

DOI: 10.3901/JME.2022.24.289

基于激光点云与图像融合的3D目标检测研究*

刘永刚^{1,2} 于丰宁¹ 章新杰² 陈 峥³ 秦大同¹

(1. 重庆大学机械与运载工程学院/机械传动国家重点实验室 重庆 400044;

2. 吉林大学汽车仿真与控制国家重点实验室 长春 130025;

3. 昆明理工大学交通工程学院 昆明 650500)

摘要: 目前基于激光雷达与摄像头融合的目标检测技术受到了广泛的关注,然而大部分融合算法难以精确检测行人、骑行者等较小目标物体,因此提出一种基于自注意力机制的点云特征融合网络。首先,改进Faster-RCNN目标检测网络以形成候选框,然后根据激光雷达和相机的投影关系提取出图像目标框中的视锥点云,减小点云的计算规模与空间搜索范围;其次,提出一种基于自注意力机制的Self-Attention PointNet网络结构,在视锥范围内对原始点云数据进行实例分割;然后,利用边界框回归PointNet网络和轻量级T-Net网络来预测目标点云的3D边界框参数,同时在损失函数中添加正则化项以提高检测精度;最后,在KITTI数据集上进行验证。结果表明,所提方法明显优于广泛应用的F-PointNet,在简单、中等和困难任务下,汽车、行人和骑行者的检测精度均得到较大的提升,其中骑行者的检测精度提升最为明显。同时,与许多主流的三维目标检测网络相比具有更高的准确率,有效地提高了3D目标检测的精度。

关键词: 激光雷达; 3D目标检测; 点云融合; 注意力机制; 深度学习

中图分类号: TG156

Research on 3D Object Detection Based on Laser Point Cloud and Image Fusion

LIU Yonggang^{1,2} YU Fengning¹ ZHANG Xinjie² CHEN Zheng³ QIN Datong¹

(1. State Key Laboratory of Mechanical Transmissions, College of Mechanical and Vehicle Engineering, Chongqing University, Chongqing 400044;

2. State Key Laboratory of Automotive Simulation and Control, Jinlin University, Changchun 130025;

3. Faculty of Transportation Engineering in Kunming University of Science and Technology, Kunming 650500)

Abstract: At present, 3D object detection based on the fusion of lidar and camera has received extensive attention. However, most fusion algorithms are difficult to accurately detect small target objects such as pedestrians and cyclists. Therefore, a feature fusion network based on the self-attention mechanism is proposed, which fully considers the local feature information to achieve accurate 3D object detection. Firstly, to reduce the spatial search range of the point cloud, the Faster-RCNN is improved to form a candidate box. Then, the frustum point cloud was extracted according to the projection relationship between the lidar and the camera. Secondly, a Self-Attention PointNet based on the self-attention mechanism is proposed to segment the original point cloud data within the scope of the frustum. Finally, while using the PointNet and T-Net to predict the 3D bounding box parameters, the regularization term is considered in the loss function to achieve higher convergence accuracy. The KITTI dataset is used for verification and testing. The results show that this method is obviously superior to F-PointNet and the detection accuracy of cars, pedestrians, and cyclists has been greatly improved, and it has higher accuracy than mainstream 3D object detection networks.

Key words: lidar; 3D object detection; point cloud fusion; attention mechanism; deep learning

* 国家自然科学基金(52172400)、汽车仿真与控制国家重点实验室开放基金(20201101)和重庆自主品牌汽车协同创新中心揭榜挂帅项目(2022CDJDX-004)资助项目。20220119收到初稿,20220926收到修改稿

0 前言

近年来自动驾驶汽车得到了迅速的发展,目标检测作为自动驾驶汽车感知中的一项基础性关键工作,对于保障自主车辆安全,提高智能汽车环境理解力具有重要意义^[1]。然而,单摄像头感知系统无法提供可靠的 3D 几何结构,复杂或恶劣的天气条件也限制了其全天候工作能力。与摄像头相比,激光雷达能够获取驾驶环境的三维信息且能够给出十分精确的空间位置,在障碍物检测、目标测距等方面具有无可比拟的优势。因此,利用传感器融合技术能够将摄像头与激光雷达结合起来,综合两者的优点,增强感知能力,实现精确的全天候环境感知。近年来,传感器融合技术已逐渐成为研究热点。

在三维目标检测方面,与传统的激光雷达点云目标检测算法相比,深度学习算法因避免了对点云数据进行手工特征提取等操作,在三维物体特征提取方面表现出优异的性能。基于深度学习的三维目标检测主要包括间接法、直接法和融合法^[2]。间接法的核心是先将点云进行体素化处理,再用来训练深度神经网络。WANG 等^[3]提出将点云数据转化成为含有特征向量的体素,通过在稀疏特征网络上基于投票算法进行计算,避免了卷积运算所需要的庞大计算量。ZHOU 等^[4]基于 VoxelNet 方法将点云划分为等间距的三维体素,并对体素内所有无序点云数据通过体素特征编码层继续特征提取。YAN 等^[5]在 VoxelNet 的基础上,提出了稀疏嵌入式卷积检测网络模型 SECOND,利用点云的稀疏性改善 VoxelNet 的特征提取效果,其他基于体素离散化的代表性方法还有 Voxel-FPN^[6]、Vote3Deep^[7]、3D FCN^[8]等。然而体素化方法在三维空间中划分网格对内存资源消耗大,而且划分标准往往对检测结果有很大影响,此外,点云的分布往往是稀疏的,直接对空网格进行卷积运算会造成不必要的计算资源浪费。

直接法则是直接利用原始点云数据在深度神经网络模型中进行训练。QI 提出可直接处理激光点云数据的 PointNet^[9]网络,并在后续研究中进行改进得到 PointNet++^[10]网络,该神经网络模型通过将采样层、组合层、特征提取层处理模块级联组合,获得目标的深层语义特征,保证了更准确的目标特征提

取。LI 等^[11]提出的 PointCNN 用了 K 近邻方法来学习点云中的空间信息,应用基于某种近邻方法的一维卷积让卷积网络更好地处理不规则和无序的点云数据。DENG 等^[12]提出 PPFNet,使用多个 PointNet 来生成易区分且抗旋转的局部特征,在准确率、鲁棒性以及抗平移旋转等方面都取得了不错的效果。其他基于原始点云数据处理的代表性方法还有 LaserNet^[13]、3DSSD^[14]等。然而上述 PointNet 网络只是简单地将所有点连接,没有局部特征信息,而基于 PointNet++架构的一系列网络,为了构建全局关系往往需要对点云进行分层,分层的过程又会引入新的信息损失,且这类方法处理大规模点云时计算消耗极高。

融合法以点云数据为主,图像数据为辅的方式进行目标检测,目标检测的平均精度比前述的两种处理方法要高。LI 等^[15]提出了 velo FCN 方法,采用基于二维投影的点云检测结构,将激光雷达数据呈现在二维点图上,并使用一个单一的端到端完全卷积网络来预测目标的置信度和边界框。CHEN 等^[16]提出了 MV3D 则采用基于多视图的融合检测结构,通过深度网络对各个视图进行特征提取和感兴趣区域提取,最后与深度融合方案结合起来进行目标检测。KU 等^[17]提出 AVOD 方法,使用更高分辨率的区域提议网络分别对点云、前视图和鸟瞰图中的目标生成各自的候选区域并进行融合处理,最终得到更精确的目标边界框。QI 等^[18]提出了 Frustum PointNet 网络,使用来自摄像头图像的 2D 检测结果来过滤点云并得到每个候选区域的点云,并使用 PointNet 对这些点云来执行分类和回归,该方法取得了较高的精度,其他基于点云数据与图像数据融合的代表性方法还有 F-ConvNet^[19]、MMF^[20]、ContFuse^[21]等。

综上所述,融合的方案能够结合图像和激光雷达的优势,有效提高 3D 目标检测的准确率,然而大部分融合算法不直接处理点云数据,而是通过把点云投影到特定视角的二维平面来提取特征,因此造成点云数据三维特性的丢失,这极大地影响了点云信息的空间关系。因此本文在基于原始点云数据处理的 F-PointNet 网络基础上提出了一种基于自注意力机制的点云特征融合网络,充分考虑点云空间局部特征信息以及不同点云间的特征的权重,并利用摄像头图像检测结果映射得到视锥体区域点云,缩小点云数据的空间搜索范围,减小计算规模,

大大提高行人、骑行者等较小物体的平均检测精度，实现精准的三维目标检测。

1 基于视觉的目标识别

在图像数据和点云数据融合之前，需要对感兴趣区域进行提取，所以首先需要对图像数据训练一个二维目标检测网络。Faster-RCNN 算法是由 REN 等^[22]提出的目标检测模型。在基础的图像特征提取方面，Faster-RCNN 在继承 Fast-RCNN 网络优点的基础上，将之前的全连接层替换为全卷积层，代替了选择性搜索，从而较大地提升了检测速度与算法分类的 mAP(Mean average precision)值以及对小目标检测的准确率。本文改进 Faster-RCNN 以预测最终 2D 检测边界框，在原始算法基础上融合了最新的特征金字塔网络(Feature pyramid networks, FPN)^[23]，实现低层特征高分辨率和高层特征的高语义信息的结合，提高检测精度。改进后的 Faster R-CNN 网络主要由多尺度特征提取网络、区域特征提取网络(Region proposal network, RPN)、ROI pooling 层、分类和回归层等 4 部分组成，如图 1 所示。算法流程为首先输入图像，经过 Resnet101 与 FPN 融合得到多尺度的卷积特征图，然后利用 RPN 在特征图上提取可能包含目标的区域候选框，最后 ROI pooling 部分对区域候选框进行池化操作，将不同尺寸的输入转换为固定尺寸的输出，再利用检测网络进行回归和分类处理，得到目标边框和得分。训练过程的试验步骤为首先在 ImageNet 分类和 COCO 目标检测数据集上对 2D 目标检测模型权重进行预训练，然后在 KITTI 2D 目标检测数据集上对其进行微调，以分类和预测 2D 目标边界框。

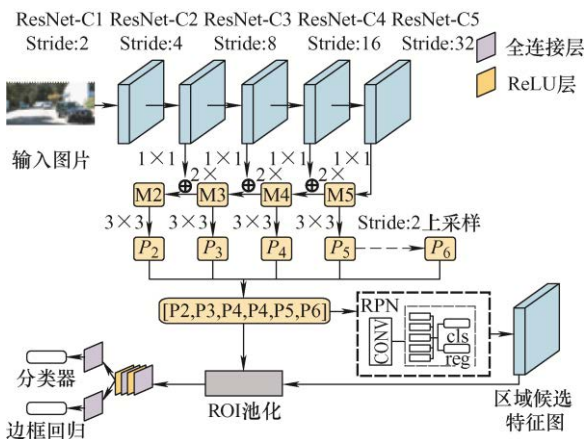


图 1 改进的 Faster-RCNN 网络结构

2 基于自注意力机制的点云特征提取网络设计

点云特征提取主干网络采用的 PointNet，其结构是直接使用原始点云数据作为输入的点云特征提取网络。在目标检测任务中，由于 PointNet 丢失了空间局部信息使得网络检测小目标物体和概括复杂场景的能力是有限的，不利于精确的三维检测。因此，我们在其基础上通过引入注意力机制^[24]构成了 Self-Attention PointNet，以达到很好地获取点云空间局部特征信息的目的。

其中 Self Attention Block 架构如图 2 所示，该架构基于 Multi-Head Attention 机制^[25]对点云进行编码，该机制是在原始 Attention 机制的基础上将 Q, K, V 经矩阵参数进行映射并对原始输入进行多次不共享参数的 Attention 操作，最终拼接输出结果。为了充分挖掘点云局部特征信息，我们利用自注意力机制实现模型。所谓的自注意力机制就是三部分输入设为同一矩阵 X ，为避免混淆，方便概念解释，下文仍使用 Q, K, V 代指 Self Attention Block 中的三部分输入。首先主要完成输入向量 Q, K, V 的线性映射，其次通过尺度缩放点积与 Softmax 归一化完成点云间特征的融合工作，最后通过残差将输入与 V 的值连接，残差学习可以提升网络的拟合能力，提高对点云数据的识别精度。具体的模型如下所示。

$$Y = \text{MultiHead}(X, X, X) \quad (1)$$

$$\text{MultiHead}(Q, K, V) = \text{head}_1 \oplus \text{head}_2 \oplus \dots \oplus \text{head}_m \quad (2)$$

$$\text{head}_i = \text{Attention}(QW_Q, KW_K, VW_V) \quad (3)$$

式中， X 为三维点云集合， Y 为对点云的编码结果， (W_Q, W_K, W_V) 为可学习的参数矩阵， \oplus 表示特征拼接。对某一点云 $x_i \in X$ 而言，其编码过程如式(4)所示

$$\text{Attention}(x_i, X, X) = \text{softmax} \left(\frac{QK^T}{\sqrt{d_k}} \right) \cdot V = \sum_{s=1}^m \frac{1}{Z} \exp \left(\frac{\langle x_i, x_s \rangle}{\sqrt{d_k}} \right) \cdot x_s \quad (4)$$

从中容易观察到每次对 x_i 进行编码的过程，其他的点云 x_s 也作为变量，影响到 x_i 的输入结果，得到的编码结果既蕴含本地信息又包括全局信息。

该点云特征提取网络的整体框架示意图如图 3 所示，输入点云的形式为 $(n \times 4)$ ，其中 n 为点云

的数量, 4 为点云的维度。首先点云数据通过一个共享参数的双层感知机模型进行特征提取, 其次通过特征空间变换矩阵预测网络实现对特征的对齐, 然后经搭建的 Self Attention Block, 通过

编码即可得到可用于点云分类的特征向量, 将得到的特征向量与最大池化后的全局特征进行拼接实现点云数据的实例分割, 并输出每个点云所属类别的评估分数。

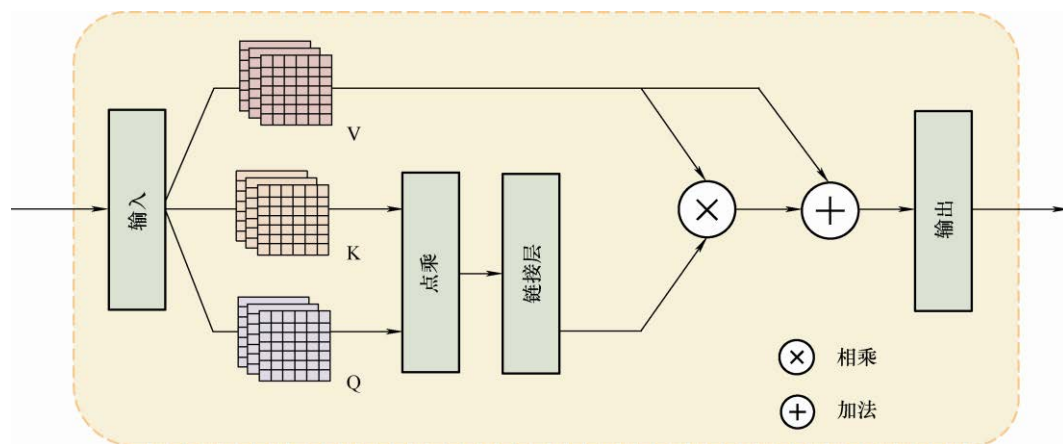


图 2 Self-Attention Block 结构

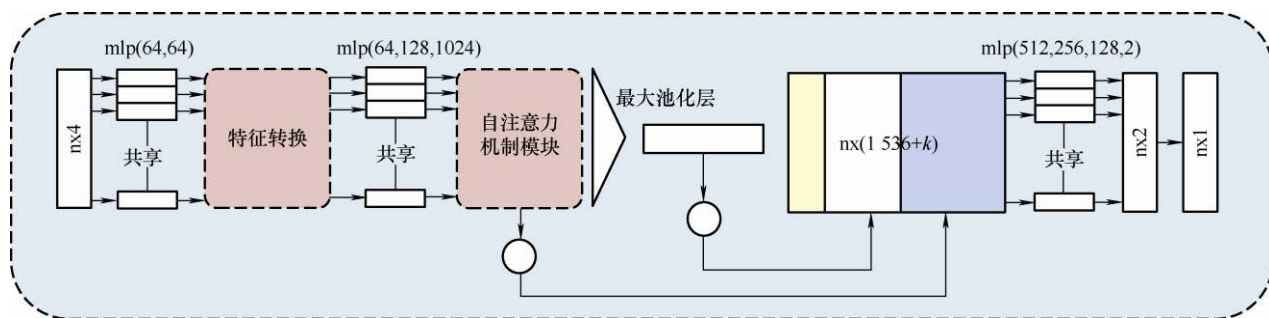


图 3 Self-Attention Pointnet 网络结构

3 基于特征融合的三维点云目标检测网络设计

3.1 整体网络架构

本文构建的基于自注意力机制的特征融合的三

维点云目标检测网络主要由坐标系校准与目标锥形候选区域提取、点云实例分割和非模态 3D 边界框参数预测等 3 个部分构成, 如图 4 所示。该网络使用 RGB 图像与原始点云数据作为输入, 首先利用上述改进后的 Faster-RCNN 算法在给定图像中生成目标检测框, 实现二维目标检测, 并根据激光雷达和

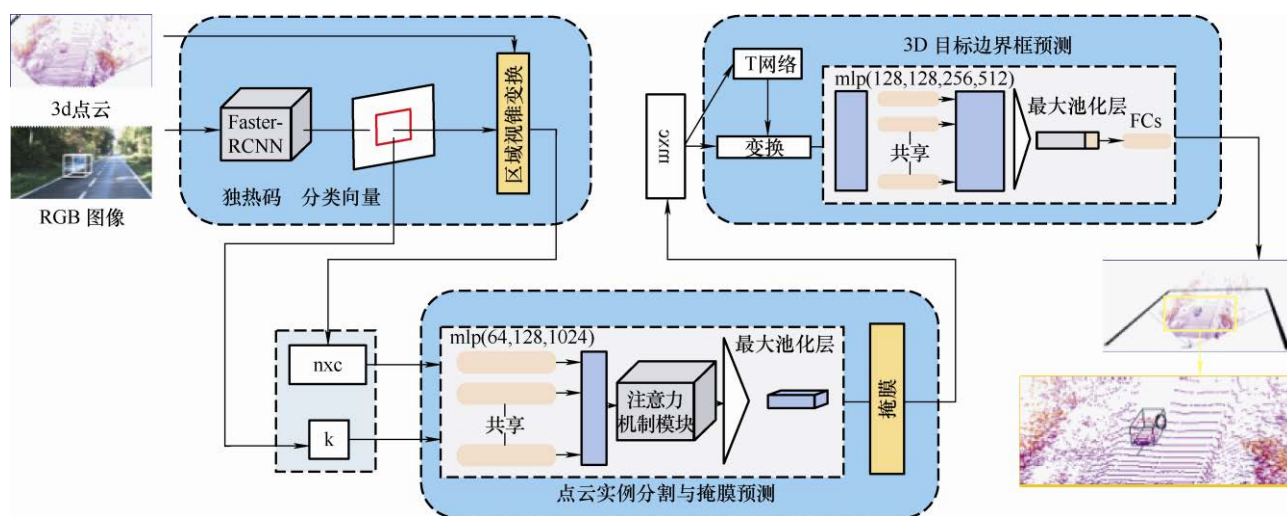


图 4 基于特征融合的三维点云目标检测网络结构

相机的投影关系提取出图像目标的视锥点云数据; 其次通过基于自注意力机制的 Self-Attention PointNet 网络预测每个点所属类别的评估分数并提取目标实例的点云; 最后由 3D 边界框评估网络输出目标的三维边界框的中心坐标、尺寸和航向角, 同时在损失函数中添加正则化项以提高检测精度, 实现在三维环境下准确的目标检测。

3.2 坐标系校准与目标锥形候选区域提取

在对目标视锥候选区域点云进行提取之前, 需要校准图像与点云数据, 具体坐标系校准方程如式(5)所示^[26], 首先通过刚体变换将激光雷达坐标系中的点 (X_w, Y_w, Z_w) 变换到相机坐标系下的 (X_c, Y_c, Z_c) , 其次通过小孔成像原理中的中心透视投影法实现相机坐标系到图像坐标系之间的转换, 最后通过伸缩和平移变换实现图像坐标系到像素坐标系的转换, 从而完成激光雷达点云到相机图像的空间对齐与配准, 如图 5 所示。

$$Z_c \begin{pmatrix} u \\ v \\ 1 \end{pmatrix} = \begin{pmatrix} f_x & 0 & u_0 & -f_x b_x \\ 0 & f_y & v_0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} \mathbf{R} & \mathbf{T} \\ 0 & 1 \end{pmatrix} \begin{pmatrix} X_w \\ Y_w \\ Z_w \\ 1 \end{pmatrix} \quad (5)$$

式中, \mathbf{R} 和 \mathbf{T} 分别表示相机坐标系相对于激光雷达坐标系的旋转和平移矩阵, f_x 和 f_y 分别为相机横、纵焦距, u_0 和 v_0 为图像原点坐标, u 和 v 为像素坐标系下的坐标, b_x 表示在 KITTI 数据集中目标相机坐标系相对于参考相机坐标系的偏移。

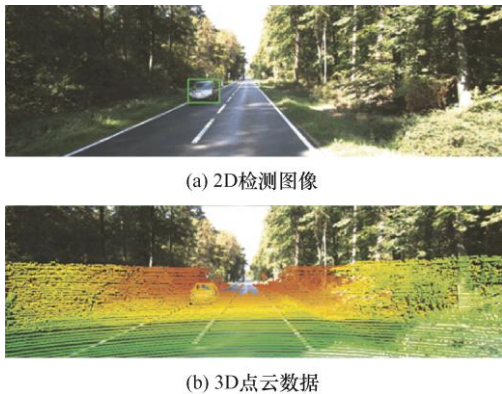


图 5 2D 检测图像与 3D 点云数据的校准效果图

采用基于改进后 Faster-RCNN 的 2D 目标检测模型, 获取 RGB 图像中的 2D 目标边界框, 对于一个二维检测框, 将其四个角点通过坐标系转换映射为三维世界中的四条射线, 以此确定一个顶点为相机光心的四面锥, 收集视锥中的所有点以形成视锥体点云, 通过将输入三维定位模块

的点云限制为一个视锥, 以此来减少三维网络的学习难度与计算规模, 提高三维信息的预测精度, 如图 6 就展示了视锥切割的效果。由于每个提取出的视锥点云在相机坐标系下拥有不同的方向, 为了便于对点云数据进行处理, 将视锥朝向中心视角旋转使得视锥体的中心轴正交于图像平面, 从而实现其从相机坐标系到视锥坐标系下的转换, 如图 7 所示。

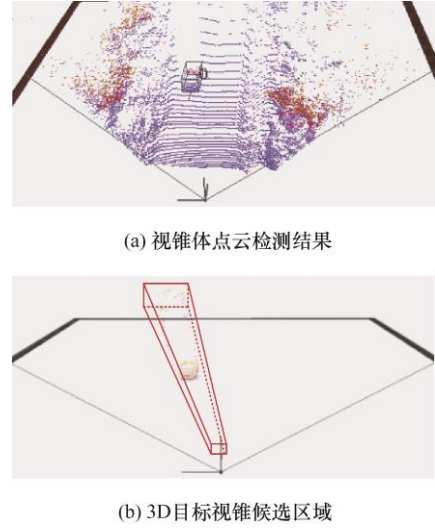


图 6 视锥切割与目标视锥候选区域提取图

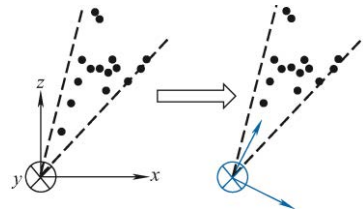


图 7 视锥朝向调整图

3.3 点云实例分割与 3D 目标边界框预测

通过构建的基于自注意力机制的 Self-Attention PointNet 实现视锥点云的实例分割, 首先将锥形区域提取出的点云数据作为输入, 结合二维目标检测部分的 one-hot 分类向量输出每个点云所属类别的评估分数, 其次通过掩模操作剔除背景、杂乱点云等非目标点云, 并利用掩模结果提取目标实例的点云。为提高算法的平移不变性, 我们将获得的目标点云坐标做进一步归一化处理, 将所有的目标点云减去其质心的坐标, 从而形成在掩模坐标系下的点云数据, 并使用与 STN (Spatial transformer network)^[27]相似的 T-Net 网络通过残差回归的方式进一步调整质心的位置, 从而将目标实例点云从掩模坐标系转换至 3D 目标边界框中心坐标系下, 转换过程如图 8 所示。

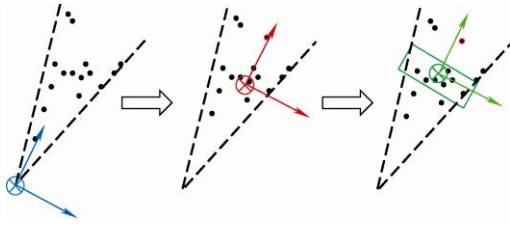


图 8 目标点云坐标系转换

最后利用 3D 边界框回归 PointNet 网络以及多层感知机处理实现目标实例点云的边界框回归预测,其中回归参数包括 3D 目标边界框的中心坐标、边界框的长宽高、边界框长宽高残差、目标的航向角和目标航向角残差。为了轻量化该网络结构,对此模块中 T-Net 中多层感知器的卷积核数进行了相应地修改,如图 9 所示。

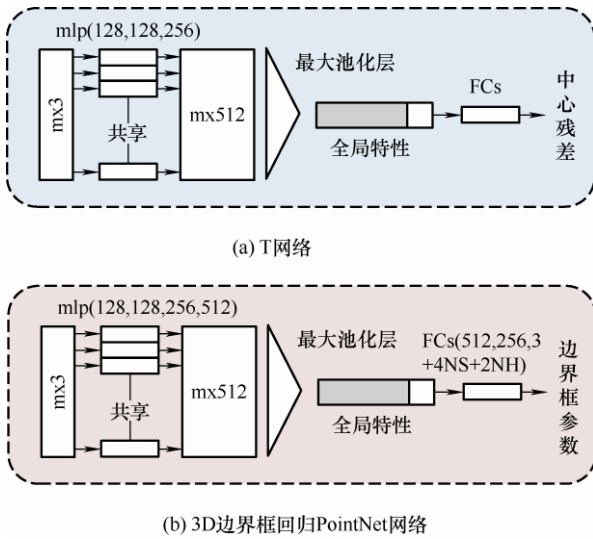


图 9 T-Net 网络与边界框回归网络结构

3.4 多任务损失函数与正则化

本文提出的改进后的网络模型损失函数分为分类损失和回归损失。我们同时优化了涉及的三种具有多任务损失的神经网络,包括改进的 Self-Attention PointNet 的点云特征提取网络,以及轻量级 T-Net 网络和非模态 3D 边界框回归 PointNet 网络。则总体网络模型的损失函数为

$$L_{multi-task} = L_{seg} + \alpha (L_{c1-reg} + L_{c2-reg} + L_{h-cls} + L_{h-reg} + L_{s-cls} + L_{s-reg} + \beta L_{corner}) \quad (6)$$

$$L_{corner} = \sum_{i=1}^{NS} \sum_{j=1}^{NH} L_{\delta} \min \left\{ \sum_{k=1}^8 \|P_k^{ij} - P_k^*\|, \sum_{k=1}^8 \|P_k^{ij} - P_k^{**}\| \right\} \quad (7)$$

式中, $L_{multi-task}$ 为整个网络的损失; α 和 β 为模型参数, 分别设为 1 和 10; L_{seg} 为 Self-Attention PointNet 网络产生的实例分割损失; L_{c1-reg} 为轻量

级 T-Net 网络产生的质心回归损失; L_{c2-reg} 为非模态 3D 边界框回归 PointNet 网络的质心回归损失; L_{h-cls} 和 L_{h-reg} 分别为网络模型产生的航向角分类损失和回归损失; L_{s-cls} 和 L_{s-reg} 分别为网络模块预测 3D 边界框产生的分类损失和回归损失; L_{corner} 为预测的 3D 边界框 8 个角的损失之和, 上述中对于分类损失, 使用 Softmax 交叉熵损失函数, 在回归损失中使用 Smooth- l_1 损失函数。

由于改进后的网络模型对点云数据的学习能力很强, 在训练中为了避免出现过拟合现象, 在原 $L_{multi-task}$ 中增加一项权重衰减项, 采用了几组不同的正则化系数进行训练, L_2 正则化如式(8)所示

$$L_{total_loss} = L_0 + \frac{\lambda}{2n} \sum w^2 \quad (8)$$

式中, L_{total_loss} 为添加权重衰减项之后的总损失; L_0 为增加 L_2 正则化之前的损失; $\frac{\lambda}{2n} \sum w^2$ 为增加的 L_2 正则化损失项, 其中 w 为神经网络中的权重, λ 为 L_2 正则化系数, n 为训练集中 batchsize 的大小。

4 试验验证与分析

4.1 试验数据集与环境配置

本文使用 KITTI 公开数据集^[26]进行性能评估, KITTI 3D 目标检测数据集包括了不同时间下室外多种场景的相机图像和 64 线激光雷达的点云数据, 包含 7 481 帧训练数据和 7 518 帧测试数据, 对 7 481 帧训练数据进行约 1:1 划分, 其中 3 712 帧用于训练、3 769 帧用于验证, 并在划分出的验证集上完成算法性能的评估。在训练数据集根据目标在相机视场内的遮挡和成像大小划分为简单(完全可见)、中等(部分遮挡)和困难(较难看到)3 种等级。

本文网络模型的训练和测试过程的试验环境基于 Linux Ubuntu 16.04 操作系统、Intel i7 8700k CPU、内存 12 核 32 GB、GTX 1080Ti GPU, 运算平台为 CUDA 10.0, 采用 Cudnn 7 作为网络的 GPU 加速库, 深度学习框架为 Tensorflow-GPU, 版本号 1.13.1。

4.2 试验训练参数分析

在试验中, 本网络模型训练过程中的主要评价指标是 IoU(Intersection over union)、汽车、行人与骑行人类别的阈值分别设置为 0.7、0.5、0.5。其中, 模型训练最优化方法采用 ADAM 算法, 设置动量为 0.9, 初始学习率为 0.001, 学习率衰减系数为 0.5,

每 60k 次迭代衰减一半；网络参数初始化采用 Xavier 优化器，模型的批处理大小设置为 32，最大迭代次数设置为 181。由于 KITTI 属于数据量中等的数据集，为提高对数据的利用，对部分点云进行数据增强。首先，对视锥提取模块提取到的视锥点云沿着 Z 轴，即深度方向，做稍微的扰动，从而增加点云的深度来增加训练数据数量；其次，在相机坐标(Z 为向前，Y 为向下)中沿着 YZ 平面对视锥点云做随机镜像翻转。

4.3 试验结果及分析

4.3.1 目标检测精度对比分析。

客观评价指标主要对给定交并比 (IOU=0.7,0.5,0.5) 阈值，分别计算不同准确率 (Precision) 和召回率 (Recall) 的对应关系得到 P-R 曲线，计算公式如式(9)~(10)所示，其中 X_{TP} 表示正确检出的目标， X_{FN} 表示漏检的目标， X_{FP} 表示误检的目标。然后按照 KITTI 官方要求通过对准确率

召回率的分段积分计算不同难度系数下的平均准确率(mAP)。

$$Precision = \frac{X_{TP}}{X_{TP} + X_{FP}} \quad (9)$$

$$Recall = \frac{X_{TP}}{X_{TP} + X_{FN}} \quad (10)$$

为确定最佳权重衰减项，进行了多次试验，分别设置正则化系数 λ 为 0.01、0.001、0.000 1，如表 1 所示。当 $\lambda=0.01$ 时，对骑行人的检测精度最高；当 $\lambda=0.001$ 时，对行人的检测精度最高；而 λ 为 0.000 1 时，对汽车的检测精度最高，但是对行人和骑行人的检测精度都比未加衰减率时低。综合考虑，选择正则化系数为 0.001，其对汽车和骑行人的检测精度都比未加衰减率时高，且对目标检测整体精度的提升较为可观。

表 1 不同衰减率下的目标检测精度

正则化系数	Car			Pedestrian			Cyclist		
	Easy	Moderate	Hard	Easy	Moderate	Hard	Easy	Moderate	Hard
0	82.18	68.93	61.02	64.45	55.36	48.43	68.52	50.53	47.14
0.01	83.75	69.17	62.38	64.54	55.71	48.75	71.11	53.36	50.29
0.001	82.93	69.16	61.95	67.54	57.62	50.84	68.61	51.83	48.12
0.0001	84.25	69.75	63.07	62.89	54.54	48.09	67.82	52.51	48.32

同时，为了验证各个处理部分对原网络的影响程度，进行了如表 2 的对比试验。可以看出添加正则化项对汽车和行人三种难度下的检测精度均略有提高，对行人的检测精度有较高的提升，三种难度下分别增加了 3.09%，2.26%，2.41%；添加提出的 Self-Attention PointNet 网络后，汽车和行人的检测精度出现较好的提升效果，其检测精度在困难难度下分别从 61.02% 增加到 63.54%，从 48.43% 增加到 51.25%，其中骑行人的检测精度提升明显，在三种难度下分别增

加了 8.52%，7.38%，7.02%，进一步证明了本文提出的基于自注意力编码机制的点云特征网络能够有效地结合点云空间的局部特征信息；对于整合后的网络模型，汽车和行人的检测精度与原模型比均有很好的提高，特别是骑行人的检测精度提升明显，在三种难度下分别增加了 12.25%，8.08%，7.23%，可见在添加 Self-Attention PointNet 网络基础上引入正则化项可以实现更高精度的收敛。综合所有结果来看，本文提出的网络模型有效地提高了 3D 目标的检测精度。

表 2 各处理部分对 3D 目标检测 AP 值的影响

正则化损失	Self-Attention	Car			Pedestrian			Cyclist		
		Easy	Moderate	Hard	Easy	Moderate	Hard	Easy	Moderate	Hard
—	—	82.18	68.93	61.02	64.45	55.36	48.43	68.52	50.53	47.14
✓	—	82.93	69.16	61.95	67.54	57.62	50.84	68.61	51.83	48.12
—	✓	84.79	71.50	63.54	67.18	58.15	51.25	77.04	57.91	54.16
✓	✓	84.10	71.01	63.39	67.07	58.08	51.21	80.77	58.61	54.37

最后，将本文模型与近年来主流的 3D 目标检测模型(仅汽车类别)进行比较，不同模型的目标检测结果如表 3 所示，表中的数据来自文献

[4,16,18,28,29]以及本文模型训练的试验检测结果。F-PointNet(v2) 表示主干特征提取网络是基于 PointNet++ 架构，从表中可知与原 F-PointNet 系列

网络相比, 本文模型提升了整体检测精度; 与基于 3D 点云和 RGB 图像融合的 PointFusion 相比, 本文模型在三种难度下的检测精度均表现出色; 与基于融合图像与点云框架的多视角聚合投影的 MV3D、RT3D 网络相比, 检测精度有明显的提升; 与基于点云数据体素网格化的 VoxelNet 网络相比, 三种难度下的准确率均得到不错的提高。

表 3 本模型与其他模型的 3D 目标检测 AP 值对比
(仅汽车类别)

Method	Car		
	Easy	Moderate	Hard
F-PointNet(v2)	83.76	70.92	63.56
PointFusion	77.92	63.00	53.27
RT3D	72.85	61.64	64.38
MV3D	71.29	62.68	56.56
VoxelNet	81.97	65.46	62.85
Ours	84.79	71.50	63.54

4.3.2 试验结果可视化分析。

试验过程中得到的总损失函数收敛曲线和点云实例分割的准确率曲线如图 10 所示, 从图 10a 可知, 当训练集的权重更新次数超过 15 000 次, 损失值在 2.2 左右波动, 并趋于平缓, 从图 10b 可看出点云实例分割的准确率在权重更新超过 20 000 次后也稳定在 95% 左右, 可见本文的网络模型具有快速、良好的收敛性能。

经过改进后的网络模型输出的部分 3D 目标边界框预测结果如图 11 所示, 其中上半部分为二维图像目标检测的可视化结果, 下半部分为点云中的检测结果, 红色框表示真实的边界框, 绿色框表示网络预测的边界框。从图中可以看出对于场景中合理距离的非遮挡物体, 本文的网络可以输出非常精确的边界框。从图 11b、11e 与图 11f 可以看出, 即使在图片中右下角汽车只能看到小部分, 在点云中仍旧能精确地检测出目标物体的边界和航向; 从

图 11c 中可以看到提出的网络模型在复杂的有大量行人的环境下依旧能够输出很不错的结果; 而从图 11b 中可以看到有漏检、失效的现象, 这是由于图片中存在不同程度的遮挡导致二维检测没有识别出目标, 且受遮挡处几乎没有点云, 导致目标边界框最终预测不准。

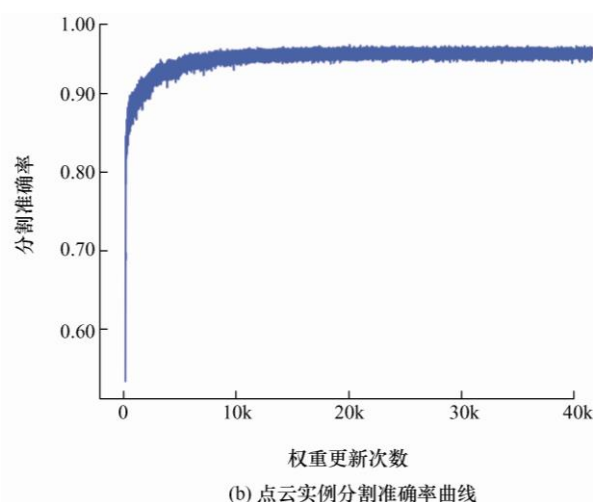
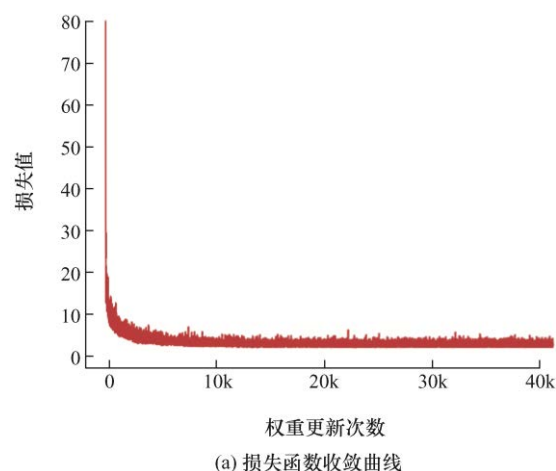
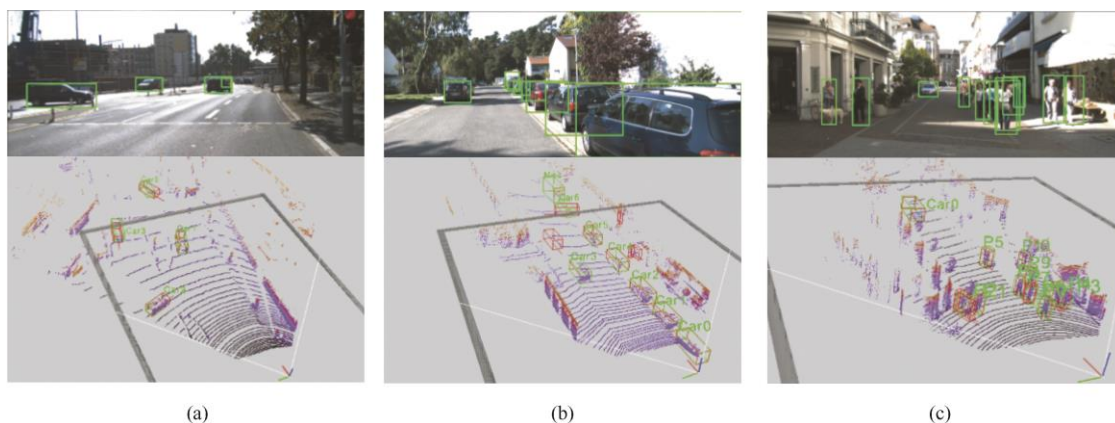


图 10 3D 目标检测训练过程中的损失函数收敛曲线和测试准确率曲线



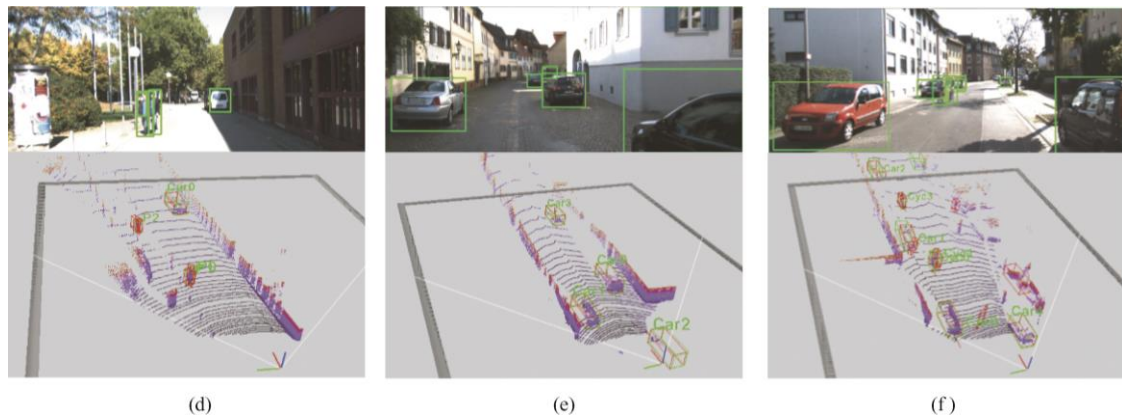


图 11 KITTI 数据集上检测结果可视化

5 结论

本文研究了一种基于深度学习的摄像头和激光雷达的特征融合算法，并针对处理点云数据过程中存在信息损失过多与检测小目标物体时精度较低的问题，重点构建了一种基于自注意力机制的点云特征提取网络结构，充分考虑点云空间局部特征信息以实现精准的三维目标检测。在 KITTI 数据集上的试验结果表明，本文研究的算法相对于其他方法具有更高的精度，特别是对小目标物体的点云处理，3D 目标检测提升效果更好。然而由于网络的点云目标检测模块依赖于图像的二维检测结果，这就要求图像的目标检测模块具有足够强的检测能力，后续的研究中将进一步改善网络结构，避免这些问题造成网络整体性能的损失。

参 考 文 献

- [1] 薛培林, 吴愿, 殷国栋, 等. 基于信息融合的城市自主车辆实时目标识别[J]. 机械工程学报, 2020, 56(12): 165-173.
XUE Peilin, WU Yuan, YIN Guodong, et al. Real-time target recognition of urban autonomous vehicles based on information fusion[J]. Chinese Journal of Mechanical Engineering, 2020, 56(12): 165-173.
- [2] 彭育辉, 郑玮鸿, 张剑锋. 基于深度学习的道路障碍物检测方法[J]. 计算机应用, 2020, 40(8): 2428-2433.
PENG Yuhui, ZHENG Weihong, ZHANG Jianfeng. Road obstacle detection method based on deep learning[J]. Journal of Computer Applications, 2020, 40(8):

2428-2433.

- [3] WANG D L, POSNER I. Voting for voting in online point cloud object detection[C]//Robotics: Science and Systems Xi, Sapienza Univ Rome: MIT PRESS, 2015: 13-22.
- [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(CVPR), Salt Lake City UT: IEEE Comp Soc, 2018: 4490-4499.
- [5] YAN Yan, MAO Yuxing, LI Bo. SECOND: Sparsely embedded convolutional detection[J]. Sensors, 2018, 18(10): 3337-3354.
- [6] 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): 704-723.
- [7] ENGELCKE M, RAO D, ZENG D, et al. Vote3Deep: fast object detection in 3d point clouds using efficient convolutional neural networks[C]//2017 IEEE International Conference on Robotics and Automation (ICRA), Singapore: IEEE, 2017: 1355-1361.
- [8] LI B. 3D fully convolutional network for vehicle detection in point cloud[C]//2017 IEEE/RSJ International Conference on Intelligent Robots and Systems(IROS), Vancouver: IEEE, 2017: 1513-1518.
- [9] QI C R, SU Hao, MO Kaichun, et al. PointNet: Deep learning on point sets for 3d classification and segmentation[C]//30th IEEE Conference on Computer Vision and Pattern Recognition(CVPR), Honolulu: IEEE, 2017: 77-85.
- [10] QI C R, YI Li, SU Hao, et al. PointNet plus plus : Deep

- hierarchical feature learning on point sets in a metric space[C]//Proceedings of Advances in Neural Information Processing Systems 30, Long Beach CA: NIPS, 2017: 5099-5108.
- [11] LI Yangyan, BU Rui, SUN Mingchao, et al. PointCNN: Convolution on x-transformed points[C]//Proceedings of Advances in Neural Information Processing Systems 31, Montreal: NIPS, 2018: 820-830.
- [12] DENG Haowen, BIRDAL T, ILIE S, et al. PPFNet: Global context aware local features for robust 3D point matching[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR), Salt Lake City UT: IEEE, 2018: 195-205.
- [13] MEYER G P, LADDHA A, KEE E, et al. LaserNet: An efficient probabilistic 3D object detector for autonomous driving[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR), Long Beach CA: IEEE, 2019: 12669-12678.
- [14] 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(CVPR), Seattle: IEEE, 2020: arXiv: 2002.10187.
- [15] LI Bo, ZHANG Tianlei, XIA Tian. Vehicle detection from 3 D lidar using fully convolutional network[C]//Proceedings of Robotics: Science and Systems (RSS), Ann Arbor: MIT PRESS, 2016: 42-50.
- [16] CHEN Xiaozhi, MA Huimin, WAN Ji, et al. Multi-view 3 D object detection network for autonomous driving[C]//30th IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR), Honolulu: IEEE, 2017: 6526-6534.
- [17] 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), Madrid: IEEE, 2018: 5750-5757.
- [18] 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(CVPR), Salt Lake City UT: IEEE, 2018: 918-927.
- [19] WANG Zhixin, JIA Kui. Frustum convNet: sliding frustums to aggregate local point-wise features for amodal 3D object detection[C]//IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Macau: IEEE, 2019: 1742-1749.
- [20] 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 (CVPR), Long Beach CA: IEEE, 2019: 7337-7345.
- [21] LIANG Ming, YANG Bin, WANG Shenlong, et al. Deep continuous fusion for multi-sensor 3D object detection[C]//15th European Conference on Computer Vision (ECCV), Munich: Springer-Verlag Berlin, 2018: 663-678.
- [22] REN Shaoqing, HE Kaiming, GIRSHICK R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence(TPAMI), 2016, 36(6): 1137-1149.
- [23] LIN T Y, DOLLAR P, GIRSHICK R, et al. Feature pyramid networks for object detection[C]//30th IEEE Conference on Computer Vision and Pattern Recognition(CVPR), Honolulu: IEEE, 2017: 936-944.
- [24] WOO S, PARK J, LEE J Y, et al. CBAM: Convolutional block attention module[C]//15th European Conference on Computer Vision (ECCV), Munich: SPRINGER-VERLAG BERLIN, 2018: 3-19.
- [25] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Proceedings of Advances in Neural Information Processing Systems 30, Long Beach CA: NIPS, 2017: 1049-1064.
- [26] GEIGER A, LENZ P, STILLER C, et al. Vision meets robotics: The kitti dataset[J]. International Journal of Robotics Research, 2013, 32(11): 1231-1237.
- [27] JADERBERG M, SIMONYAN K, ZISSERMAN A, et al. Spatial transformer networks[C]//Proceedings of Advances in Neural Information Processing Systems 28, Montreal: NIPS, 2015: 2017-2025.
- [28] XU Danfei, ANGUELOV D, JAIN A. PointFusion: deep sensor fusion for 3d bounding box estimation[C]//31st IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City UT: IEEE, 2018: 244-253.
- [29] ZENG Yiming, HU Yu, LIU Shice, et al. RT3D:

Real-time 3D vehicle detection in lidar point cloud for autonomous driving[J]. IEEE Robotics And Automation Letters, 2018, 3(4): 3434-3440.

作者简介：刘永刚(通信作者)，男，1982 年出生，博士，教授，博士研究生导师。主要研究方向为智能汽车决策与控制关键技术、新能源汽车动力系统优化与控制、车辆自动变速传动及综合控制。

E-mail: andyliuyg@cqu.edu.cn

于丰宁，男，1997 年出生，硕士研究生。主要研究方向为智能汽车激光雷达 3D 目标检测。

E-mail: yufengning@cqu.edu.cn

章新杰：男，1984 年出生，博士，教授，博士研究生导师。主要研究方向为车辆动力学及控制、智能运载测试与评价、驾驶员模型。

E-mail: x_jzhang@jlu.edu.cn

陈峥，男，1982 年出生，博士，教授，博士研究生导师。主要研究方向为动力电池管理、智能车辆控制及混合动力汽车能量管理。

E-mail: chen@kust.edu.cn

秦大同，男，1956 年出生，博士，教授，博士研究生导师。主要研究方向为机械传动系统、车辆动力传动及其智能控制。

E-mail: dtqin@cqu.edu.cn