伪点云的特征是独立的、无关联的,特征只包含坐标信息,比较单一,但输出伪点云特征包含了伪点云内部以及点与点之间的位置坐标关联关系,特征中嵌入了语义关联,可用于后续3D检测框的精细回归.

4 实验分析比较

4.1 数据集及软硬件平台

本文采用KITTI 3D目标检测数据集,该数据集覆盖了乡村、市区以及高速公路等多种场景,可用于3D目标检测任务. 该数据集对前置摄像机采集到的图片内物体进行3D框标注,其中7 481张已公开标注信息,另外7 518张未公开标注信息,共包含带标注的物体80 256个. 对于公开标注的7 481对图片,采用VoxelNet[36]中的划分方式,选取3 712对图片作为训练集样本,3 769对图片作为验证集样本,本文实验均基于该划分方式,在训练集上进行训练,在验证集上进行性能评估. 在训练阶段,可以同时利用LiDar点云、伪点云数据进行训练;在验证阶段,只输入图像数据,利用图像数据生成伪点云.

实验采用Intel(R) Core(TM) i9-10980XE CPU @ 3.00 GHz,GPU为单卡NVIDIA GeForce 3090,内存256 GB,操作系统Ubuntu 18.04,训练框架Pytorch 1.6.

4.2 数据增强处理

KITTI 3D目标检测数据集仅有3 712张标注图片作为训练集,样本偏少,模型容易过拟合,且不同点云场景中目标数量相差较大,存在样本不均衡. 因此,本文采用随机采样、全局变换、样本空间变换进行数据增强,扩充训练样本. 随机采样是指当某类目标数量不足时,随机将存储在样本库中的目标框及包含的点云加入场景中. 全局变换包括旋转变换和尺度缩放变换,对象是整个点云场景以及所有的标注框. 样本空间变换针对标注的样本而非整个点云场景进行随机的空间变换,目的是得到各种姿态的目标,提升模型的鲁棒性.

此外,考虑到深度估计中的难样本,比如遮挡严重的目标,其伪点云无法准确表征该目标,若采用上述数据增强方法进行样本扩充会引入“坏”样本,反而影响模型精度,因此,在伪点云的样本库生成时,需要考虑大误差样本的剔除,避免在数据增强时引入“坏”样本,导致模型学习错误、精度下降. KITTI数据集对所有的标注目标都标注了遮挡程度,分为4个层次:完全可见0、部分遮挡1、遮挡严重2、完全遮挡3. 本文在样本库生成时根据训练集中标注目标的遮挡程度,将遮挡程度大于某个阈值(实验中阈值设置为1)的目标抛弃,否则将其加入样本库中.

4.3 伪点云重构实验

本节主要验证深度相关伪点云稀疏化方法对3D目标检测的有效性. 本节实验训练的epoch设置为100,BatchSize设置为3,优化器选用Adam优化器,momentum设置为0.9,学习率设置为0.001. 所有对比实验超参设置保持一致. 实验比较了角度稀疏化和本文提出的深度相关稀疏化方法,将不同方法得到的重构伪点云作为输入,进而将它输入到CT3D目标检测模型,对检测结果进行对比分析,进而评价伪点云重构效果. 检测结果以在KITTI 3D目标检测验证集上的指标3D AP,BEV AP和AOS作为评价标准,仅列出车辆目标的检测性能. 此外,考虑到本文提出的深度相关稀疏化方法,不同距离区间下稀疏化程度由参数$(D_1,W_1,D_2,W_2,D_3,W_3)$决定,本文取不同参数形成了4种方案,最后选取出效果最好的一组参数形成本文的稀疏化方法. 典型的参数值及结果如表1所示(加粗数据表示最优结果).

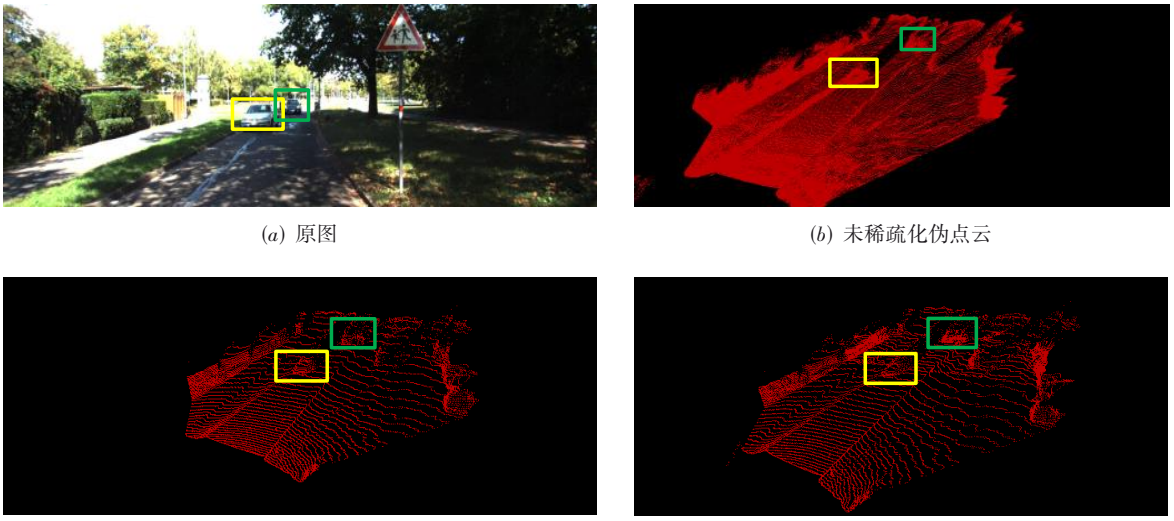
根据表1的实验结果可以看出,本文提出的深度相关稀疏化方法相比角度稀疏化效果更好,在主要检测

表1　伪点云稀疏化实验结果　　单位:%

IoU (阈值0.7)	3D AP(↑)			BEV AP(↑)			AOS(↑)		
	Easy	Mod	Hard	Easy	Mod	Hard	Easy	Mod	Hard
角度稀疏化	82.33	63.77	56.61	**89.97**	74.94	67.68	96.10	86.78	81.37
稀疏化方案1:D_1=32,W_1=512; D_2=128,W_2=512; D_3=256,W_3=2 048	82.02	65.32	58.37	89.48	75.05	67.85	96.26	87.18	81.81
稀疏化方案2:D_1=32,W_1=512; D_2=128,W_2=1 024; D_3=256,W_3=2 048	84.71	65.95	58.85	89.52	75.10	67.87	96.33	87.28	81.89
稀疏化方案3:D_1=64,W_1=512; D_2=128,W_2=1 024;d>40时无操作	83.16	66.11	59.15	89.61	75.20	68.12	**96.42**	**87.47**	82.11
稀疏化方案4:D_1=64,W_1=512; D_2=128,W_2=1 024; D_3=256,W_3=2 048	**84.80**	**66.28**	**59.30**	89.56	**75.28**	**68.21**	96.41	87.40	**82.14**

指标上均有提升. 深度相关稀疏化方法在有效去除冗余伪点云的同时,根据场景深度对稀疏程度进行区分,更多地保留了中远距离的伪点云,提高了中远距离小目标检测的准确性. 综合各项指标比较,本文提出的稀疏化方案3和方案4在各项检测指标上效果更好,尤其针对3D目标检测领域大家比较关注的"Moderate"的3D AP指标,方案4达到了66.28%,相比于角度稀疏化操作方案提升了2.51%. 因此. 本文选取稀疏化改进方案4作为本文最终的深度相关伪点云稀疏化方案.

图5展示了同一幅图对应的原始未稀疏化伪点云、角度稀疏化、本文稀疏化的可视化结果. 可以看出原始伪点云稠密,存在大量冗余点,不易于分析出车辆目标位置. 角度稀疏化和本文稀疏化方法均能去除大量冗余背景点,同时保留伪点云自身的优势. 考虑不同距离的目标,黄色矩形框对应的车辆目标距离较近,无论采用角度稀疏化还是本文深度相关稀疏化方法都能较好地保留目标点云,便于检测识别该车辆;而绿色矩形框对应的车辆目标距离较远,角度稀疏化保留的目标点偏少,本文深度相关稀疏化方法能够保留更多中远距离的目标点,有助于后续的3D目标检测.



(a) 原图　(b) 未稀疏化伪点云　(c) 角度稀疏化结果　(d) 本文方法稀疏化结果

图5　不同方法伪点云稀疏后的可视化结果

4.4　数据增强实验

下面进一步验证数据增强的效果. 如表2所示(加粗数据表示最优结果),在角度稀疏化基础上加入伪点云数据增强方法后,大部分检测指标均有提升,尤其针对3D目标检测领域大家比较关注的"Moderate"的3D AP指标,提升了0.39%. 在本文提出的深度相关稀疏化基础上加入本文伪点云数据增强方法,除"Hard"的AOS指标之外,其余指标均有提升,且相比原有的角度稀疏化方案,"Moderate"的3D AP指标提升了2.61%. 可见,本文提出的数据增强方案是有效的.

同时在表2中,本文伪点云数据增强后"Hard"的AOS指标有所下降. 我们猜测这是因为:AOS指标主要评价目标检测框的朝向,而朝向学习任务相对简单,针对"Hard"类别虽然存在标注错误的"坏"样本,但是这些"坏"样本并不能对相对简单的朝向任务产生太大影响;反观本文提出的数据增强方案,被剔除的误差较大的"坏"样本属于"Hard"级别,从而减少了模型中"Hard"级别朝向学习的样本数,因此AOS指标反而可能降低.

4.5　LiDar点云指导下伪点云目标特征生成实验

本节实验是为了验证本文提出的LiDar点云指导下特征分布趋同的伪点云目标特征生成子网络的有效

性. 实验训练集和验证集的划分方式与文献 Pseudo-LiDAR[16]一致. 实验中采用 CT3D 网络对 LiDar 点云分支和伪点云分支单独训练,得到二者的预训练模型,然后再采用本文的改进网络对伪点云分支进行微调. 预训练网络的 epoch 设置为 100,学习率设置为 0.001;本文改进网络训练的 epoch 设置为 50,学习率设置为 0.000 5. 训练过程中的 BatchSize 设置均为 3,优化器选用 Adam 优化器,momentum 设置为 0.9. 所有对比实验超参设置保持一致,检测结果如表 3 所示(加粗数据表示最优结果).

表 2 数据增强实验结果 单位:%

IoU 阈值(0.7)	3D AP(↑)			BEV AP(↑)			AOS(↑)		
	Easy	Mod	Hard	Easy	Mod	Hard	Easy	Mod	Hard
角度稀疏化	82.33	63.77	56.61	**89.97**	74.94	67.68	96.10	86.78	**81.37**
角度稀疏化+数据增强	**82.85**	**64.16**	**56.82**	89.74	**75.19**	**67.79**	**96.40**	**87.34**	79.77
本文稀疏化	84.80	66.28	59.30	89.56	75.28	68.21	96.41	87.40	**82.14**
本文稀疏化+数据增强	**84.84**	**66.38**	**59.39**	**89.86**	**75.47**	**68.26**	**96.42**	**87.56**	80.11

表 3 LiDar 点云指导下伪点云目标特征生成实验结果 单位:%

IoU 阈值(0.7)	3D AP(↑)			BEV AP(↑)			AOS(↑)		
	Easy	Mod	Hard	Easy	Mod	Hard	Easy	Mod	Hard
CT3D[32]	84.84	66.38	59.39	89.86	75.47	68.26	96.42	87.56	80.11
CT3D[32]+LiDar 点云指导	**85.27**	**66.49**	**59.40**	**90.15**	**75.83**	**68.51**	**96.47**	**87.82**	**82.48**

表 3 实验中输入统一采用经过本文提出的稀疏化和数据增强处理方法重构后的伪点云. 根据表 3 的实验结果可知,本文提出的 LiDar 点云指导下分布趋同的伪点云目标特征生成子网络是有效的. 在原有 CT3D 网络中引入 LiDar 点云指导下分布趋同的伪点云目标特征生成子网络,各项检测指标均有提升,例如“Moderate”的 3D AP 指标提升了 0.11%.

4.6 注意力感知实验结果

本节验证提出的注意力感知模块的有效性. 训练集和验证集的划分方式与 Pseudo-LiDAR[16]一致. 点云的 3D 目标检测模型同样采用 CT3D. 输入统一采用经过本文提出的深度相关稀疏化和数据增强方法重构后的伪点云. 本节实验训练的 epoch 设置为 100,BatchSize 设置为 3,优化器选用 Adam 优化器,momentum 设置为 0.9,学习率设置为 0.001. 所有对比实验超参设置保持一致. 实验结果如表 4 所示(加粗数据表示最优结果).

根据表 4 的结果可知,在原有 CT3D 网络中加入本文提出的注意力感知模块后,各项指标均有所提升,在“Moderate”的 3D AP 指标上提升了 0.16%. 在 CT3D 网络中同时引入 LiDar 点云指导下分布趋同的伪点云目标特征生成子网络以及注意力感知模块,即本文提出的目标特征分布趋同与语义关联的 3D 目标检测网络,“Moderate”的 3D AP 指标上提升了 0.57%. 可见,本文提出的注意力感知模块是有效的,能够有效地构建伪点云之间的位置关联关系,使原来相互独立的伪点云特征嵌入更多的空间结构信息和语义关联关系,具有更强的表征能力.

表 4 注意力感知实验结果 单位:%

IoU 阈值(0.7)	3D AP(↑)			BEV AP(↑)			AOS(↑)		
	Easy	Mod	Hard	Easy	Mod	Hard	Easy	Mod	Hard
CT3D	84.84	66.38	59.39	89.86	75.47	68.26	96.42	87.56	80.11
CT3D[32]+注意力感知模块	**84.97**	**66.54**	**59.43**	**90.04**	**75.79**	**68.65**	**96.48**	**87.86**	**82.39**
CT3D[32]+LiDar 点云指导+注意力感知模块	**85.67**	**66.95**	**59.76**	**90.26**	**76.06**	**68.77**	**96.63**	**87.95**	**82.60**

4.7 相关方法比较

为了与其他基于视觉的 3D 目标检测方法进行对比,表 5 给出了本文提出的特征分布趋同与语义关联的 3D 目标检测网络在 KITTI3D 目标检测官方评测集上的检测指标,并与其他方法的指标进行对比分析. 其中,Stereo R-CNN[3]和 Disp R-CNN[4]是直接基于图像的 3D 目标检测方法. 前者在立体区域建议网络(RPN)之后添加额外的分支来预测稀疏的关键点、视点和目标维数,并结合二维左右框来计算粗略的三维目标边界框,作为对比方法中的早期方案,其指标为同类方法的最低基线;后者结合了额外的实例分割掩码和部件位置图来估计感兴趣目标的视差,通过学习识别特定的形状先验以获得更精确的视差估计来提高检测精度,该网络引入感兴趣目标掩膜这种强语义信息,和本文方法在注意力感知模块中引入 RGB 图像语义分类结果的方案有相似性,因此我们也将其作为一种比较方法. 此

外,Pseudo-LiDAR[16]、OC Stereo[37]、Pseudo-LiDAR++[17]是基于图像深度转伪点云的3D目标检测方法.首次提出伪点云的网络Pseudo-LiDAR是该类方法的基线方法.Pseudo-LiDAR++考虑利用成本较低但稀疏的LiDAR传感器来消除深度估计偏差,得到位置更精确的Pseudo-LiDAR数据,使各个指标均有提升,而本文方法在训练阶段同样引入LiDAR点云来指导特征分布趋同.因此,Pseudo-LiDAR和Pseudo-LiDAR++这两类方法也作为本文的比较方法.OC Stereo提出了一种新的二维框关联和以物体为中心的立体匹配网络.该网络只估计感兴趣物体之间的视差,以解决典型深度立体匹配方法中出现的问题,进而将视差图转换为点云,采用基于LiDAR的3D目标检测网络来预测3D边界框.该方法一定程度上解决了深度立体匹配方法存在的问题,提升了AP指标,因此也作为对比方法.

由表5(加粗数据表示最优结果)可以看出,本文提出的特征分布趋同与语义关联的3D目标检测网络在KITTI3D目标检测官方评测集上达到了较好的检测效果,各项检测指标相较于表中列出的大多数模型都有了明显的提升,例如针对"Moderate"的3D AP,本文方法相较于直接基于双目图像的3D检测网络Stereo R-CNN提升了19.05%,相较于首次提出伪点云的Pseudo-LiDAR网络提升了15.23%,相较于近三年提出的其他基于双目转伪点云的3D目标检测网络OC Stereo、Pseudo-LiDAR++、Disp R-CNN也分别提升了11.68%、6.85%、3.50%.

表5　KITTI官方评测集上的3D目标检测指标　　单位:%

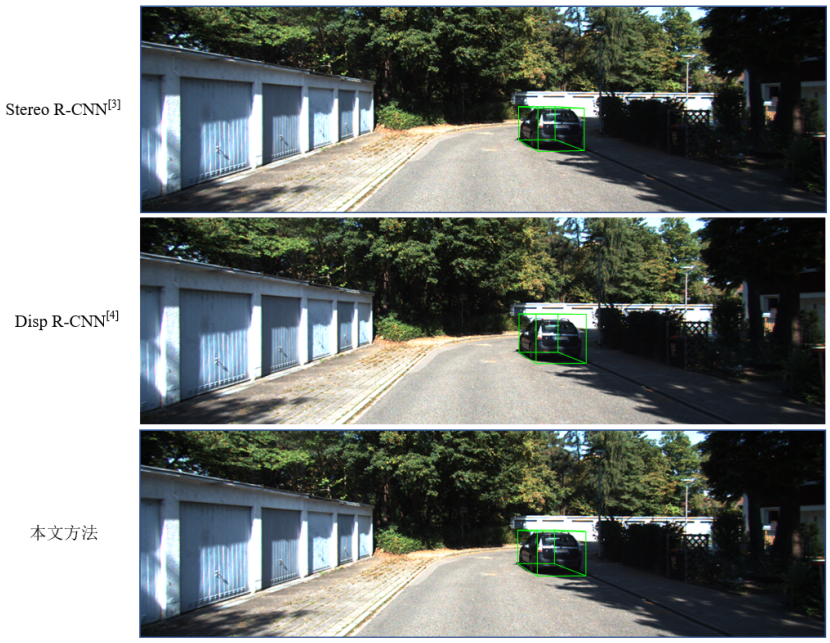
IoU阈值(0.7)	3D AP(↑)			BEV AP(↑)		
	Easy	Mod	Hard	Easy	Mod	Hard
Stereo R-CNN[3](CVPR 2019)	47.58	30.23	23.72	61.92	41.31	33.42
Pseudo-LiDAR[16](CVPR 2019)	54.53	34.05	28.25	67.30	45.00	38.40
OC Stereo[37](ICRA 2020)	55.15	37.60	30.25	68.89	51.47	42.97
Pseudo-LiDAR++[17](ICLR 2020)	61.11	42.43	36.99	78.31	58.01	51.25
Disp R-CNN[4](CVPR 2020)	68.21	45.78	37.33	79.61	57.98	47.09
本文方法	**72.16**	**49.28**	**42.56**	**81.93**	**60.56**	**51.75**

分析性能提升的原因,我们认为,Stereo R-CNN利用双目图像的语义信息和物体本身的几何信息得到的空间关系并不准确,深度估计不准造成检测精度不高;Disp R-CNN未能充分利用3D几何信息,且对2D检测算法的依赖限制了最终3D检测的性能;Pseudo-LiDAR因双目深度估计网络对远距离的物体以及物体的边缘估计准确性不高,导致Pseudo-LiDAR的位置发生偏移并影响了后续基于LiDAR的3D检测网络的精度,尤其在"Hard"等级下指标较差;Pseudo-LiDAR++仅仅是在深度估计阶段引入LiDAR点云,虽然提升了深度估计的准确性,但所得的伪点云依然不能很好地适应3D目标检测任务;OC Stereo并未深入考虑伪点云的特性,缺乏对Pseudo-LiDAR的有效重构处理,也没有针对性地设计适合伪点云的3D目标检测网络,因此与本文方法相比还存在差距.

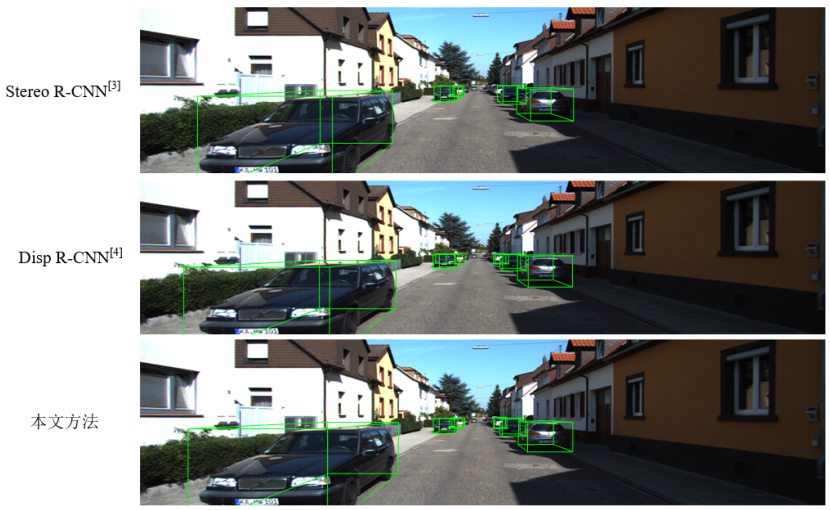
图6给出了本文所提出的3D目标检测算法与Stereo R-CNN[3]和Disp R-CNN[4]在KITTI3D数据集上的部分检测可视化结果.3D定位是3D目标检测任务最重要的指标,包括了中心点位置、长宽高以及偏向角.为了方便,我们仅计算不同方法得到的3D框中心点距离真实中心点的位置偏差,来衡量不同方法定位结果的准确性.假设当前方法得到的中心点坐标为(x_1,y_1,z_1),真值中心点坐标为(x,y,z),则位置偏差计算公式为$d=\sqrt{(x_1-x)^2+(y_1-y)^2+(z_1-z)^2}$.

观察图6(a),我们发现Stereo R-CNN和本文方法的检测结果和GT更加贴合,Disp R-CNN结果稍差.根据上述位置偏差公式可以得到$d_{\text{Stereo R-CNN}}=0.15$、$d_{\text{本文方法}}=0.28$、$d_{\text{Disp R-CNN}}=0.41$,和我们的观察一致.图6($b$)中本文方法和Disp R-CNN都检测到了6辆车,而Stereo R-CNN只检测到了5辆车.3种方法都检测到的5辆车中,各种方法的检测目标框都比较贴合.但是本文方法对小目标更有优势,例如车尾朝向我们最远的那辆被本文方法检测到的车,Stereo R-CNN并未检测到该车,Disp R-CNN虽然检测到了该车,但是位置偏差更大,具体地,$d_{\text{本文方法}}=0.05$、$d_{\text{Disp R-CNN}}=0.11$.图6($c$)的结果中,本文方法检测到了4辆车,而Stereo R-CNN只检测到了3辆车,漏检了近处被截断的车,而Disp R-CNN虽然检测到了该被截断的车,但是检测不准导致目标框投影到图片上出现了坐标混乱,出现了绿色长条直线.此外,3种方法都检测到的最远目标,$d_{\text{本文方法}}=0.25$、$d_{\text{Stereo R-CNN}}=0.91$、$d_{\text{Disp R-CNN}}=0.53$,本文得到的定位更准确.在图6($c$)中,本文方法不仅精准锁定了远、近距离车辆,还准确检测到了存在截断的车辆,比其他两个方法的结果更准确.

显然,Stereo R-CNN和Disp R-CNN由于深度估计不准造成的3D目标检测框定位不准的问题较为突出,而本



(*a*) 第20帧



(*b*) 第39帧

(*c*) 第50帧

图6 本文方法和其他方法的检测结果比较

文方法由于设计了深度相关的伪点云稀疏化方法,因此能更好地适应深度估计不准、远距离小目标的情况;本文方法由于设计了LiDar点云指导下特征分布趋同与语义关联的3D目标检测网络,因此能更好地利用伪点云进行3D目标检测,而在伪点云特征表示中嵌入目标丰富的语义关联关系,也能更好地利用图像上下文关系准确定位目标并适应目标遮挡情况. 因此本文方法在精确定位目标位置、远距离小目标方面具有优势.

实验中发现本文方法还存在少量漏检. 图7中绿色3D框内为本文算法检测到的目标,红色矩形框内为算法漏检目标. 如图7(*a*)所示,本文方法对于一些极远距离小目标仍然会出现漏检现象. 此时,目标在画面中仅有12×12像素左右大小,大概位于77 m远的地方. 此外,对于图7(*b*)中被严重遮挡的车辆目标也存在漏检现象,究其原因:一是深度信息仍然不够准确;二是远距离小目标、被严重遮挡目标,其伪点云数量较少,难以再表征出一辆车的结构信息. 因此,这些难样本的检测,仍是基于视觉的3D目标检测算法需要去重点研究和突破的方向. 我们认为可能的解决方案有以下两点:(1)设计伪点云位置修正网络,自适应地对不同距离的伪点云目标进行位置偏移修正,得到更准确的位置信息,从而提高检测精度;(2)针对小目标和遮挡目标,在训练阶段进行数据增强,增加小目标的数量、随机掩膜掉目标部分区域,迫使网络学习小目标和遮挡目标,同时,保留更多RGB图像特征,以克服LiDAR点云的稀疏性影响.

此外,本文方法的实验中,在伪点云分支、激光雷达点云分支上预训练耗时分别为0.12 h/epoch、0.13 h/epoch,在所提网络上进行联合微调训练耗时0.14 h/epoch;推理时当batch-size等于1时效率为38.1 FPS,推理阶段耗时与基于LiDar点云的CT3D目标检测网络耗时相当,能够满足实时性.

(*a*) 远距离小目标　　(*b*) 遮挡目标

图7　失败的例子

5　结束语

基于伪点云的3D目标检测往往直接采用现有的LiDar点云的3D目标检测网络进行检测,精度不高. 本文针对伪点云的特性,提出了深度相关的伪点云稀疏化方法实现伪点云重构,能够在有效去除冗余点的同时,更多地保留中远距离的小目标点云;提出了LiDar点云指导下特征分布趋同与语义关联的3D目标检测网络,在3D目标检测模型训练过程中引入LiDar点云分支来指导伪点云的目标特征编码,尽可能地让二者编码后的特征表示在分布上保持一致,减小数据源不一致造成的检测精度损失;设计了注意力感知模块,在伪点云特征表示中嵌入目标丰富的语义关联关系,提升了3D检测精度.

本文方法仅仅在训练阶段要求输入LiDar点云、伪点云数据,而在验证阶段,网络的输入只有伪点云,因此满足实际应用只需配置相机、无需激光雷达这一高造价设备的要求. 低成本的投入、高精度的检测效果使其极具应用价值.

参考文献

[1] 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.

[2] QIN Z Y, WANG J L, LU Y. Triangulation learning network: From monocular to stereo 3D object detection[C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2019: 7607-7615.

[3] LI P L, CHEN X Z, SHEN S J. Stereo R-CNN based 3D object detection for autonomous driving[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2019: 7636-7644.

[4] 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]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway:

IEEE, 2020: 10545-10554.

[5] XU Z B, ZHANG W, YE X Q, et al. ZoomNet: Part-aware adaptive zooming neural network for 3D object detection [J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(7): 12557-12564.

[6] CHEN Y L, LIU S, SHEN X Y, et al. DSGN: Deep stereo geometry network for 3D object detection[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2020: 12533-12542.

[7] WANG Y, YANG B, HU R, et al. PLUMENet: Efficient 3D object detection from stereo images[EB/OL]. (2021-01-17)[2022-10-10]. https://arxiv.org/abs/2101.06594.

[8] YAN C, SALMAN E. Mono3D: Open source cell library for monolithic 3-D integrated circuits[J]. IEEE Transactions on Circuits and Systems I: Regular Papers, 2018, 65 (3): 1075-1085.

[9] LIMAYE A, MATHEW M, NAGORI S, et al. SS3D: Single shot 3D object detector[EB/OL]. (2020-04-30) [2022-10-10]. https://arxiv.org/abs/2004.14674.

[10] 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 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2017: 2040-2049.

[11] LEPETIT V, MORENO-NOGUER F, FUA P. EPnP: An accurate O(n) solution to the PnP problem[J]. International Journal of Computer Vision, 2009, 81(2): 155-166.

[12] HE T, SOATTO S. Mono3D++: Monocular 3D vehicle detection with two-scale 3D hypotheses and task priors [J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33(1): 8409-8416.

[13] MOUSAVIAN A, ANGUELOV D, FLYNN J, et al. 3D bounding box estimation using deep learning and geometry[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2017: 5632-5640.

[14] FANG J J, ZHOU L T, LIU G Z. 3D bounding box estimation for autonomous vehicles by cascaded geometric constraints and depurated 2detections using 3D results [EB/OL]. (2019-09-01)[2022-10-10]. http://arxiv.org/abs/1909.01867v1.

[15] LIU Y X, YI Y X, LIU M. Ground-aware monocular 3D object detection for autonomous driving[J]. IEEE Robotics and Automation Letters, 2021, 6(2): 919-926.

[16] 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]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2019: 8437-8445.

[17] YOU Y R, WANG Y, CHAO W L, et al. Pseudo-LiDAR++: Accurate depth for 3D object detection in autonomous driving[EB/OL]. (2019-06-14)[2022-10-10]. http://arxiv.org/abs/1906.06310.

[18] VIANNEY J M U, AICH S, LIU B B. RefinedMPL: Refined monocular PseudoLiDAR for 3D object detection in autonomous driving[EB/OL]. (2019-11-21) [2022-10-10]. http://arxiv.org/abs/1911.09712.

[19] ZOU Z X, CHEN K Y, SHI Z W, et al. Object detection in 20 years: A survey[EB/OL]. (2019-05-13) [2022-10-10]. http://arxiv.org/abs/1905.05055.

[20] GIDARIS S, KOMODAKIS N. Object detection via a multi-region and semantic segmentation-aware CNN model[C]//2015 IEEE International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2015: 1134-1142.

[21] SHRIVASTAVA A, GUPTA A. Contextual priming and feedback for faster R-CNN[C]//European Conference on Computer Vision. Cham: Springer, 2016: 330-348.

[22] BRAHMBHATT S, CHRISTENSEN H I, HAYS J. StuffNet: Using ‘stuff’ to improve object detection[C]//2017 IEEE Winter Conference on Applications of Computer Vision (WACV). Piscataway: IEEE, 2017: 934-943.

[23] GÜNEY F, GEIGER A. Displets: Resolving stereo ambiguities using object knowledge[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2015: 4165-4175.

[24] 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 (IROS). New York: ACM, 2018: 1-8.

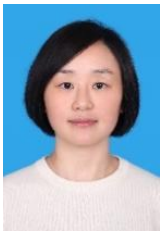
[25] VORA S, LANG A H, HELOU B, et al. PointPainting: Sequential fusion for 3D object detection[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2020: 4603-4611.

[26] HU J, SHEN L, SUN G. Squeeze-and-excitation networks [C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 7132-7141.

[27] JADERBERG M, SIMONYAN K, ZISSERMAN A, et al. Spatial transformer networks[C]//Proceedings of the 28th International Conference on Neural Information Processing Systems. Cambridge: MIT Press, 2015: 2017-

2025.
[28] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc., 2017: 6000-6010.
[29] FENG M T, ZHANG L, LIN X F, et al. Point attention network for semantic segmentation of 3D point clouds[J]. Pattern Recognition, 2020, 107:107446.
[30] QIU S, ANWAR S, BARNES N. Geometric back-projection network for point cloud classification[J]. IEEE Transactions on Multimedia, 2021, 24: 1943-1955.
[31] 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. Piscataway: IEEE, 2012: 3354-3361.
[32] SHENGA H L, CAI S J, LIU Y, et al. Improving 3D object detection with channel-wise transformer[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2021: 2723-2732.
[33] YAN Y, MAO Y X, LI B. SECOND: Sparsely embedded convolutional detection[J]. Sensors, 2018, 18(10): 3337.
[34] RUBINSTEIN R. The cross-entropy method for combinatorial and continuous optimization[J]. Methodology And Computing In Applied Probability, 1999, 1(2): 127-190.
[35] LIN T Y, GOYAL P, GIRSHICK R, et al. Focal loss for dense object detection[C]//2017 IEEE International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2017: 2999-3007.
[36] ZHOU Y, 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. Piscataway: IEEE, 2018: 4490-4499.
[37] PON A D, KU J, LI C Y, et al. Object-centric stereo matching for 3D object detection[C]//2020 IEEE International Conference on Robotics and Automation (ICRA). Piscataway: IEEE, 2020: 8383-8389.

作者简介



郑　锦　女,1978年生,四川乐山人. 于北京航空航天大学获博士学位. 现为北京航空航天大学计算机学院副教授. 主要研究方向为计算机视觉、目标检测跟踪.
E-mail: JinZheng@buaa.edu.cn

蒋博韬　男,2000年生,河北石家庄人. 2022年于武汉大学获学士学位. 现为北京航空航天大学在读硕士研究生. 主要研究方向为2D目标检测与单目3D目标检测.

彭　微　女,1997年生. 2022年于北京航空航天大学获硕士学位. 主要研究方向为2D目标检测与单目3D目标检测.

王　森　男,1998年生,江苏扬州人. 于苏州大学获学士学位. 2023年于北京航空航天大学获硕士学位. 主要研究方向为2D目标检测与单目3D目标检测.