doi:10.19562/j.chinasae.qcgc.2024.09.008

基于激光雷达点云的动态驾驶场景多任务分割网络*

王 海[1],李建国[1],蔡英凤[2],陈 龙[2]
(1. 江苏大学汽车与交通工程学院,镇江 212013; 2. 江苏大学汽车工程研究院,镇江 212013)

[摘要] 在自动驾驶场景理解任务中进行准确的可行驶区域以及动静态物体分割对于后续的局部运动规划和运动控制至关重要。然而当前基于激光雷达点云的通用语义分割方法并不能在车端边缘计算设备上实现实时且鲁棒的预测,且不能预测当前时刻的物体运动状态。为解决该问题本文提出一种可行驶区域及动静态物体多任务分割网络 MultiSegNet。该网络利用激光雷达输出的深度图及处理后得到的残差图像作为编码空间特征和运动特征的表征输入到网络用于特征学习,从而避免直接处理无序高密度点云。针对深度图在不同方向视角内目标分布数量差异较大的特点,本文提出了变分辨率分组输入策略。该方法能在降低网络计算量的同时提高网络的分割精度。为适配不同尺度目标所需要的卷积感受野尺寸本文提出了深度值引导的分层空洞卷积模块。同时本文为有效关联并融合不同时域下物体的空间位置和姿态信息提出了时空运动特征增强网络。为验证所提出 MultiSegNet 的有效性,本文在大规模点云驾驶场景数据集 SemanticKITTI 及 nuScenes 上进行验证。结果表明:可行驶区域、静态物体和动态物体的分割 IoU 分别达到 98%、97% 和 70%,性能优于主流网络,且在边缘计算设备上实现实时推理。

关键词:无人驾驶;激光雷达;多任务点云分割网络;动态物体分割

A LiDAR-Based Dynamic Driving Scene Multi-task Segmentation Network

Wang Hai[1], Li Jianguo[1], Cai Yingfeng[2] & Chen Long[2]
1. *School of Automotive and Traffic Engineering, Jiangsu University, Zhenjiang* 212013;
2. *Automotive Engineering Research Institute, Jiangsu University, Zhenjiang* 212013

[**Abstract**] In autonomous driving scene understanding task, accurate segmentation of drivable areas, dynamic and static objects is essential for subsequent local motion planning and motion control. However, the current general semantic segmentation method based on lidar point cloud cannot achieve real-time and robust prediction on vehicle-end edge computing devices, and cannot predict the motion state of objects at the current moment. In order to solve this problem, a multi-task segmentation network MultiSegNet for driving areas and dynamic and static objects is proposed in this paper. The network uses the depth map output by the lidar and the processed residual image as the representation of the encoded spatial features and motion features to input to the network for feature learning, so as to avoid directly processing disordered high-density point clouds. For the large difference in the number of target distributions in different directions of the depth map, a variable resolution grouping input strategy is proposed, which can reduce the amount of network computation and improve the segmentation accuracy of the network. In order to adapt to the size of the convolutional receptive field required for targets at different scales, a depth-value-guided hierarchical dilated convolution module is proposed. At the same time, in order to effectively correlate and fuse the spatial position and attitude information of objects in different time domains, a spatiotemporal motion feature enhancement network is proposed. The effectiveness of the proposed MultiSegNet is verified on the large-scale point cloud driving scene datasets SemanticKITTI and nuScenes. The results show that the segmentation IoU of driving area, static object and dynamic object reaches 98%, 97% and 70%, respectively, which is better than that of main-

* 国家重点研发计划项目(2023YFB2504401)资助。
原稿收到日期为 2024 年 01 月 23 日,修改稿收到日期为 2024 年 04 月 23 日。
通信作者:王海,教授,博士,E-mail:wanghai1019@163.com。

stream networks, with real-time inference realized on edge computing devices.

Keywords: autonomous driving; lidar; multi-task point cloud segmentation network; dynamic object segmentation

前言

环境感知模块作为自动驾驶系统的重要组成部分,须准确预测周围环境中的可行驶区域、静态物体和动态物体的空间位置信息,并输出到后续的决策规划和控制等下游任务,如跟踪和预测物体的运动轨迹,进行安全高效的路径规划和底盘控制等。

依据感知模块的传感器类型,当前主流的环境感知方法分为基于摄像头的视觉方法,基于激光雷达点云的方法,以及基于多传感器融合的方法。由于基于视觉的网络主要依靠输入图片的纹理和颜色特征进行预测,缺乏三维空间位置信息导致网络对物体在真实世界的空间位置预测的鲁棒性较差。例如,纯视觉网络难以区分与天空相同颜色的货车,且基于纯视觉的方法易受到光照变化的影响,在极端情况下难以保证行驶的安全性。而基于激光雷达点云的感知方法对环境光照强度变化有较强的鲁棒性,且能获取准确的三维空间信息。多传感器融合的感知方法通常精度更高,但是网络对设备的算力要求也较高。基于激光雷达点云的感知方法与基于视觉的方法相比具有更高的鲁棒性,与多传感器融合的方法相比对应用成本更低。因此基于激光雷达点云的方法具有环境变化鲁棒性强、对设备算力要求低、易于实现实时性等优势。在面向自动驾驶的激光雷达点云分割领域,为满足下游任务的不同需求,细分为语义分割任务[1]和动静态物体分割任务[2]等。为满足下游的运动规划需要,语义分割结果主要提供了可行驶区域分割信息、动态和静态物体分割结果。图1为所提方法的预测结果,其中绿色为可行驶区域,粉色为动态物体,橙色为静态物体。具体而言,可行驶区域和静态物体指明了当前区域的可通行性,动态物体时刻影响自动驾驶车辆后续的决策控制,具有局部不确定性。因此准确高效地分割可行驶区域和动静态物体的能力对于提高自动驾驶汽车的环境感知能力十分关键[3]。

然而,当前的研究方法都是以独立的网络推理点云语义分割与动静态物体分割结果。而采用独立

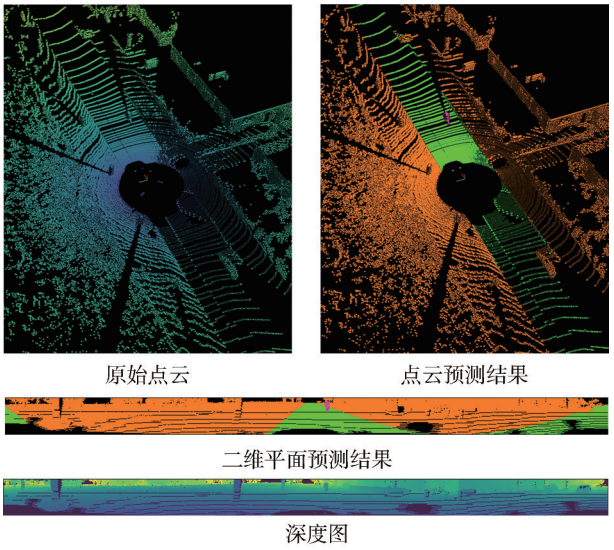


图1　MultiSegNet的输入输出效果对比

的网络分别预测相似的特征会导致计算资源的浪费,尤其对于车端计算平台而言是不可接受的。因此,对于自动驾驶感知系统而言,设计能够在车端计算平台实时运行并同时推理语义信息和物体运动状态信息的多任务网络是非常必要的。

依据基于深度神经网络的激光雷达点云分割方法主流表征的形式,可以分为如下方法:基于点的方法[4]、基于体素的方法[5]和基于深度图的方法。其中,基于点的方法其表征包含了物体清晰的3D轮廓,但是网络处理此类庞大无序的数据须消耗较大的算力,难以实现在边缘计算设备上的实时推理。而基于稀疏体素网格的方法将点云表征为一个个体素,该方法能在消耗较低算力的情况下提取信息,但是体素表征往往是残缺的,使其难以较高的细粒度的编码物体的几何形状特征。因此,该方法难以实现较高精度的分割。深度图被视为是一种相对轻量化的中间表征[6],该表征可以直接利用2D卷积神经网络提取特征,所以适用于对实时性要求较高的任务,且由于其编码了3D空间深度信息有利于网络学习到物体的空间位置及形状信息。如基于深度图的点云语义分割方法RangeForme[7]取得了优异的精度和实时性,但是由于Transformer模型的复杂网络结构使其难以在边缘计算设备上实现高效部署运行。

为解决上述问题,本文提出了一种可行驶区域和动静态物体多任务分割网络:MultiSegNet。网络的整体结构如图2所示。该网络以深度图和残差图像作为表征输入。具体而言,首先分别将编码空间形状和位置特征的深度图和残差图像输入到空间运动特征增强的双分支主干网络中进行特征提取。分别学习物体的空间位置变化特征及几何形状特征。紧接着融合两类特征并经过解码器输出2D像素空间上预测结果。最后将网络预测的2D分割结果逆投影到3D点云即得到了点云的可行驶区域即动静态物体联合分割结果。为有效地提取深度中不同尺度的目标提出了深度值引导的分层卷积模块,根据目标在深度方向的分布特点为其分配相应大小感受野的卷积块。最后针对深度图输入不同方向视角内的目标分布差异较大的特点,提出了深度图分组变分辨率输入的策略,在有效降低网络计算量的同时提升了网络预测的精度。本工作在大规模驾驶场景数据集SemanticKITTI[8]及nuScenes[9]上验证所提网络的有效性,并在边缘计算平台Nvidia Jetson TX2实现实时推理。

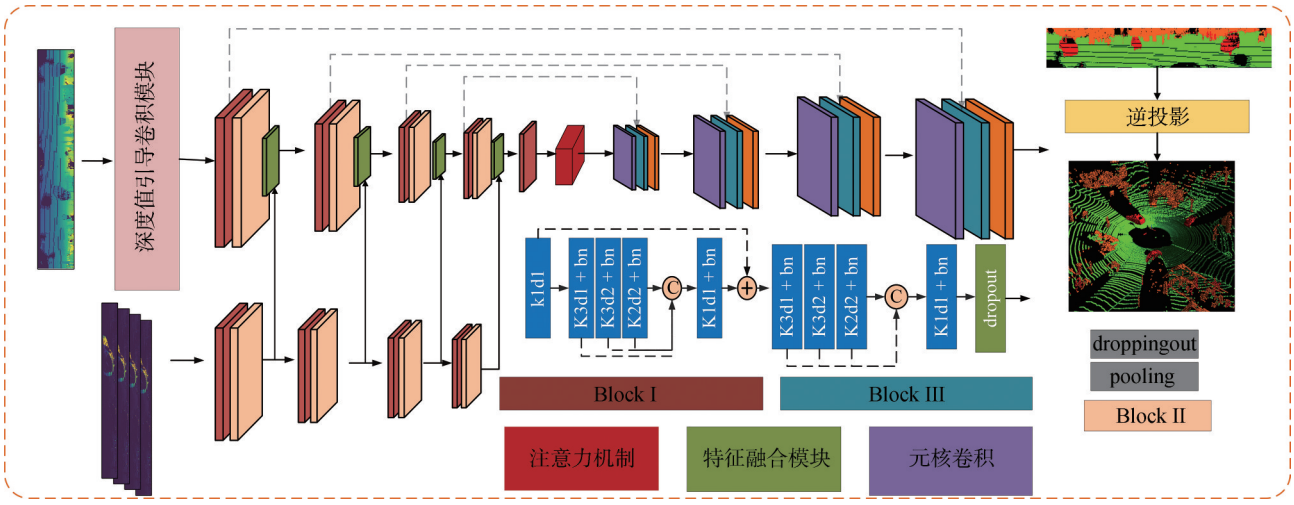


图2 本文提出的网络结构

1 多任务语义分割网络设计

1.1 基础模块

1.1.1 深度图表征

深度图[1]是一类由点云投影到二维空间的中间表征。它的优点是能够在获取三维空间信息的同时避免由于直接处理无序点云数据,且能够直接使用现有应用于视觉任务成熟的卷积神经网络。

对于每个在笛卡尔坐标系下的雷达点$p=(x,y,z)$,通过球形投影$\Pi:\mathbb{R}^3\mapsto\mathbb{R}^2$将激光点云转换到图像坐标系,定义如下:

$$\begin{pmatrix}u\\v\end{pmatrix}=\begin{pmatrix}\frac{1}{2}\left[1-\arctan(y,x)\pi^{-1}\right]w\\\left[1-\left(\arcsin(z/r)+f_{\text{up}}\right)f^{-1}\right]h\end{pmatrix}\tag{1}$$

式中:(u,v)是图像坐标;(h,w)是目标深度图表征的高和宽;$f=f_{\text{up}}+f_{\text{down}}$是传感器竖直方向的视野高度;$r=\|p_i\|_2$是每个点到传感器的距离。利用上述索引,可以提取每个点的深度信息r。本文所提网络采用深度信息r、点的笛卡尔坐标(x,y,z)和反射强度e作为深度图的5个通道值。

1.1.2 残差图像表征

本文中的网络利用残差图像[2]从时序点云中提取物体的运动状态特征。生成当前帧和历史帧之间的残差图像需要3步。首先,使用相对位姿矩阵将历史帧点云坐标l转换到当前帧点云k坐标系。然后,将转换到当前帧点云坐标的历史帧点云重投影为深度图。最后,通过计算历史帧与当前帧的归一化深度绝对差值得到每个像素的空间深度差$d_{k,i}^{l}$,定义如下:

$$d_{k,i}^{l}=\begin{cases}|r_i-r_i^{k\to i}|/r_i, & \text{真实像素点}\\0, & \text{其他}\end{cases}\tag{2}$$

式中:l是当前帧;k是历史帧;r_i是由当前帧图像坐标系对应点p_i的深度值;$r_i^{k\to l}$是历史帧坐标变换到当前帧坐标系下对应像素坐标点的深度值。

1.1.3 元核卷积模块

元核卷积(meat-kernel convolution)[10]用于捕捉

空间姿态和位置信息。在笛卡尔坐标系下,标准2D卷积难以有效提取深度图像中的深度信息。因此,元核卷积被放置在主干网络的输入位置。该方法能够通过比较笛卡尔坐标系中相邻像素的深度值,从深度图像中提取3D空间几何形状信息。具体而言,元核卷积将相对距离值输入到共享的多层感知机(MLP)中,输出9个权重向量与之对应的9个特征向量的点积,以获取相对距离向量。最后,9个向量被连接并输入到卷积中,从几个通道的不同采样位置提取信息,以更新主要的特征向量。

1.2　深度值引导的分层卷积模块

深度值引导的分层卷积模块(depth-guided lamination convolution module, DLCM)与传统的卷积操作相比,空洞卷积在卷积核中引入了一个或多个间隔(空洞)以增大感受野,同时保持计算参数数量相对较少。这使得网络能够捕捉更广泛范围的上下文信息,从而更好地理解深度图像中编码的三维空间信息。在语义分割任务中,空洞卷积可以在保持输入特征分辨率的同时,捕获更广泛的上下文信息,提升模型学习像素之间关系的能力。尽管空洞卷积可以捕捉更大范围的上下文信息,但也可能导致一些局部特征的丢失。在某些情况下,对于任务需要更加细致的局部特征的情况,可能不如普通卷积效果好。类似的,现有基于深度图的2D网络,如LMNet[2]和MotionSeg3D[11]都在主干网络中采用了大量空洞卷积。但是当使用较大的膨胀率时,卷积核在感受野内的采样点之间的距离增大,这可能导致对局部小目标特征的采样不足。为解决这个问题,本文提出了一个深度值引导的分层卷积模块。

在相机平面下空间尺寸相同的物体距离视点越近表明物体在视界中占据更大的区域。因此本文发现在深度图表征中物体的占据区域大小与深度值可以建立关联。因此,可以对不同距离的物体(不同大小)采用不同的膨胀率,以更好地适应不同大小的目标。该方法可兼顾近处目标适配大感受野,同时解决远处小目标易被忽略的问题。具体而言,首先将输入的深度图以深度值范围分为3层,0-30 m为近距离层,30-50 m为中间距离层,50 m-infinite为远距离层。分别对不同距离的深度图层的像素点采用不同大小的膨胀率的空洞卷积。也就是对近处的大目标层采用较大膨胀率的空洞卷积,对中等大小的目标层采用较小膨胀率的空洞卷积,对远距离小目标物体采用标准卷积。该模块结构如图3所示。不同大小感受野的卷积模块结构相同,先采用元核卷积模块(meta-kernel module)建立物体内相邻点空间深度关联,然后采用一个3×3的卷积块做特征提取,紧接着用一个3×3的卷积块做下采样处理,最后将元核卷积输出的特征与下采样后的特征做加和操作。不同大小感受野的卷积块膨胀率由感受野从大到小分别设置为6、3、1。

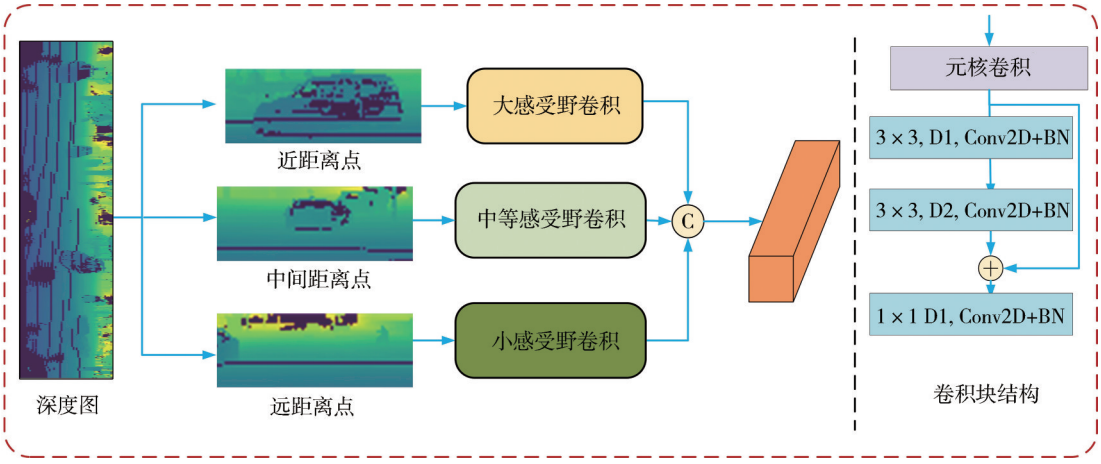


图3　本文提出的深度值引导的分层卷积模块

1.3　空间运动特征增强主干网络

空间运动特征增强主干网络(spatial-motion feature augment backbone, SFAB)在文献[11]的基础上改进得到。本文采用的主干网络采用独立的分支分别学习物体的运动特征和空间特征,并引入了空间多尺度特征池化注意力[12]解决空洞卷积容易忽略局部小目标特征的问题。在深度图表征输入空间特征提取分支前采用了深度引导的分层卷积模块,对不同大小的特征采用相应大小感受野的空洞卷积模块进行特征提取,以有效学习不同尺度的特征。

具体网络结构以双分支编码器结构和单分支解码器结构为框架。对于分割任务而言最重要的是有足够的上下文语义信息,而采用较大感受野卷的积模块可以通过建立类别间的复杂关联而获取到上下文语义信息。所以在Block Ⅰ中采用了膨胀率为2和3的空洞卷积组合,本文进一步连接每个卷积输出,并应用1 × 1卷积进行残差连接合并,以便网络有效利用来自不同网络深度融合特征中的丰富语义信息。

为编码器中关联空间特征和运动特征,本文在空间特征提取分支中采用运动注意力模块关联并融合物体的空间位置变化特征。f_a记为空间特征提取分支从深度图中提取到的空间特征,f_m记为运动特征提取分支从残差图像提取到的运动特征,融合特征可以记为

$$f_a' = f_a \otimes \mathrm{Sigmoid}\left(\mathrm{Conv}_{1\times1}\left(f_m\right)\right) \tag{3}$$

$$f_a'' = f_a' \otimes \left[\mathrm{Softmax}\left(\mathrm{Conv}_{1\times1}\left(\mathrm{APoll}\left(f_a'\right)\right)\right)\cdot C\right] + f_a \tag{4}$$

式中:f代表特征图尺寸$C \times h \times w$;APool(·)记为空间特征维度上的平均池化操作。

该模块可采用空间注意力机制增强空间位置特征并生成不同时空下的物体运动特征f_m。然后采用通道注意力机制增强空间特征和运动特征关联,最后生成融合特征。

紧接着在提出的骨干网络的颈部,引入空间池化金字塔注意力模块来解决空洞卷积导致的遗漏局部特征的问题。空间池化金字塔注意力模块由1个1 × 1和3个3 × 3的卷积块组成,其膨胀比为(6,12,18)。

1.4 深度图变分辨率分组输入策略

深度图变分辨率输入(variable resolution input,VRInput)。为更进一步降低网络的计算量,Rangeformer提出了分组深度图训练范式。具体而言,将点云分割到多个等间隔的水平方位角的点云子集。不同的Z个子集组成了无重叠的360°全景视角。然后将多个子集的点云分别栅格化投影,得到的高分辨率深度图会缓解投影点重合和物体变形的问题。通过将输入深度图按视角分组输入训练,模型训练时的水平分辨率就能降为之前的1/Z,$W_{\_train} = W/Z$。同时深度图网格的投影细粒度也能较好地保持。

本文发现点云在投影到深度图后两侧视角和前后视角内的物体数量及视角占据比例分布差异较大,两侧多为距离较近的车辆和行人且目标数量较少,单个目标占视场的比例较大,前向视角则多为密集且分布的距离差异较大的目标。针对该问题,本文期望通过提高前视视角深度图块的分辨率同时降低两侧视角深度图块的分辨率以提升计算效率,并提升模型对前向视角内目标的分割精度。如图4所示,在分组深度图训练的基础上增大前向视角A范围内深度图块的分辨率,减小两侧视角B、D范围内深度图块的分辨率,后向视角C范围内的深度图分辨率不变。为保证计算资源的利用效率最优,本文采用A视角宽度方向分辨率为1 024,两侧B、D视角水平分辨率为256,后侧视角C分辨率为512,高度方向分辨率均为512。

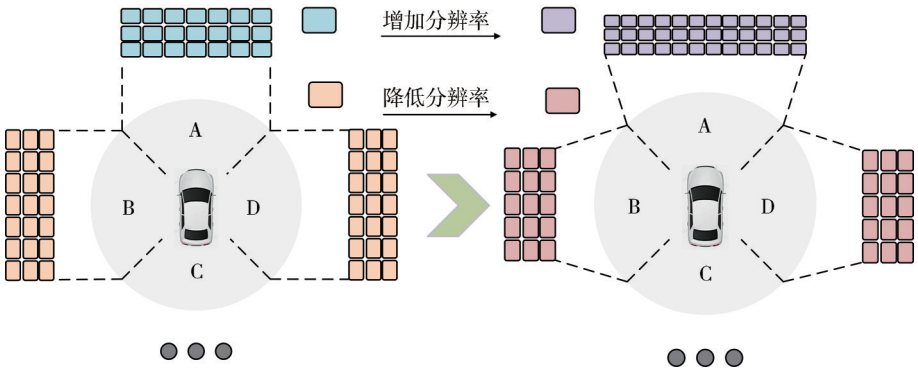


图4 深度图变分辨率输入

训练时的水平分辨率可降为之前的1/4,同时深度图的投影颗粒度也能较好保持。训练时将深度图分为4部分堆叠输入网络进行训练,A视角下的深度图块分为两部分,B、C视角下的深度图块合并为一部分,D视角下的深度图块作为一部分,每一部分的深度图块分辨率为512×512。训练时在每帧点云中随机选取一个深度图部分参与训练,而在推理时将一帧点云的所有的深度图块堆叠作为一个batch用

于推理。尽管是一个从经验出发设计的训练策略,但是实验发现这个训练策略相较于普通的训练策略在相同的迭代次数下是更容易收敛的同时内存消耗降为原来的1/4。

1.5　后处理

基于投影的点云表征的主要缺陷是离散化错误和模糊响应的卷积层引起的信息丢失。例如,当深度图像重新投影回原始的3D空间时,就会出现此问题。原因是在逆投影过程中,许多点云可能被分配到同一像素点,导致误分类某些物体边缘。当物体在背景场景中投下阴影时,这个问题会变得更加明显。

为解决反投影中投影点重合的问题,本文采用了基于kNN的后处理方法[1]。通过利用每个匹配图像像素周围的窗口来处理每个LiDAR点,然后将其转换为点云的子集,再使用kNN确定最近邻的集合。

1.6　损失函数

采用两个损失函数来监督网络。总的损失函数结合权重交叉熵和Lovàsz-Softmax损失函数[13]:$\mathcal{L}=\mathcal{L}_{\text{wce}}+\mathcal{L}_{\text{ls}}$。训练数据集类别不平衡问题给神经网络训练带来了挑战。在动静态以及可行驶区域分割任务的分类标准下,SemanticKITTI中静态物体、动态物体的类别数相差较大;这使得网络更加偏向于学习在数据集中出现次数较多的类别,导致网络在出现次数较少的类别数上的学习表现很差。

本文采用Focal loss[14]来解决样本不平衡的问题,使网络更加关注难以被学习到的物体类别。具体来说,采用softmax交叉熵损失函数$\mathcal{L}_{\text{wce}}(y,\hat{y})$依照不同类别在数据集中出现的频率的倒平方根给不同的物体类别分配不同的学习权重,即

$$\mathcal{L}_{\text{wce}}(y,\hat{y})=-\sum_{1}^{i}\alpha_i p(y_i)\log\left(p(\hat{y}_i)\right) \tag{5}$$

式中:$\alpha_i=1/\sqrt{f_i}$;y_i和$\hat{y}_i$是真实类别标签和预测到的类别标签;f_i代表频率,例如第i类物体的点云个数与总点云个数的比值。这增强了网络在类别数占比较少的物体上特征学习的性能。Lovàsz-Softmax损失函数表达式为

$$\mathcal{L}_{\text{ls}}=\frac{1}{|C|}\sum_{c\in C}\bar{\triangle}_{J_c}(m_i(c)) \tag{6}$$

$$m_i(c)=\begin{cases}1-x_i(c), & c=y_i(c)\\ x_i(c), & \text{其他}\end{cases} \tag{7}$$

式中:$|C|$是类别个数;$\bar{\triangle}_{J_c}$定义为Jaccard index的Lovàsz扩展;$x_i(c)\in[0,1]$和$y_i(c)\in\{-1,1\}$分别代表每个像素点的预测类别概率和真值标签。

2　实验验证

2.1　实验操作细节

2.1.1　数据集

为验证所提方法在点云大规模复杂场景下的性能,本文在SemanticKITTI数据集和nuScenes数据集上进行了充分的实验以评估所提方法的性能。

SemanticKITTI涉及了城区道路、居民区和高速公路场景及乡村道路场景。SemanticKITTI数据集总共包含22个序列,其中有10个序列(19 130帧)用于训练,1个序列(4 071帧)用于验证,以及11个序列(20 351帧)用于测试。nuScenes数据集包括了城区、高速公路、各类型的交叉口如普通交叉口、环形交叉口、立交交叉口等复杂场景用于训练和测试。在所提模型训练和验证时将所有的语义类别重新映射为只有3类:可行驶区域、动态物体和静态物体。具体将原类别中的未占用道路及停车场映射为可行驶区域,将动态物体类别如行驶的车辆、行人、骑行者映射为动态物体,其他类别映射为静态物体,由于原类别标注相较重映射类别细粒度更高,所以类别重新映射无须重新标注数据。

2.1.2　评估指标

对于衡量动静态物体以及可行驶区域分割性能,使用交并比(IoU)指标。IoU的定义如下:

$$\text{IoU}=PT/(PT+FP+FN) \tag{8}$$

式中PT、FP和FN分别表示移动类别的真正例、假正例和假负例的预测。

2.1.3　实验参数配置

与Rangenet++类似,在SemanticKITTI上的深度图尺寸为512×512。实验采用了Adam优化器[15]和OneCycle-Scheduler,学习率设置为$lr=1e-3$。生成的残差图像采用当前帧及前2帧和前5帧的扫描点云。对于深度图变分辨率输入,本文将环视深度图分为4部分,大小为512×512。网络在SemanticKITTI上共训练60个epoch,训练的batchsize为32。与SalsaNext[16]相同,该方法采用了Cross-entropy dice loss和Lovasz-softmax loss监督模型训练。模型训练使用4张Nvidia RTX3090 GPU进行训练。

2.2　消融实验

2.2.1　网络结构组成模块消融

在SemanticKITTI数据集的验证集和测试集上

进行消融实验,并根据实验结果分析所采用不同模块和结构改进的有效性。对于每个设置的验证网络训练了3次并报告平均结果。如表1所示,首先对本文中所提出的模块和结构改进进行消融研究。消融实验以MotionSeg3D作为基线模型。改进后的空间运动特征增强网络(SANet)与基线模型相比mIoU提升了0.6%;动态物体的分割IoU提升了1.0%,静态物体分割IoU提升了0.6%,表明在特征融合模块中加入的空间金字塔池化注意力能够有效关联物体的空间位置和运动模式特征,从而有效地提升网络的动态物体和静态物体分割性能。在基线模型的基础上加入深度值引导的分层卷积模块(DLCM),mIoU提升了0.7%,动态物体分割IoU提升了1.3%,静态物体分割IoU提升了0.7%。表明分层卷积能够依照不同的深度范围对不同尺寸的物体特征采用合适大小感受野的卷积进行特征提取,从而有效提升模型的分割精度。采用变分辨率深度图表征输入(VRInput),网络分割的mIoU提升了0.7%,动态物体分割IoU提升了1.4%,静态物体分割mIoU提升了0.7%,由此说明变分辨率输入能够有效提升小目标物体输入表征的细粒度,并更加清晰地反映前向视角小目标外部轮廓,从而提升网络的分割精度。

表1 所提方法的消融实验

项目	Baseline	SANet	DLCM	VRInput	Free Space/%	Static Object/%	Moving Object/%	mIoU/%	性能提升/%
(a)	√				97.6	96.3	66.6	86.8	
(b)	√	√			97.8	96.9	67.6	87.4	0.6
(c)	√		√		97.6	97.0	67.9	87.5	0.7
(d)	√			√	97.8	96.7	68.0	87.5	0.7
(e)	√	√	√		98.0	97.3	68.4	87.9	1.1
(f)	√	√	√	√	**98.1**	**97.8**	**69.3**	**88.4**	**1.6**

改进后的整体网络模型分割mIoU提升了1.6%,可行驶区域分割IoU提升了0.5%。与其他类别相比提升较少,主要原因为可行驶区域的点云占整体点云的比例较高,而特征相对简单,因此网络已经能够较好地学习该特征,所以基线模型的分割精度已经达到了97.6%,而所提模型在此基础上提升了0.5%,说明本文的模型对于简单特征有着更优的学习能力。而静态物体分割IoU提高了1.5%,动态物体分割IoU提升了2.8%,验证了本文所提方法的有效性。

2.2.2 深度引导的分层卷积模块消融

为有效提取不同尺寸的物体特征,深度引导的分层卷积模块采用3个不同大小感受野的空洞卷积模块提取相应大小的特征。小感受野卷积块膨胀率SDR=1,大感受野卷积膨胀率与基线模型采用的膨胀率保持一致,为LDR=5。由于中等尺寸特征的物体占比最多,中等尺寸感受野的空洞卷积模块的精度提升相较其他尺寸的空洞卷积对模型性能提升最明显。

本节验证了不同膨胀率的空洞卷积对模型分割性能的影响。如表2所示,随着中等尺度感受野卷积膨胀率由2增加到3,模型的分割性能提升最大,增大到4后模型的分割性能开始下降,说明膨胀率为3的中等尺度感受野卷积能较好地适应30-50 m距离内物体分布特征,使得深度值引导的分层卷积模块可发挥较优的特征提取能力。

表2 深度值引导的分层卷积模块的消融

项目	Convolution Dilatation Rate	Free Space/%	Static Object/%	Moving Object/%	Mean IoU /%
(a)	2	97.6	96.5	67.6	87.2
(b)	3	**98.1**	**96.8**	**68.4**	**87.7**
(c)	4	97.7	96.6	67.7	87.3
(d)	5	97.6	96.7	67.5	87.2

2.3 网络性能比较

如表3和表4所示,所提方法在SemanticKITTI及nuScenes均超过目前主流的点云语义分割方法2DPASS,及主流的点云动态物体分割方法MotionSeg3D。LMNet和MotionSeg3D由于具有针对动态特征提取而设计的结构,所以上述方法的动态物体分割精度相较2DPSS的精度更优,但是静态物体的分割性能较差。而本文所提方法的静态物体和动态物体分割精度在两个数据集上均为最高。其中静态物体在SemanticKITTI和nuScenes的分割精度分别达到97.5%和98.0%,说明本文所提方法在处理静态场景元素时具有良好的稳定性和准确性,可保证网络模型在复杂环境下预测结果的鲁棒性同时提升整体场景分割的精度。动态物体的分割精度分别为70.6%和70.8%。动态物体的分割是3D分割

任务中的一个难点,由于动态物体通常具有更高的动态性和不确定性。所提模型在动态场景下的良好性能,表明本文所提网络模型具备处理复杂动态场景时的能力。现有可行驶区域分割方法精度已经达到98.0%,而所提方法分割精度在SemanticKITTI和nuScenes与主流点云语义分割方法相比,性能提升达到了0.4%,表明本文所提方法具有较好的结构化特征提取能力。

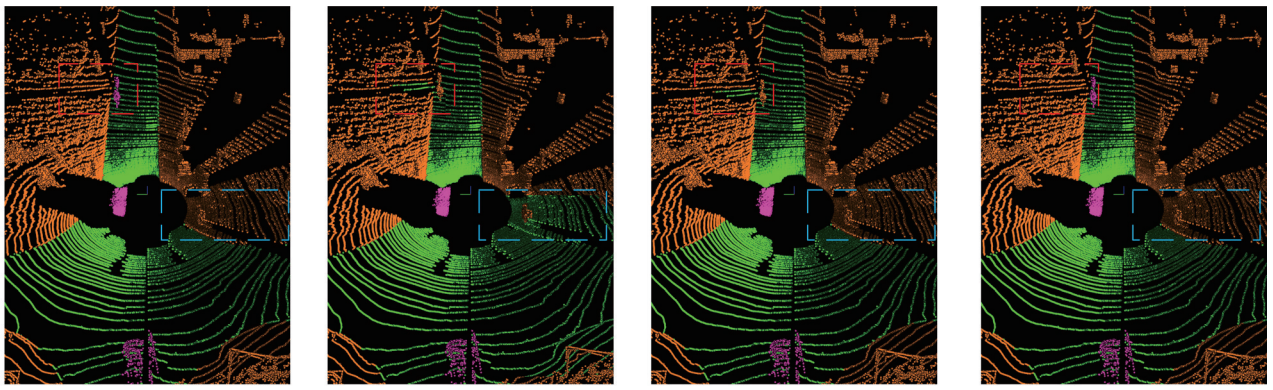
表3　在SemanticKITTI验证集对分割IoU评估和比较

方法	Modality	Free Space /%	Static Object /%	Moving Object /%	mIoU /%	推理耗时 /ms
Cylinder3D[17]	Voxel-based	97.8	95.3	60.6	84.5	170
PolarNet[18]	BEV-based	97.1	95.7	62.3	85.0	62.43
LMNet[2]	RV-based	96.4	95.2	68.4	86.6	14.26
2DPASS[19]	RGB&RV-based	98.0	96.5	61.8	85.4	62.85
4DMOS[20]	RV-based	97.3	95.7	68.6	87.4	51.23
MotionSeg3D[11]	RV-based	97.2	96.1	69.3	87.8	50.27
Ours	RV-based	**98.4**	**97.5**	**70.6**	**88.8**	**46.34**

表4　在nuScenes验证集对分割IoU评估和比较

方法	Modality	Free Space /%	Static Object /%	Moving Object /%	mIoU /%	推理耗时 /ms
Cylinder3D[17]	Voxel-based	98.0	95.8	60.9	84.9	170
PolarNet[18]	BEV-based	97.3	95.9	62.5	85.2	62.49
LMNet[2]	RV-based	96.7	95.4	68.6	87.0	14.23
2DPASS[19]	RGB&RV-based	98.4	96.9	61.8	85.7	62.87
4DMOS[20]	RV-based	97.6	95.8	68.9	87.5	51.26
MotionSeg3D[11]	RV-based	97.8	96.6	69.7	88.2	50.24
Ours	RV-based	**98.8**	**98.0**	**70.8**	**89.2**	**46.37**

更多定性比较如图5所示。图5(b)显示LMNet将道路右侧的行人通道误检为可行驶区域,且将正在运动的骑行者误检为静态物体。类似的如图5(c)中,MotionSeg3D错误预测正在运动的骑行者。如图5(d)所示,本文中提出的方法能够准确完整地预测静止车辆、运动车辆及骑行者的运动状态,说明所提方法在动态驾驶环境下能较好地分割可行驶区域、静态物体及动态物体。



(a) Ground Truth Labels　(b) LMNet-V2　(c) Cylinder3D　(d) MultiSegNet

图5　不同方法分割性能的定性比较

2.4　网络运行效率

网络运行耗时的测试实验配置为Intel Xeon Silver 4210R CPU @ 2.40 GHz和NVIDIA RTX 3090 GPU。表3和表4分别示出所提方法与几类基准方法在SemanticKITTI及nuScenes数据集的平均推理时间对比。结果表明所提方法与相近精度的方法相比,推理耗时最低为46 ms。本文所提模型部署在边缘计算平台Nvidia Jetson TX2推理耗时为35 ms。

3 结论

提出一种基于LiDAR点云的可行驶区域、静态物体和动态物体多任务分割的网络。所提方法采用深度值引导的分层卷积模块和空间特征增强网络及变分辨率深度图分组输入策略。上述方法能够有效建立物体的空间位置和运动模式间的关联,从而提升多任务分割模型的精度。所提方法基于深度图像,对计算资源的需求与基于点云的分割方法相比更少,有利于实现模型的离线实时部署运行。同时在SemanticKITTI及nuScenes数据集上的实验结果表明,所提方法分割性能优于主流的点云分割方法。

由于基于视觉的方法能够提取物体的颜色和纹理特征,基于视觉的方法更适合学习更为复杂的语义特征。在未来的研究中可结合相机RGB图像利用多模态模型预测更加细致的语义输出,也可同时推理物体的运动状态信息。后续工作可融合图像输入提取物体的颜色及纹理特征,用于分析道路交通标识、地面指示标、行人及车辆的运动状态等信息,进而预测更完整的场景信息。

参考文献

[1] MILIOTO A, VIZZO I, BEHLEY J, et al. RangeNet++: fast and accurate lidar semantic segmentation[C]. 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2019: 4213-4220.

[2] CHEN X, LI S, MERSCH B, et al. Moving object segmentation in 3D LiDAR data: a learning-based approach exploiting sequential data[J]. IEEE Robotics and Automation Letters, 2021, 6(4): 6529-6536.

[3] 蔡英凤, 王海, 陈小波, 等. 驾驶辅助系统基于融合显著性的行人检测算法[J]. 汽车工程, 2015, 37(10): 1215-1220.
CAI Yingfeng, WANG Hai, CHEN Xiaobo, et al. A pedestrian detection algorithm based on fusion saliency for driving assistance systems [J]. Automotive Engineering, 2015, 37 (10): 1215-1220.

[4] LANG A H, VORA S, CAESAR H, et al. Pointpillars: fast encoders for object detection from point clouds[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019: 12697-12705.

[5] MATURANA D, SCHERER S. VoxNet: a 3D convolutional neural network for real-time object recognition[C]. 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2015: 922-928.

[6] WANG S, ZHU J, ZHANG R. Meta-RangeSeg: lidar sequence semantic segmentation using multiple feature aggregation[J]. IEEE Robotics and Automation Letters, 2022, 7(4): 9739-9746.

[7] KONG L, LIU Y, CHEN R, et al. Rethinking range view representation for lidar segmentation[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023: 228-240.

[8] BEHLEY J, GARBADE M, MILIOTO A, et al. SemanticKITTI: a dataset for semantic scene understanding of lidar sequences[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019: 9297-9307.

[9] CAESAR H, BANKITI V, LANG A H, et al. NuScenes: a multimodal dataset for autonomous driving[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 11621-11631.

[10] FAN L, XIONG X, WANG F, et al. Rangedet: in defense of range view for lidar-based 3D object detection[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021: 2918-2927.

[11] SUN J, DAI Y, ZHANG X, et al. Efficient spatial-temporal information fusion for LiDAR-based 3D moving object segmentation [J]. arXiv preprint arXiv:2207.02201, 2022.

[12] TANG H, LIU Z, ZHAO S, et al. Searching efficient 3D architectures with sparse point-voxel convolution[C]. European Conference on Computer Vision. Springer, 2020: 685-702.

[13] BERMAN M, TRIKI A R, BLASCHKO M B. The lovász-softmax loss: a tractable surrogate for the optimization of the intersection-over-union measure in neural networks[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 4413-4421.

[14] LIN T Y, GOYAL P, GIRSHICK R, et al. Focal loss for dense object detection[C]. Proceedings of the IEEE International Conference on Computer Vision, 2017: 2980-2988.

[15] ZHANG Z. Improved adam optimizer for deep neural networks [C]. 2018 IEEE/ACM 26th International Symposium on Quality of Service (IWQoS). IEEE, 2018: 1-2.

[16] CORTINHAL T, TZELEPIS G, ERDAL AKSOY E. SalsaNext: fast, uncertainty-aware semantic segmentation of LiDAR point clouds [C]. International Symposium on Visual Computing. Springer, 2020: 207-222.

[17] ZHU X, ZHOU H, WANG T, et al. Cylindrical and asymmetrical 3D convolution networks for lidar segmentation[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021: 9939-9948.

[18] ZHANG Y, ZHOU Z, DAVID P, et al. PolarNet: an improved grid representation for online lidar point clouds semantic segmentation[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 9601-9610.

[19] YAN X, GAO J, ZHENG C, et al. 2DPASS: 2D priors assisted semantic segmentation on lidar point clouds[C]. Computer Vision-ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXVIII. Springer, 2022: 677-695.

[20] CHOY C, GWAK J, SAVARESE S. 4D Spatio-Temporal ConvNets: Minkowski convolutional neural networks[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019: 3075-3084.