第 50 卷 第 10 期/2023 年 5 月/中国激光 研究论文

中国激光

基于激光点云的深度语义和位置信息融合的三维目标检测

胡杰[1,2,3*]，安永鹏[1,2,3]，徐文才[1,2,3]，熊宗权[1,2,3]，刘汉[1,2,3]

[1]武汉理工大学现代汽车零部件技术湖北省重点实验室，湖北 武汉 430070；

[2]武汉理工大学汽车零部件技术湖北省协同创新中心，湖北 武汉 430070；

[3]武汉理工大学湖北省新能源与智能网联车工程技术研究中心，湖北 武汉 430070

摘要 提出一种高性能的基于深度语义和位置信息融合的双阶段三维目标检测(DSPF-RCNN)算法。在第一阶段提出深度特征提取－区域选取网络(DFE-RPN)，使网络在俯视图中能够提取目标更深层次的纹理特征和语义特征。在第二阶段提出逐点语义和位置特征融合(ASPF)模块，使网络能够自适应地提取目标最有差异性的特征，增强中心点在特征提取时的聚合能力。算法在KITTI数据集上进行测试，结果显示，测试集中Car类目标在Easy、Moderate和Hard水平的检测精度均优于现有的主流算法，检测精度分别为89.90%，81.04%和76.45%；验证集中Car和Cyclist类目标在Moderate水平的检测精度分别为84.40%和73.90%，相对于主流算法提升了4%左右，推理时间为64 ms。最后将算法部署在实车平台上实现了在线检测，验证了其工程价值。

关键词 遥感；自动驾驶；激光雷达；三维目标检测；特征融合

中图分类号 TN958.98/TN249 **文献标志码** A **DOI**：10.3788/CJL220811

1 引 言

随着人工智能和计算机科学技术的日渐成熟，自动驾驶技术也得到了快速发展。自动驾驶功能的实现离不开高效可靠的感知系统，基于激光雷达的三维目标检测是感知系统的重要部分。目标检测最早是基于二维图像，通过对像素特征的提取来实现目标的分类和二维边检框的检测。在自动驾驶的实际场景中，更多是带有深度信息的三维目标，只使用视觉来进行二维检测，既不能利用丰富的空间信息，也无法满足复杂场景下的工程需要。激光雷达能产生更丰富的空间三维信息，且受光照等自然条件的影响较小，有着更强的鲁棒性，因此基于激光雷达的三维目标检测算法研究具有重要意义。

目前，基于激光雷达的三维目标检测方法主要分为两类：单阶段方法和双阶段方法。单阶段方法[1-14]是对输入网络的点云进行特征学习，直接预测出目标的边界框和分类置信度。单阶段网络简化了很多特征提取的过程，使得算法更加高效，如Voxelnet[5]、Pixor[6]、Second[7]、3DSSD[8]、SASSD[9]、Pointpillar[10]等，但在下采样和对二维俯视图(BEV)进行特征提取时损失了大量重要信息。单阶段方法通过特征融合弥补了一些特征损失，但检测精度依然不高[11-12]。双阶段方法[15-23]增加了对三维下采样空间特征的进一步提取，用获取的特征对上一阶段粗糙候选框进行细化，得到更加精细的分类和边界框定位，如PointRCNN[15]、STD[19]、Pv-RCNN[20]、Voxel R-CNN[22]等。双阶段方法因为三维特征信息的补充细化，能够得到更高的检测精度，但消耗的时间也会增多。与传统方法相比[24-27]，基于深度学习的三维目标检测方法更多考虑点云特征提取的有效性，俯视图中有目标丰富的特征信息。但目前的双阶段方法在二维特征提取过程中没有对俯视图特征进行深层次利用，导致在粗回归和分类时不能对Cyclist、Pedestrian等较小目标进行准确的识别。三维特征提取通常是对下采样空间特征进行无差别提取，因此网络不能对距离较远或者点云形状相似的目标进行有效区分，增加了网络推理时间，但没有明显提升Cyclist、Pedestrian等目标的检测精度，甚至还会出现误检。

基于所述问题，本文提出融合深度语义和位置特征的双阶段三维目标检测(DSPF-RCNN)算法。在第一阶段设计了深度特征提取-区域选取网络(DFE-RPN)，将下采样三维空间特征转化为二维俯视图，进行更加细致和深入的特征提取，对二维图像的浅层纹

收稿日期：2022-05-05；**修回日期**：2022-07-07；**录用日期**：2022-09-07；**网络首发日期**：2022-09-17

基金项目：湖北省科技重大专项(2020AAA001，2022AAA001)

通信作者：*auto_hj@163.com

理信息和深层语义特征进行多维度和深层次的提取融合，增强网络对目标特征信息的捕获能力。在第二阶段设计逐点语义和位置特征融合（ASPF）模块，在利用关键点对三维稀疏下采样空间特征进行聚合之前，使每个中心点先获取周边点云的三维空间语义特征和位置特征，这样网络能够自适应地聚合目标最有差异性的特征，从而更好描述目标更突出的特征信息，提升目标的检测准确度并减少误检。在多尺度特征聚合中，网络只对后两个下采样体素空间特征进行聚合，相对于一般的双阶段网络，减少了推理时间。通过改进，本文提出的双阶段算法能够更好地利用目标的语义特征和空间位置信息，提高了目标的检测精度，小目标的检测精度也获得了大幅提升。

2　方法原理

DSPF-RCNN算法的组成如图1所示。在第一阶段，将激光雷达产生的无序原始点云数据划分为均匀大小的体素网格，即体素化（voxelization），使用三维稀疏卷积（3D sparse convolution）对非空体素网格进行下采样特征提取，然后将下采样最后一层的输出转化为二维的俯视图并输入到DFE-RPN模块中进行深层次的特征提取。对俯视图中的纹理信息和语义信息进行深层次的融合拼接，在提升每个像素点丰富度的同时，更深层次融合不同尺度下的特征信息，最后进行目标的粗边框回归和分类，第一阶段结束。在第二阶段，首先在第一阶段后两个三维下采样空间中，通过最远点采样选取一些点云作为中心点（center points），将中心点输入到逐点语义和位置特征融合模块，使选取的中心点能够融合周围点云的三维空间语义特征和位置信息，这样网络能够自适应地提取目标更有差异性的特征，中心点在聚合近邻点云时具备更强的特征聚合能力，增强了网络对目标不同特征信息的聚合能力。使用中心点在三维体素空间中聚合周围点的特征，即特征聚合（FA），将聚合到的特征和第一阶段的目标候选框进行感兴趣区域池化（ROI pooling）处理，最后通过全连接（FC）层对目标进行更加精细的分类和边界框回归，输出结果包括目标的三维空间尺寸、中心点坐标和朝向角。

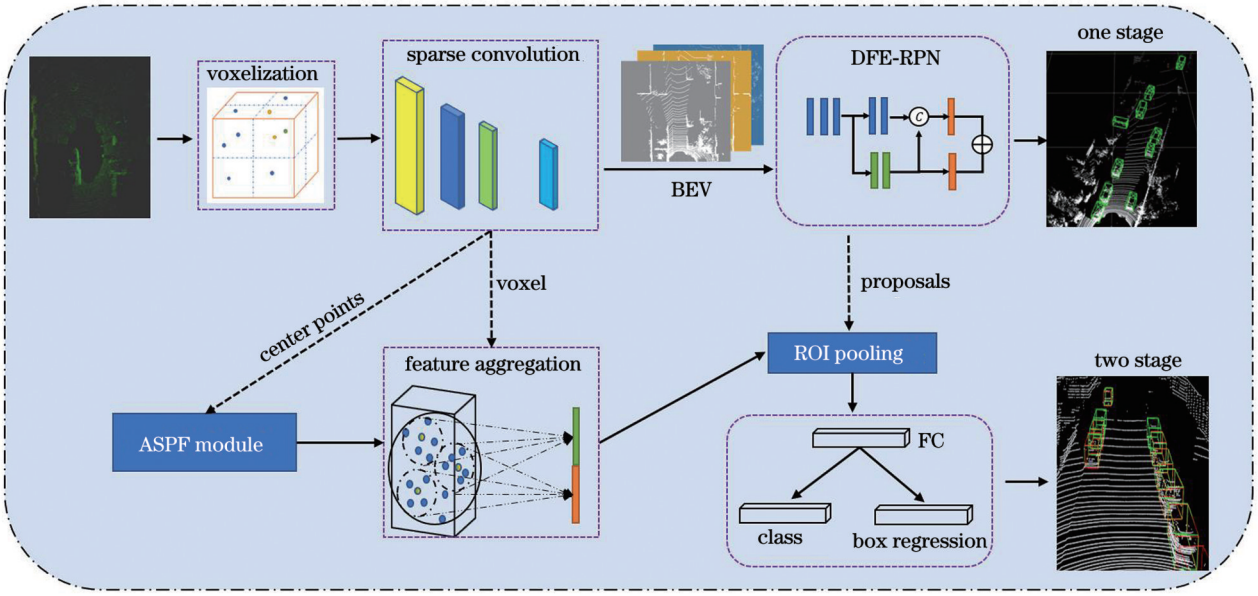

图1　DSPF-RCNN算法的流程图

Fig. 1　Flow chart of DSPF-RCNN algorithm

接下来对网络的组成模块和创新部分进行详细的介绍。

2.1　体素化和三维稀疏卷积

目前对原始点的处理有基于点（point-based）的处理[13-14]和基于体素（voxel-based）的处理[5-6]，由于point-based处理时需要消耗更多的时间，本文采用体素化方法，将原始非均匀分布的点云划分到长、宽、高相同的体素网格空间中，以达到规则化原始数据的目的。在每个体素中点云数量也是随机分布的。将逐点特征（point-wise）转化成逐体素特征（voxel-wise）时有两种常用方式：体素特征编码（VFE）和对每个体素中的所有点特征求平均值。由于后者更加简洁高效，本文采用求平均值的转化方式。对转化后的非空逐体素特征进行三维稀疏卷积操作（即卷积核尺寸为$k\times k\times k$）。通过三维空间下采样的特征提取，特征量不断增加，空间维度逐渐缩小。

2.2　DFE-RPN模块

俯视图包含目标丰富的二维特征信息，对俯视图特征进行深层次提取有利于获取目标的语义信息并提升检测精度。针对常规方法中二维俯视图特征提取不充分的问题，本文提出了DFE-RPN。网络处理过程如图2所示，将BEV（包含多张沿Y轴压缩的高度

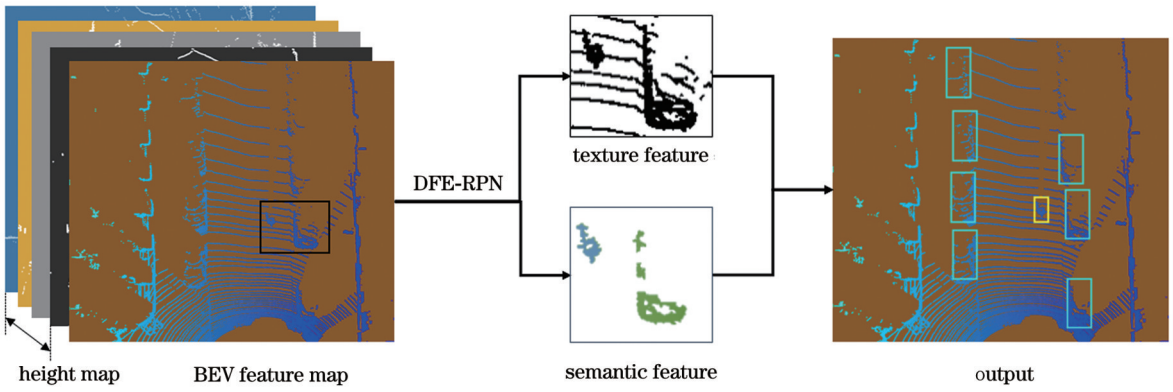

图 2　DFE-RPN 处理可视化图

Fig. 2　DFE-RPN processing visualization

特征图和强度特征图)作为输入,以提取俯视图中的浅层纹理特征和深层语义特征。纹理特征包含目标低层结构信息,语义特征包含目标深层抽象信息,进一步将两种信息进行深层次融合,使网络能够自适应地聚焦更加有价值的二维特征,最终增加网络在第一阶段对目标的检测精度,尤其是较小目标。

DFE-RPN 的详细组成如图 3 所示,包括两个部分,第一部分是多层特征提取网络,主要进行二维图像浅层纹理特征和深层语义特征的提取,包括卷积核尺寸为 3×3 且输出通道为 128 和 256 维的卷积层(conv)、卷积核尺寸为 1×1 且输出通道为 1 维的卷积层、输出通道为 256 维的反卷积层(deConv)。具体步骤如下:以二维的 BEV 作为输入,首先用三个输出通道为 128 维的卷积层进行特征初步提取,利用三个输出通道为 256 维的卷积层进行更深层特征提取,与反卷积层输出的特征进行相加(addition),输出特征称为 $\boldsymbol{X}_0$;对两个反卷积层输出的特征分别进行输出特征维度为 256 的卷积操作,对输出的特征进行像素级别相加和维度级别拼接(concat),特征输出分别称为 $\boldsymbol{X}_1$ 和 $\boldsymbol{X}_2$;两个输出通道为 1 维的卷积操作是对二维图像中重要特征信息进行权重提取,将获取的权重与卷积层输出特征相乘(multiply),使特征提取网络能够自适应地关注和提取图像中更有价值的目标信息,增加目标特征信息的利用率,最后将两个支路的权重特征相加,输出记为 $\boldsymbol{X}_3$,特征的多层提取和初步融合结束。DFE-RPN 的第二部分是深层次特征融合,即对第一部分提取到的多层特征进行更深层次拼接融合,使浅层纹理特征和深层语义特征相互补充。深层次特征融合网络的特征输入为 $\boldsymbol{X}_0$、$\boldsymbol{X}_1$、$\boldsymbol{X}_2$ 和 $\boldsymbol{X}_3$,特征维度分别 256、256、512 和 256,经过深层融合最终特征输出为 512 维。基于最后输出的二维特征,产生锚框并生成候选框,完成第一阶段检测目标的粗分类和粗回归。

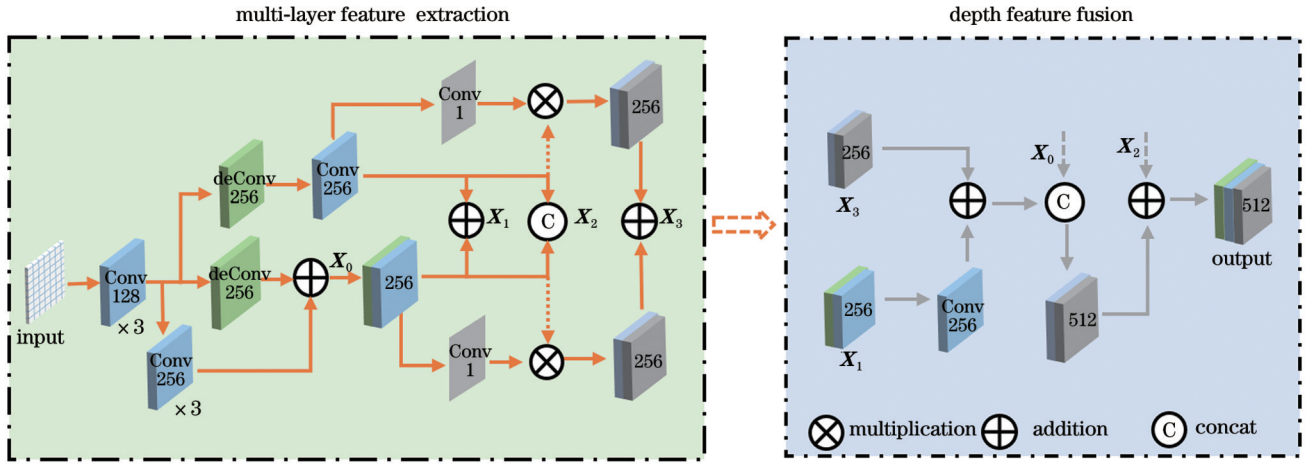

图 3　DFE-RPN 模块示意图

Fig. 3　Schematics of DFE-RPN module

2.3　逐点语义和位置特征融合模块

第一阶段完成后,获得的目标分类和回归结果比较粗糙。第二阶段主要是利用三维空间信息特征来进一步优化补充第一阶段的检测结果,所以三维特征的提取效果会显著影响最终的目标检测效果。为了更好地聚合目标的三维空间特征,本文设计了一个逐点语义和位置特征融合模块,具体过程如图 4 所示,其中 $\triangle \boldsymbol{p}$ 为位置变化量,$\triangle \boldsymbol{f}$ 为特征变化量。

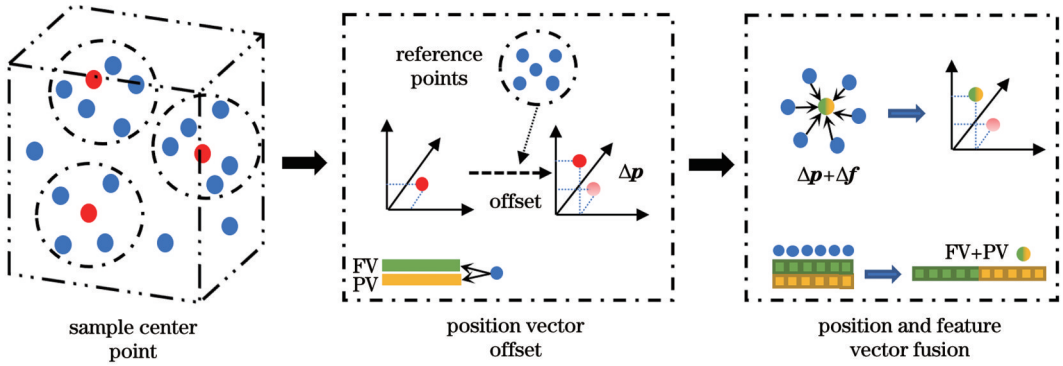

图 4　逐点特征向量和位置向量的融合示意图

Fig. 4　Fusion diagrams of aware-point feature vector and position vector

在第一阶段三维下采样体素空间中，使用最远点采样算法采样一定数量点云，这些采样点称为中心点（center point），每个中心点都包含一个三维特征向量（FV）和一个三维位置向量（PV），虚线圆圈是每个中心点需要检索的空间球形范围，球形区域内的剩余点云是中心点进行位置偏移和特征变形的参考点。首先以偏移前的中心点为基准，以参考点的三维位置向量为参考，进行中心点位置偏移。偏移过程[28]为

$$\boldsymbol{p}'_{v(i)}=\frac{1}{n}R'\left[\sum_{j\in N(i)}\boldsymbol{Q}_2\left(\boldsymbol{p}_{v(i)}-\boldsymbol{p}_{v(j)}\right)\right], \tag{1}$$

$$\boldsymbol{p}''_{v(i)}=\boldsymbol{p}'_{v(i)}+\tanh\left[\boldsymbol{Q}_3\left(\boldsymbol{f}'_{v(i)}\right)\right], \tag{2}$$

式中：n为实际参考点的数量；$\boldsymbol{p}_{v(i)}$为第i个中心点的位置向量；$v(i)$为第i个中心点的向量；$R'(\cdot)$为激活函数；i为所选取的中心点编号；$N(i)$为第i个中心点周围的参考点；j为参考点编号；$\boldsymbol{Q}_2$、$\boldsymbol{Q}_3$为需要学习的偏移和变形权重矩阵；$\boldsymbol{f}'_{v(i)}$，$\boldsymbol{p}'_{v(i)}$为获取到的加权后的偏移语义特征向量和位置向量；$\boldsymbol{p}''_{v(i)}$为对加权激活处理后的特征向量进行初步融合的位置向量。

根据参考点特征向量对中心点特征进行变形操作，最后将偏移位置向量和变形特征向量进行拼接和融合[28]：

$$\boldsymbol{f}'_{v(i)}=\frac{1}{n}R'\left[\sum_{j\in N(i)}\boldsymbol{Q}_1\left(\boldsymbol{f}_{v(i)}-\boldsymbol{f}_{v(j)}\right)\right], \tag{3}$$

$$F_i=\psi\left(\boldsymbol{f}'_{v(i)}\right)+\psi\left(\boldsymbol{p}''_{v(i)}\right), \tag{4}$$

式中：$\boldsymbol{f}_{v(i)}$为第i个中心点的特征向量；$\boldsymbol{Q}_1$为需要学习的偏移和变形权重矩阵；$\psi(\cdot)$为多层感知机（MLP）操作；F_i是第i个中心点获取到的三维空间特征向量和位置向量经过更深层融合后的特征。偏移权重和变形权重是网络需要学习的参数。对中心点进行位置偏移和特征变形，使中心点特征聚合阶段能够自适应地调节特征提取的感受野，提取到的特征更加贴合目标原始形状，从而增强网络对目标自身特有信息的提取和对形状相似目标的辨别能力。

2.4　特征聚合和池化

将ASPF模块获取的位置与语义特征融合后的中心点称为关键点（key point），通过关键点来聚合周围的点云特征，如图5所示。具体步骤如下：1）根据每个关键点设置一定的聚合范围，如图5中的球形区域，球形半径为0.4 m和0.8 m，目的是在不同的分辨率下进行特征聚合，这一过程称为组化，即为每个关键点分配聚合范围；2）通过Ac-Pointnet网络（即加速的Pointnet网络）[22]，进行特征提取；3）对不同尺度和分辨率下的特征进行拼接，获取最后聚合特征，特征维度为128。

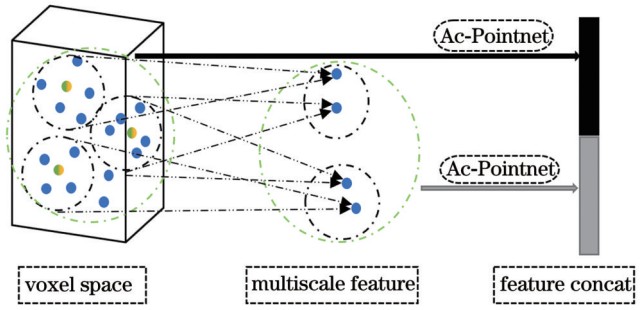

图 5　特征聚合示意图

Fig. 5　Schematic of feature aggregation

对第一阶段产生的512个候选框（proposals）进行筛选，选出128个候选框作为ROI pooling的对象。将每个目标候选框均匀划分为216个网格，以每个网格中心作为关键点池化上层获取的特征，输出三维张量（128，216，128），具体含义为共有128个proposals，每个proposal被划分为216个网格，每个网格聚合128个特征。通过最大池化操作来池化特征，将池化后的特征输入到两个全连接层中进行精细的分类和边框回归。最后确定每个目标的类别和位置信息，包括中心点坐标(X,Y,Z)、边框尺寸(L,W,H)和朝向。

2.5　损失函数

网络的损失函数包含两个部分：第一阶段的DEF-RPN损失函数和第二阶段检测头的损失函数。DEF-RPN的损失函数包括分类损失[29]和边框回归损失[10]，具体为

$$L_{\text{DFE-RPN}}=\frac{1}{N_{\text{fg}}}\left[\sum_i L_{\text{cls}}(p_i,c_i^*)+1(c_i^*\geqslant 1)\sum_i L_{\text{reg}}(\delta_i,t_i^*)\right],\tag{5}$$

$$L_{\text{cls}}=\begin{cases}-\alpha(1-y')^{\gamma}\log\ y', & y=1\\ -(1-\alpha)y'^{\gamma}\log(1-y'), & y=0\end{cases},\tag{6}$$

式中：N_{fg}为所有前景锚框（锚框内有目标）的数量；p_i和δ_i分别为网络预测的类别和回归输出；c_i^*和t_i^*分别为检测目标真实的类别值和回归参数值；$1(c_i^*\geqslant 1)$为检测目标的分类置信度；L_{cls}为分类损失函数，使用Focal损失函数[29]；α、γ为超参数，选取默认值$\alpha=0.25$，$\gamma=2$；y'为目标的预测值；y为目标真实值；L_{reg}为边框回归损失函数，使用Huber损失函数。

第二阶段检测头损失函数L_{head}[22]包括最终的分类置信度预测损失函数$L_{\text{confidence}}$和边框回归函数L_{reg}。$L_{\text{confidence}}$函数使用交叉损失熵函数，L_{reg}函数和第一阶段中的回归函数相同，均使用Huber损失函数，具体计算公式为

$$L_{\text{head}}=\frac{1}{N_{\text{s}}}\left\{\sum_i L_{\text{confidence}}\left[p_i,l_i^*(I_{\text{IoU},\,i})\right]+1(I_{\text{IoU},\,i}\geqslant\theta_{\text{reg}})\sum_i L_{\text{reg}}(\delta_i,t_i^*)\right\},\tag{7}$$

$$L_{\text{confidence}}=-\sum_{i=1}^{N}y^{(i)}\log\ \hat{y}^{(i)}+(1-y^{(i)})\log(1-\hat{y}^{(i)}),\tag{8}$$

式中：N_{s}为在训练阶段采样的候选框数量；p_i为分类网络的预测值；δ_i和t_i^*分别为回归网络的预测值和真实值；$I_{\text{IoU},\,i}$为第i个目标候选框与其真实值之间的交并比（IoU）；$l_i^*(I_{\text{IoU},\,i})$为关于IoU的函数，与前景IoU阈值$\theta_{\text{H}}$和背景IoU阈值$\theta_{\text{L}}$有关（当$I_{\text{IoU},\,i}<\theta_{\text{L}}$时，$l_i^*(I_{\text{IoU},\,i})=0$；当$\theta_{\text{L}}\leqslant I_{\text{IoU},\,i}<\theta_{\text{H}}$时，$l_i^*(I_{\text{IoU},\,i})=\dfrac{I_{\text{IoU},\,i}-\theta_{\text{L}}}{\theta_{\text{H}}-\theta_{\text{L}}}$；当$I_{\text{IoU},\,i}>\theta_{\text{H}}$时，$l_i^*(I_{\text{IoU},\,i})=1$）；$\theta_{\text{reg}}$为回归损失的IoU阈值；$1(I_{\text{IoU},\,i}\geqslant\theta_{\text{reg}})$为仅使用$I_{\text{IoU},\,i}\geqslant\theta_{\text{reg}}$的候选框进行回归计算；$N$为候选框数量；$y^{(i)}$和$\hat{y}^{(i)}$分别为目标的真实值和预测值。

3 实验验证

3.1 数据集和评价指标

本文采用的数据集是KITTI公开数据集[22]，此数据集包括7481个训练样本和7518个测试样本，其中训练样本被分为包含3712个样本的训练集（training set）和3769个样本的验证集（validation set）。采用64线激光雷达，并在相机坐标系下进行标注，标注范围为前方90°，标注7类目标种类，矩阵框数量约为15000。按照每类目标被遮挡和被截断的程度，检测难度分为容易（Easy）、中等（Moderate）和困难（Hard）三个等级。评价指标是平均精度（AP）、准确率、召回率。

3.2 网络参数介绍

网络处理每帧点云的三维空间范围是0 m≤X≤70.4 m，−40 m≤Y≤40 m，−3 m≤Z≤1 m，将每帧原始点云划分成均匀的体素网格[5]，体素网格的尺寸为0.05 m×0.05 m×0.10 m。在三维稀疏卷积操作中分别进行4次下采样操作[7]，空间网格大小为原来的1、1/2、1/4和1/8，每层对应的特征输出通道维数为16、32、64、64。在DFE-RPN中最初输入的每个卷积层分别包括三个卷积块，两个分支的输出通道维数分别为128和256，两个反卷积块输出通道维数为128，最终的融合特征输出通道维数为256。第一阶段共选取512个候选框。在ASPF模块中设置每个中心点周围点云的球形范围半径为0.5 m。在特征聚合阶段，在两个下采样空间尺度范围中，设置的聚合半径为（0.4，0.8）和（0.8，1.6）。在感兴趣区域池化阶段，在第一阶段进一步选取128个候选框，将每个候选框划分成216个均匀大小的网格，结合特征聚合阶段输出的128个聚合特征进行特征池化操作。

3.3 训练和推理参数

在训练过程中，采用端到端的方式，优化器选取Adam，在四张显卡上进行训练，batch size设置为8，学习率设置为0.01，整个训练过程在训练数据集上迭代80次，训练总时间约为9 h。前景IoU阈值θ_{H}设置为0.75，背景IoU阈值θ_{L}设置为0.25，边框回归IoU阈值θ_{reg}设置为0.55，本文采用文献[20，22]中的默认值。在第二阶段选取了128个感兴趣候选框，其中有一半是正样本，即IoU大于等于θ_{reg}。数据增强方法[30]包括整体点云绕X轴翻转、按照一定的角度自由旋转及整体点云空间的自由缩放。在推理过程中，第一阶段使用非极大值抑制（NMS）操作选出前100个感兴趣区域候选框，此时的IoU阈值设置为0.7，即当检测框的IoU置信度大于0.7时，才被视为有效的检测结果；第二阶段将选出的100个候选框再次进行NMS操作以去除冗余的预测结果，此时的IoU阈值设置为0.1。

3.4 测试集和验证集上的算法性能对比

利用训练集训练DSPF-RCNN，在KITTI官方测试集上进行测试，具体评价指标为40个不同召回位置处的平均精度，Car类目标的IoU阈值设置为0.7，Pedestrian和Cyclist的IoU阈值设置为0.5。同时，也在验证集上进行算法性能比较，IoU阈值设置方式相同。

表1是DSPF-RCNN在KITTI官方测试集上与主流方法的性能对比，分别包括Car、Cyclist和Pedestrian三类目标在Easy、Moderate和Hard水平的

表 1 3D模式下KITTI测试集上的算法性能比较

Table 1 Algorithm performance comparison on KITTI test set in 3D mode unit: %

Modality	Method	Car			Cyclist			Pedestrian		
		Easy	Moderate	Hard	Easy	Moderate	Hard	Easy	Moderate	Hard
RGB+LIDAR	AVOD[1]	73.59	65.78	58.38	60.11	44.90	38.80	38.28	31.51	26.98
	F-PointNet[31]	82.19	69.79	60.59	72.27	56.12	49.01	51.21	44.89	40.23
	AVOD-FPN[1]	83.07	71.76	65.73	63.76	50.55	44.93	50.80	42.81	40.88
LIDAR-only	SECOND[7]	83.34	72.55	65.82	71.33	52.08	45.83	51.07	42.56	37.29
	VoxelNet[5]	77.82	64.17	57.51	61.22	48.36	44.37	39.48	33.69	31.51
	PointRCNN[15]	86.96	75.64	70.70	74.96	58.82	52.53	–	–	–
	STD[19]	87.95	79.71	75.09	78.69	61.59	55.30	53.08	44.24	**41.97**
	Part-A^2[18]	87.81	78.49	73.51	79.17	63.52	**56.93**	53.10	43.35	40.06
	3DSSD[8]	88.36	79.57	74.55	**82.48**	**64.10**	56.90	**54.64**	**44.27**	40.23
	DVFENet[32]	86.20	79.18	74.58	78.73	62.00	55.18	43.55	37.50	35.33
	DA-3DSSD[33]	88.27	79.51	74.25	–	–	–	–	–	–
	CIA-SSD[12]	89.59	80.28	72.87	–	–	–	–	–	–
	SVGA-Net[34]	87.33	80.47	75.91	78.58	62.28	54.88	48.48	40.39	37.92
	DSPF-RCNN	**89.90**	**81.04**	**76.45**	77.95	62.73	56.24	49.10	41.45	38.21

检测结果。其中，DSPF-RCNN对Car类目标的检测结果均优于现有的主流算法，三个难度水平的检测AP分别为89.90%、81.04%和76.45%。表2是DSPF-RCNN在KITTI验证集上与主流方法的性能对比，此时的评价指标为11个不同召回位置处的平均精度。可以看出，DSPF-RCNN对Car、Cyclist和Pedestrian三类目标在Moderate水平上的检测结果分别为84.40%、73.90%和60.10%，相较于SVGA-Net和Part-A^2网络分别提升了4.17%、4.0%和0.05%。在3D和BEV模式下，DSPF-RCNN在KITTI验证集上的检测结果如表3、4所示，此时的评价指标是40个召回位置处(R_{40})的平均精度。

表 2 3D 模式下KITTI验证集上的算法性能对比

Table 2 Algorithm performance comparison on KITTI validation set in 3D mode unit: %

Method	Modality	Car			Cyclist	Pedestrian
		Easy	Moderate	Hard	Moderate	Moderate
SECOND[7]	LiDAR only	88.61	78.62	77.22	67.75	52.98
VoxelNet[5]	LiDAR only	81.97	65.46	62.85	–	–
PointRCNN[15]	LiDAR only	88.88	78.63	77.38	–	–
STD[19]	LiDAR only	89.70	79.80	**79.30**	–	–
3DSSD[8]	LiDAR only	89.71	79.45	78.67	–	–
SASSD[9]	LiDAR only	90.15	79.91	78.78	–	–
Part-A^2[18]	LiDAR only	89.47	79.47	78.54	69.90	60.05
CIA-SSD[12]	LiDAR only	90.04	79.81	78.80	–	–
DVFENet[33]	LiDAR only	89.81	79.52	78.35	–	–
SVGA-Net[34]	LiDAR only	**90.59**	80.23	79.15	–	–
DSPF-RCNN	LiDAR only	89.59	**84.40**	78.99	**73.90**	**60.10**

表 3 3D 模式下 KITTI 验证集上 40 个召回位置处的检测结果

Table 3 Detection results at 40 recall locations on KITTI validation set in 3D mode unit: %

IoU threshold value	Car			Cyclist			Pedestrian		
	Easy	Moderate	Hard	Easy	Moderate	Hard	Easy	Moderate	Hard
0.7	92.49	85.25	82.91	93.01	75.20	72.01	67.53	60.07	54.50

表 4 BEV 模式下 KITTI 验证集上 40 个召回位置处的检测结果

Table 4 Detection results at 40 recall locations on KITTI validation set in BEV mode unit: %

IoU threshold value	Car			Cyclist			Pedestrian		
	Easy	Moderate	Hard	Easy	Moderate	Hard	Easy	Moderate	Hard
0.7	95.77	91.49	90.81	94.01	77.30	74.20	70.27	63.22	58.28

算法在验证集上的可视化检测结果如图 6 所示，绿色框为真值框，红色框为检测框。结果显示，DSPF-RCNN 可以对目标进行准确的检测和定位。

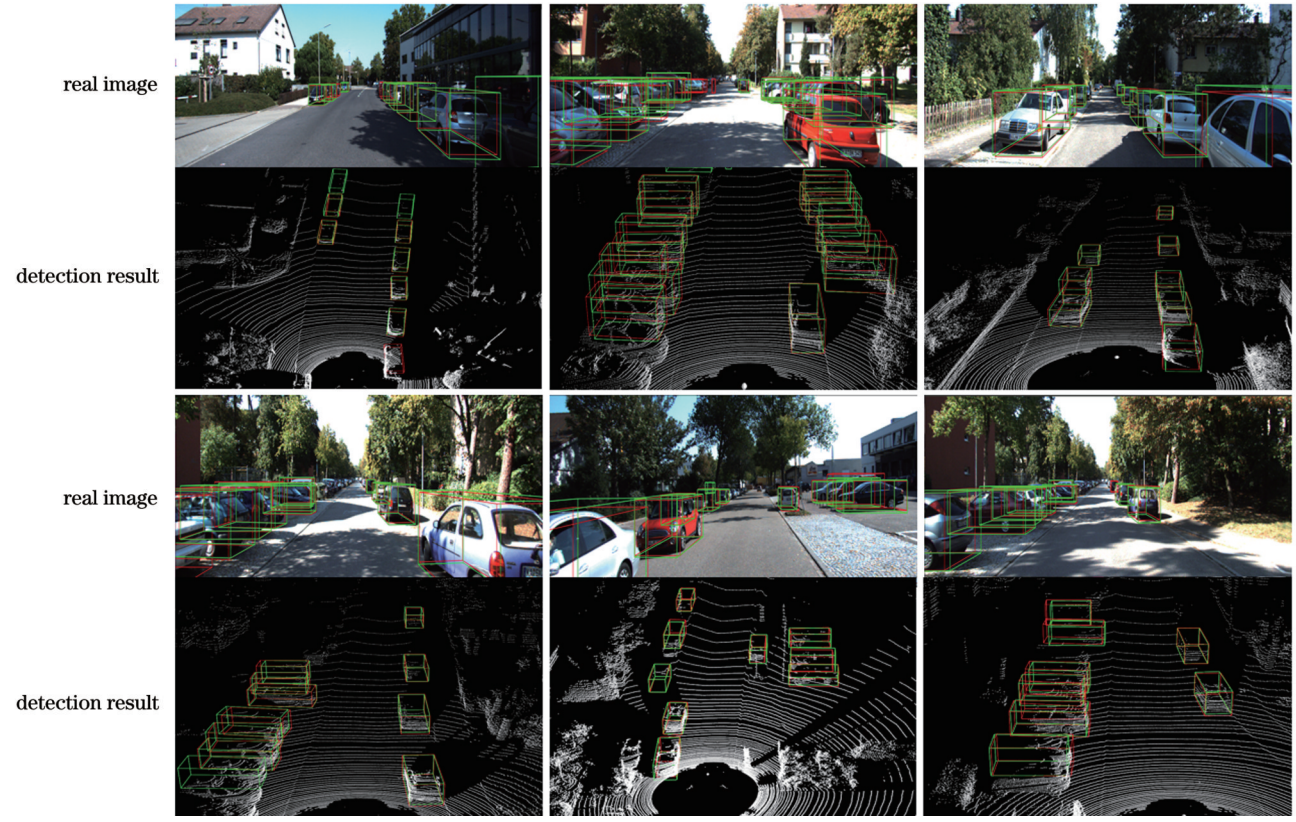

图 6 DSPF-RCNN 算法在验证集上的可视化检测结果

Fig. 6 Visual detection results of DSPF-RCNN algorithm on KITTI validation set

3.5 消融实验

为了验证每个模块在整个网络模型中的作用，本文进行模块有效性对比实验，具体的评价指标为 KITTI 验证集上 40 个召回位置处的平均精度，包括 Car、Cyclist、Pedestrian 在中等难度水平下的检测结果，如表 5 所示。

当在网络中仅加入 ASPF 模块时，结果表明，在融合周围点云三维语义特征和位置特征后，网络可以自适应地调整特征提取的感受野，能够更加有效地聚合点云特征信息；当在网络中仅使用 DFE-RPN 模块时，

表 5 ASPF 和 DFE-RPN 模块的有效性验证

Table 5 Validation of ASPF and DFE-RPN modules

ASPF	DFE-RPN	Average accuracy /%		
		Car	Cyclist	Pedestrian
		84.23	73.20	58.90
√		84.69	73.78	59.10
	√	84.72	74.30	59.43
√	√	85.25	75.20	60.07

网络能够提取到更深层次的融合浅层纹理特征和深层语义特征的目标信息，网络对特征的捕获能力得到增强，尤其是对行人和骑手等较小目标特征信息的捕获能力；在加入两个模块后，网络的检测能力得到了显著的提升。对比结果验证了本文所设计模块的有效性。

3.6 时耗对比分析

DSPF-RCNN在KITTI数据集上的运行时间如表6、7所示。表6是网络各个模块在推理1 frame点云数据时消耗的时间，包括三维稀疏卷积、DFE-RPN模块、ASPF模块、特征聚合(FA)模块、后处理(PP)模块等。其中，三维稀疏卷积和特征聚合阶段的消耗时间较长，两者时间和为40 ms，超过总时间的1/2。主要原因是三维稀疏卷积需要对整个三维空间的非空区域进行卷积处理，处理的点云数据比较庞大，所以耗时较多。在特征聚合阶段，需要在不同尺度下应用Ac-Pointnet进行特征提取，因此会有一些点云在不同维度水平下被重复提取，增加了时耗。本文只在最后两个三维下采样空间中进行特征聚合，节约了一定时间。DFE-RPN和ASPF模块分别消耗了7 ms和5 ms，占比不到总时间的1/5。

表6 各算法模块的运行时间

Table 6 Running time of each algorithm module

Module	3D Conv	DFE-RPN	ASPF	FA	PP	Other modules
Running time /ms	22	7	5	18	6	6

DSPF-RCNN与其他双阶段算法的运行时间如表7所示。DSPF-RCNN通过统一计算设备架构(CUDA)进行加速，总推理时间为64 ms，该时间均是在Car、Cyclist和Pedestrian多目标检测的基础上计算得到的。结果表明：DSPF-RCNN算法在保证较高精度的同时，推理速度在双阶段方法中具有明显的优势。

表7 DSPF-RCNN与其他双阶段方法的运行时间对比

Table 7 Comparison of running time among DSPF-RCNN and other two-stage methods

Method	PointRCNN	Part-A^2	STD	DA-3DSSD	SVGA-Net	Ours
Running time /ms	100	80	80	110	62	64

3.7 算法在本地数据集上的部署验证

为了验证算法的实际工程性能，使用实验车WHUT-E70，通过机器人操作系统(ROS)通信平台在校园场景下进行图像和雷达数据的采集，平台车的硬件部署如图7所示，安装了32线激光雷达，采集频率为10 Hz，感知范围为360°。将采集的雷达数据按照指定的格式进行标注，用标注好的本地数据集对算法进行训练。

图7 实验平台车的硬件部署

Fig. 7 Hardware deployment of experimental platform vehicle

算法在实验平台车上的检测结果如图8所示，绿色框就是网络的检测框，包括了点云数据和图像数据。其中，图像数据弥补了点云数据不能清晰分辨目标的缺点，有利于更加直观地观察图像。显示视角不一样是因为两种传感器安装的位置和硬件自身参数存在差异。可以看出，网络可以较为准确地检测出路上的行人和车辆，但是当目标被严重遮挡时也会出现漏检的情况。

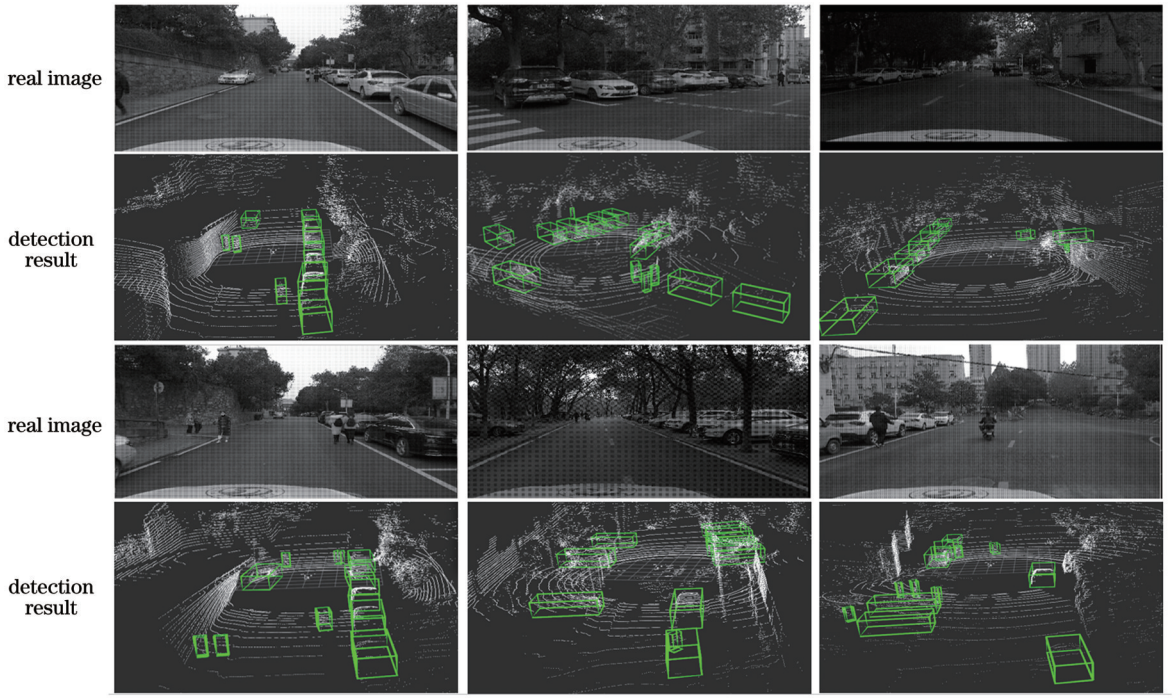

图 8 算法在实验平台车上的检测结果

Fig. 8 Test results of algorithm on experimental platform car

4 结 论

提出一个基于激光点云的双阶段目标检测算法DSPF-RCNN。在第一阶段,DFE-RPN模块使网络能够自适应地聚焦二维俯视图的纹理特征和语义特征,为候选框的精细化处理提供了更加丰富的表征信息;在第二阶段,ASPF模块对中心点进行位置偏移和特征变形,融合周围点云的三维空间语义特征和位置信息,使网络能够自适应地提取目标最有差异性的特征,提升了目标的检测精度。在KITTI测试集和验证集上进行测试并与主流方法作对比,DSPF-RCNN表现出了更有优势的性能,能够对不同大小的目标进行精确的检测,对小目标的检测也表现出更加突出的性能。实验表明:在KITTI测试集上,对Car类目标的检测精度均优于现有的主流算法;在KITTI验证集上,Car和Cyclis类目标在中等难度水平下的检测精度分别提高了4%左右,网络推理时间为64 ms。最后将DSPF-RCNN算法部署在实车上,验证了其工程价值。

参 考 文 献

[1] 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), October 1-5, 2018, Madrid, Spain. New York: IEEE Press, 2018.

[2] Graham B, Engelcke M, Maaten L V D. 3D semantic segmentation with submanifold sparse convolutional networks[C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 9224-9232.

[3] Du L, Ye X Q, Tan X, et al. Associate-3Ddet: perceptual-to-conceptual association for 3D point cloud object detection[C]// 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 13-19, Seattle, WA, USA. New York: IEEE Press, 2020: 13326-13335.

[4] Shi W J, 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 (CVPR), June 13-19, 2020, Seattle, WA, USA. New York: IEEE Press, 2020: 1708-1716.

[5] 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, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 4490-4499.

[6] Yang B, Luo W J, Urtasun R. PIXOR: real-time 3D object detection from point clouds[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 7652-7660.

[7] Yan Y, Mao Y X, Li B. SECOND: sparsely embedded convolutional detection[J]. Sensors, 2018, 18(10): 3337.

[8] Yang Z T, Sun Y N, Liu S, et al. 3DSSD: point-based 3D single stage object detector[C] //2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 13-19, 2020, Seattle, WA, USA. New York: IEEE Press, 2020: 11037-11045.

[9] He C H, Zeng H, Huang J Q, et al. Structure aware single-stage 3D object detection from point cloud[C] //2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 13-19, 2020, Seattle, WA, USA. New York: IEEE Press, 2020: 11870-11879.

[10] 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 (CVPR), June 15-20, 2019, Long Beach, CA, USA. New York: IEEE Press, 2019: 12689-12697.

[11] Kuang H W, Wang B, An J P, et al. Voxel-FPN: multi-scale voxel feature aggregation for 3D object detection from LIDAR point clouds[J]. Sensors, 2020, 20(3): 704.

[12] Zheng W, Tang W L, Chen S J, 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.
[13] Charles R Q, Hao S, Mo K C, et al. PointNet: deep learning on point sets for 3D classification and segmentation[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition, July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 77-85.
[14] Qi C R, Yi L, Su H, et al. PointNet++: deep hierarchical feature learning on point sets in a metric space[C]//NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems, December 4-9, 2017, Long Beach, CA, USA. New York: ACM Press, 2017: 5105-5114.
[15] Shi S S, Wang X G, Li H S. PointRCNN: 3D object proposal generation and detection from point cloud[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 15-20, 2019, Long Beach, CA, USA. New York: IEEE Press, 2019: 770-779.
[16] Liang M, Yang B, Chen Y, et al. Multi-task multi-sensor fusion for 3D object detection[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 15-20, 2019, Long Beach, CA, USA. New York: IEEE Press, 2019: 7337-7345.
[17] Zhou D F, Fang J, Song X B, et al. Joint 3D instance segmentation and object detection for autonomous driving[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 13-19, 2020, Seattle, WA, USA. New York: IEEE Press, 2020: 1836-1846.
[18] Shi S S, Wang Z, Shi J P, et al. From points to parts: 3D object detection from point cloud with part-aware and part-aggregation network[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(8): 2647-2664.
[19] Yang Z T, Sun Y N, Liu S, et al. STD: sparse-to-dense 3D object detector for point cloud[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV), October 27-November 2, 2019, Seoul, Korea (South). New York: IEEE Press, 2019: 1951-1960.
[20] Shi S S, Guo C X, Jiang L, et al. PV-RCNN: point-voxel feature set abstraction for 3D object detection[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 13-19, 2020, Seattle, WA, USA. New York: IEEE Press, 2020: 10526-10535.
[21] Chen Y L, Liu S, Shen X Y, et al. Fast point R-CNN[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV), October 27-November 2, 2019, Seoul, Korea (South). New York: IEEE Press, 2019: 9774-9783.
[22] Deng J J, Shi S S, Li P W, 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.
[23] Lehner J, Mitterecker A, Adler T, et al. Patch refinement: localized 3D object detection[EB/OL]. (2019-10-09)[2022-02-04]. https://arxiv.org/abs/1910.04093.
[24] 邵靖滔, 杜常清, 邹斌. 基于点云簇组合特征的激光雷达地面分割方法[J]. 激光与光电子学进展, 2021, 58(4): 0428001.
Shao J T, Du C Q, Zou B. Lidar ground segmentation method based on point cloud cluster combination feature[J]. Laser & Optoelectronics Progress, 2021, 58(4): 0428001.
[25] 张长勇, 陈治华, 韩梁. 基于改进DBSCAN的激光雷达障碍物检测[J]. 激光与光电子学进展, 2021, 58(24): 2428005.
Zhang C Y, Chen Z H, Han L. Obstacle detection of lidar based on improved DBSCAN algorithm[J]. Laser & Optoelectronics Progress, 2021, 58(24): 2428005.
[26] 梅圣明, 黄妙华, 柳子晗, 等. 基于三维激光雷达的复杂场景中地面分割方法[J]. 激光与光电子学进展, 2022, 59(10): 1028003.
Mei S M, Huang M H, Liu Z H, et al. Ground segmentation method in complex scenes based on three-dimensional lidar[J]. Laser & Optoelectronics Progress, 2022, 59(10): 1028003.
[27] 李立刚, 郭玉杰, 李林, 等. 基于变尺寸栅格地图的船载激光雷达目标检测[J]. 激光与光电子学进展, 2022, 59(8): 0828002.
Li L G, Guo Y J, Li L, et al. Target detection of shipborne lidar based on variable size grid map[J]. Laser & Optoelectronics Progress, 2022, 59(8): 0828002.
[28] Gkioxari G, Johnson J, Malik J. Mesh R-CNN[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV), October 30-November 1, 2019, Seoul, Korea (South). New York: IEEE Press, 2019: 9784-9794.
[29] Lin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 42(2): 318-327.
[30] Ye M S, Xu S J, Cao T Y. HVNet: hybrid voxel network for LiDAR based 3D object detection[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 13-19, 2020, Seattle, WA, USA. New York: IEEE Press, 2020: 1628-1637.
[31] Qi C R, Liu W, Wu C X, et al. Frustum PointNets for 3D object detection from RGB-D data[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 918-927.
[32] He Y Qi, Xia G H, Luo Y K, et al. DVFENet: dual-branch voxel feature extraction network for 3D object detection[J]. Neurocomputing, 2021, 459: 201-211.
[33] Ning J M, Da F P, Gai S Y. Density aware 3D object single stage detector[J]. IEEE Sensors Journal, 2021, 21(20): 23108-23117.
[34] He Q D, Wang Z N, Zeng H, et al. SVGA-net: sparse voxel-graph attention network for 3D object detection from point clouds [J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2022, 36(1): 870-878.

3D Object Detection Based on Deep Semantics and Position Information Fusion of Laser Point Cloud

Hu Jie[1,2,3*], An Yongpeng[1,2,3], Xu Wencai[1,2,3], Xiong Zongquan[1,2,3], Liu Han[1,2,3]

[1]*Hubei Key Laboratory of Advanced Technology for Automotive Components, Wuhan University of Technology, Wuhan* 430070, *Hubei, China*;

[2]*Hubei Collaborative Innovation Center for Automotive Components Technology, Wuhan University of Technology, Wuhan* 430070, *Hubei, China*;

[3]*Hubei Research Center for New Energy & Intelligent Connected Vehicle, Wuhan University of Technology, Wuhan* 430070, *Hubei, China*

Abstract

Object Precise perception of the surrounding environment is the basis for realizing various functions in autonomous driving. The accurate identification of the location of 3D targets in real scenes is key to improving the overall performance of autonomous driving.

Lidar has become pivotal in this field because of its superiority in sensing richer 3D spatial information while being less affected by weather and other environmental factors. Current 3D target detection methods are mainly based on deep learning, which can achieve a higher detection accuracy than traditional clustering and segmentation algorithms. The key to target detection based on deep learning is the in-depth extraction and utilization of point-cloud feature information. If feature information cannot be fully utilized, the target is misdetected or missed (Fig. 1), which has a significant impact on the safety of the automatic driving function. Therefore, deep extraction and utilization of point cloud information are key to improving the accuracy of 3D target detection.

Methods This study proposes a two-stage 3D target detection network (DSPF-RCNN, Fig. 1). In the first stage, the unordered original point cloud is divided into the regular voxel space, and the point-wise feature is converted into voxel-wise feature by using convolution neural network. The down-sampling output of the last layer is transformed into a 2D bird's eye view (BEV), whereby the BEV is input into the deep feature extraction-region proposal network (DFE-RPN, Fig. 2) for depth extraction of 2D features. Through the fusion of deep and shallow texture features with deep semantic features, the ability of the network to capture 2D image features is enhanced. In the second stage, some point clouds are selected as center points in the latter two 3D down-sampling voxel spaces through the farthest point sampling, and the center points are input into the aware-point semantics and position feature fusion (ASPF) module (Fig. 3), allowing the integration of the 3D semantic features and location information of the surrounding point clouds. In this manner, the network can adaptively extract more diverse features of the target because these center points have a stronger feature aggregation ability when aggregating neighboring point clouds, which improves the network's ability to aggregate different feature information of the target. These center points are then used to aggregate the features of the surrounding point clouds in the 3D voxel space (Fig. 4). Subsequently, the region-of-interest pooling is conducted for the aggregated features and target candidate boxes generated in the first stage. Finally, the more refined classification and boundary box regression are conducted for the target through the fully connected layer.

Discussions The DSPF-RCNN is tested and evaluated using the official KITTI test and validation sets. The detection results for Car are better than those of the existing mainstream algorithms in the test set (Table 1), and the detection accuracies at the three difficulty levels are 89.90%, 81.04%, and 76.45%. In the KITTI validation set (Table 2), at the 11 recall positions, the detection accuracy is improved by 4% compared with those of the SVGA-Net and Part-A2 networks at moderate levels for Car and Cyclist. The DSPF-RCNN can accurately detect the three types of targets (Fig. 5). The effectiveness of the proposed innovation module is further compared and analyzed (Table 5). The results show that, after integrating the 3D semantic features and position features of the surrounding point cloud, the central point can better aggregate the feature information of the surrounding point cloud in the feature aggregation stage. However, when the DFE-RPN module is added, the network's ability to capture features increase further, and the ability to extract small-target feature information, such as cyclists and pedestrians, is significantly improved. Finally, a comparative analysis is performed on the network time utilization, including the time consumed by each module in reasoning through a frame of point cloud data (Table 6). The comparison between DSPF-RCNN and the other two-stage algorithms (Table 7) shows that the total inference time of DSPF-RCNN is 64 ms, which is more advantageous in terms of the inference speed of the two-stage algorithm. Finally, the algorithm is deployed on a real vehicle platform to realize online detection (Fig. 7).

Conclusions In this study, a two-stage target detection algorithm, the DSPF-RCNN, based on a laser point cloud is proposed. First, the proposed DFE-RPN module extracts abundant target feature information from 2D images. In the second stage, the proposed ASPF module allows the central points to aggregate the salient features of different targets. Through testing on the KITTI test set and validation set, and comparison with mainstream methods, it is concluded that DSPF-RCNN performance is more advantageous in accurately detecting targets with different sizes, including small targets. At moderate levels in the KITTI validation set, the detection accuracies for Car and Cyclist are improved by approximately 4%, and the total network inference time is 64 ms. Finally, the DSPF-RCNN is applied to a local dataset to verify its engineering value.

Key words remote sensing; automatic drive; LIDAR; 3D target detection; feature fusion