计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2021,30(12):211−217 [doi: 10.15888/j.cnki.csa.008249]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

# 基于注意力机制的三维点云车辆目标检测①

彭玉旭, 董胜超

(长沙理工大学 计算机与通信工程学院, 长沙 410114)
通讯作者: 董胜超, E-mail: Dongscc@163.com

**摘　要**: 针对自动驾驶场景下三维点云车辆的识别和定位问题, 提出了一种基于注意力机制的三维点云车辆目标检测算法. 算法将稀疏无序的点云空间划分成等距规则的体素表示, 用三维稀疏卷积和辅助网络同步从所有体素中提取内部点云特征, 进而生成鸟瞰图. 但在将内部三维的点云特征转化为二维的鸟瞰图后, 通常会造成目标空间特征信息丢失, 使得最终检测结果以及方向性预估差. 为进一步提取鸟瞰图中特征信息, 提出了一种注意力机制模块, 其中包含两种注意力模型, 并对其采用首、中、尾的“立体式”布局结构, 实现对鸟瞰图中特征信息的放大和抑制, 最后使用卷积神经网络和 PS-Warp 变换机制对处理过后的鸟瞰图进行三维目标检测. 实验表明, 该算法在保证实时检测效率的前提下, 与现有算法相比, 具有更好的方向预估性以及更高的检测精度.

**关键词**: 三维点云; 车辆目标; 注意力机制; 目标检测

引用格式: 彭玉旭,董胜超.基于注意力机制的三维点云车辆目标检测.计算机系统应用,2021,30(12):211–217. http://www.c-s-a.org.cn/1003-3254/8249.html

## 3D Point Cloud Vehicle Target Detection Based on Attention Mechanism

PENG Yu-Xu, DONG Sheng-Chao

(School of Computer and Communication Engineering, Changsha University of Science and Technology, Changsha 410114, China)

**Abstract**: In this study, a 3D point-cloud target detection algorithm for vehicles based on attention mechanism is proposed for the recognition and positioning of the targets in autonomous driving scenarios. The algorithm first divides the sparse and disordered point cloud space into equidistant and regular voxel representations. Then, 3D sparse convolution and auxiliary network are used to synchronously extract the internal point cloud features from all voxels. Afterward, a bird's-eye view is generated. After the internal 3D point cloud features are converted into a 2D bird's-eye view, the spatial feature information of the target will be lost generally, which makes the final detection result and the direction prediction unsatisfactory. To further extract the feature information of the bird's-eye view, this study also proposes an attention mechanism module, which contains two attention models and adopts a three-dimensional layout structure (front, middle, and back) to realize amplification and suppression of the feature information of the bird's-eye view. The convolutional neural network and PS-Warp transformation mechanism are employed to perform 3D target detection on the processed bird's-eye view. Experiments show that, under the premise of ensuring real-time detection efficiency, this algorithm has better direction prediction and higher detection accuracy than existing algorithms.

**Key words**: 3D point cloud; vehicle target; attention mechanism; target detection

① 基金项目: 长沙理工大学青年教师成长计划 (2019QJCZ014)
Foundation item: Young Teacher Growth Plan, Changsha University of Science and Technology (2019QJCZ014)
收稿时间: 2021-03-22; 修改时间: 2021-04-19; 采用时间: 2021-04-26

## 1 引言

从点云数据中进行 3D 目标检测是自动驾驶系统重要组成部分, 例如自主导航、车辆检测和路障检测. 与基于图像的检测相比, 激光雷达所采集的点云数据提供了可靠的深度信息, 在空间目标位置判定以及方向和姿态估计上更具优势. 与仅从图像中估计 2D 边界框的普通 2D 检测不同, 自动驾驶系统需要从现实世界中估计更加具有信息量的 3D 边界框, 以完成诸如避免障碍物和路径规划之类的高级任务. 这一严峻的挑战以及自动驾驶市场日益增长的需求激发了 3D 车辆目标检测算法的研究, 新近出现的检测算法主要采用卷积神经网络处理来自激光雷达传感器所采集的深度点云数据.

基于点云的 3D 车辆目标检测算法可分为单阶段检测算法和两阶段检测算法. 单阶段检测算法[1–3] 将稀疏的 3D 点云转换为规则的表现形式, 例如体素化网格或者鸟瞰图像, 并采用 CNN 以完全卷积的方式直接预测边界框. 这使得单阶段检测方法简单快速且有效. 但是直接将卷积网络生成的点云特征转化为鸟瞰图后, 不可避免的会导致空间特征的丢失, 使得单阶段检测器处理稀疏点云数据的准确性降低. 与单阶段检测算法相比, 两阶段检测算法[4–8] 可以在第二阶段利用更精确的空间信息, 这些信息仅专注于第一阶段预测的感兴趣的区域, 因此使得边界框的预测更加准确. 但第二阶段的检测是在第一阶段检测的结果上进行, 这就大大增加了计算成本, 使得实时的检测速度不理想.

为了解决两阶段检测器检测速度慢以及单阶段检测器会导致空间特征丢失的问题. 本文提出了利用注意力机制对数据特征的表征能力进行增强, 并采用单阶段检测方法, 以提高算法定位精度并同时保持单阶段检测方法的高效性. 在 KITTI 3D 目标检测基准[9] 上评估了本文提出的算法, 并与现有的方法进行比较, 在确保高效的实时检测速度的情况下, 本文提出的检测器较现有方法具有更好的方向预估性以及更高的检测精度.

## 2 相关工作

### 2.1 单阶段检测方法

单阶段检测方法通常会把稀疏的点云场景空间转换为更加规则的表现形式, 然后用卷积神经网络来处理点云数据以提高计算效率. 这种方法要么使用 3D CNN 处理基于手工的网格[10,11], 要么使用 2D CNN 从鸟瞰图和前视图全景图[12] 中提取特征. 文献 [4] 提出多层体素特征编码结构以提取体素中每个点的特征. 文献 [3] 将点云数据沿高度轴堆叠的体素改为支柱进行特征提取, 然后用卷积神经网络进行目标的定位. 文献 [1] 利用子流形稀疏卷积研究出一种稀疏三维卷积, 可优化三维卷积中的 GPU 使用, 并提升网络的检测速度和效率. 文献 [13] 提出一种辅助网络, 用点级监督的方式来增强稀疏卷积特征表示能力. 本文提出的方法就建立在通用的单阶段体系结构的基础上.

### 2.2 两阶段检测方法

两阶段检测方法是使用第二阶段的网络结构从第一阶段生成的区域提案中进行更精确的检测. 文献 [6] 提出的检测器在第一阶段将整个场景的点云分割为前景点和背景点, 以自下而上的方式直接从点云中生成少量精确 3D 提案. 文献 [5] 在输入的 RGB-D 图像上使用 Mask RCNN 网络找到一组感兴趣的区域, 结合该区域的深度信息后得到平截头体状点云提案, 然后使用 PointNet[14] 对该提案进行三维实例分割以得到更加精确的 3D 提案空间. 文献 [7] 第 1 阶段通过为每个点生成一个新的球形锚来产生准确的提案, 利用 PointsPool 生成提案特征, 第 2 阶段在盒内预测中设计一个 3D loU 预测分支以提高定位精度. 两阶段检测方法通可以得到更加优秀的提案, 从而得到更加精确的检测结果, 相对于单阶段检测方法来说具有先天的优势, 但两阶段检测方法所花的时间成本不容忽视. 最近的单阶段检测方法已经达到了与最新的两阶段检测方法相当的性能, 高效的单阶段检测方法在实时应用中具有巨大的潜力.

### 2.3 注意力模型

注意力模型最近几年在语音识别、图像处理和自然语言处理等深度学习领域中应用广泛. 注意力模型借鉴了人类视觉的注意力机制. 人类视觉通过快速扫描周围环境场景. 获得需要重点关注的目标区域, 然后对该区域集中投入注意力资源, 以获取所需要关注目标更多的细节信息, 对其它信息忽略不予关注. 注意力机制从本质上来说和人类的视觉注意力机制类似, 其核心思想也是从众多信息中选择出对当前任务目标更关键的信息. 例如$F \in R^{C\times H\times W}$为某一中间特征图的输入, 注意力模型可分为:

$$M_C \in R^{C\times 1\times 1} \tag{1}$$

$$M_S \in R^{1\times H\times W} \tag{2}$$

式 (1) 为 1D 通道注意力图, 式 (2) 为 2D 空间注意力图. 整个注意力模型的关注过程可以概括为:

$$F' \in M_C(F) \otimes F \tag{3}$$

$$F'' \in M_S(F') \otimes F' \tag{4}$$

式中, $\otimes$表示逐元素相乘; $F'$为特征图 $F$ 经过通道注意力处理过后生成的数据; $F''$为 $F'$经过空间注意力处理过后生成的数据.

## 3 基于注意力机制的单阶段 3D 车辆检测

单阶段 3D 车辆检测模型如图 1 所示, 其中包括 4 个部分: (1) 点云体素化, 规范三维点云空间的表现形式, 便于后续网络处理; (2) Backbone 模块, 使用三维稀疏卷积网络 (SPConvNet) 结合辅助网络 (Auxiliary Network) 高效提取体素中内部点云特征; (3) 注意力模块, 对压缩后鸟瞰图的特征数据进一步处理, 重点关注有效特征; (4) 检测模块, 用卷积神经网络和 PS-Warp 变换机制对车辆目标的识别以及边界框的生成与回归.

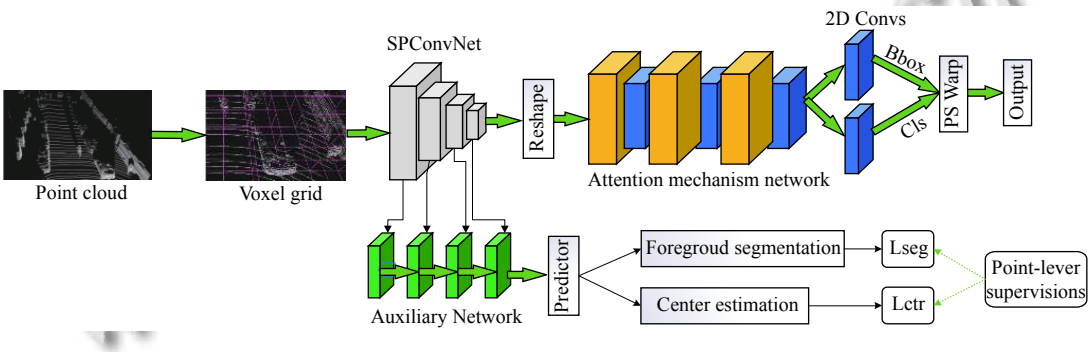


图 1 基于注意力机制的单阶段 3D 车辆目标检测模型

### 3.1 点云体素化

首先对输入的点云进行分块, 本文使用深度、高度和宽度 ($D$, $H$, $W$) 与输入点云相同尺寸的大立方体表示整个输入点云空间, 用相同尺寸的小立方体对其进行划分, 每个小立方体称为体素, 每个体素的深度、高度和宽度分别为$(u_D, u_H, u_W)$, 则整个点云空间在坐标上生成网格 (voxel grid) 个数为:

$$\left(\frac{D}{u_D}, \frac{D}{u_H}, \frac{D}{u_w}\right)$$

为了更高效的使用体素, 本文预设体素的最大数量限制 (max_voxels) 为 20 000, 根据体素的数量限制预先分配缓冲区, 接着遍历所有的点云, 将点分配给与它们相关联的体素, 并保存每个体素的点数和每个体素坐标的位置. 整个迭代过程是使用哈希表来检查体素是否存在的, 如果与点相关联的体素存在, 就将体素的数量增加一, 否则就在哈希表中设置相应的值. 一旦体素数量累加到预设的限制数值, 迭代过程将停止, 最后将所获得的每个体素的坐标和点的数量作为实际的体素数.

为了更好的检测车辆目标, 本文仅考虑沿着 Z, Y, X 轴在 [−3, 1]×[−40, 40]×[0, 70.4] 米范围内的点云. 用于车辆目标检测的每个体素的最大点数设置为 $T = 5$, 这是根据 KITTI 数据集中每个体素的点数分布决定的. 对于车辆目标检测任务, 本文使用的体素大小为 $u_D = 0.05, u_H = 0.05, u_W = 0.1$.

### 3.2 Backbone 模块

如图 1 所示, 本文使用三维稀疏卷积[1] 网络和辅助网络[13] 作为检测器的骨干网络提取特征. 三维稀疏卷积网络包含 4 个卷积块, 其中每个卷积块是由内核尺寸为 3 的子流形卷积构成的, 连接在模块最后的 3 个卷积块中附加有步幅长度为 2 的稀疏卷积, 每个卷积网络后面均接有层 ReLU 层和 RatchNorm 层, 最后会生成不同空间分辨率的多阶段的特征图. 通常, 从点云中提取的降采样多分辨率卷积特征将不可避免地丢失结构信息, 但细致的结构信息对于生成精确的目标定位至关重要.

本文采用一种具有逐点监督的可分离辅助网络, 辅助网络如图 1 所示. 它首先将稀疏卷积网络每阶段生成特征的非零索引转换为三维空间中点云坐标, 以逐点形式表示每个阶段的特征, 然后将这些点状特征连接起来, 并使用浅层预测变量来生成特定于任务的输出. 预测器是由大小为 (64, 64, 64) 神经元的共享多层感知器实现, 通过单位点卷积生成两个任务特定的输出. 最后通过点对前景的分割任务使得主干网络对

目标车辆边界框的检测更加准确. 具体来说, 用$\tilde{s}_i$表示分割分支的函数, 以预测每个点的前/后概率. 令$s_i$为指示点是否落入地面真相边界框内的二分类标签.使用 focal loss[15] 损失优化前景分割任务, 即:

$$L_{\text{seg}} = \frac{1}{N_{\text{pos}}} \sum_{i}^{N} -\alpha(1-\hat{s}_i)^{\gamma} \log_2(\hat{s}_i) \tag{5}$$

式中, $\alpha$ 和 $\gamma$ 是超参数, 本文分别设为 0.25 和 2, $\hat{s}_i$为:

$$\hat{s}_i = \begin{cases} \tilde{s}_i, & \text{if } \tilde{s}_i = 1 \\ 1-\tilde{s}_i, & \text{otherwise} \end{cases} \tag{6}$$

但是, 由于稀疏卷积产生的特征图非常稀疏, 即便是边界点被精确检测到, 在确定边界框的比例和形状时仍然存在着不确定性. 为了消除这一不确定性, 采用中心点估计的方法学习每个点到目标中心的相对位置. 假设$\Delta\tilde{p} \in \mathbb{R}^{N\times 3}$是中心估计分支的输出, $\Delta p$为点到相应中心的目标偏移量. 使用$Smooth-l_1$[16] 损失来优化中心估计任务:

$$L_{\text{ctr}} = \frac{1}{N_{\text{pos}}} \sum_{i}^{N} Smooth-l_1(\Delta\tilde{p}-\Delta p)\cdot I[s_i=1], \tag{7}$$

式中, $N$ 是前景点数, $I$ 是指标函数. 将前景分割和中心估计任务结合起来, 帮助主干网络了解更加精细的 3D 点云数据的结构信息, 此外本文只是在训练阶段采用辅助网络, 不会增加额外的计算成本.

### 3.3 注意力模块

(1) 通道注意力.通道注意力聚焦于输入数据中“有意义”的部分. 为了有效的量化通道注意力, 需要对输入特征数据的空间维度进行压缩处理, 目前, 大多采用平均池化提取有效的空间信息, 但经过研究发现最大池化也能收集到物体的一些独有特征, 因此让平均池化和最大池化两者相聚合就能提取出更加精细的空间信息.

如图 2 所示, 模型首先使用平均池化层和最大池化层聚合空间信息, 生成的平均池化特征和最大池化特征分别为$F_{\text{cavg}}$和$F_{\text{cmax}}$. 然后将这两个特征分别转发到共享网络中, 用来生成通道关注图$M_c \in R^{C\times 1\times 1}$, 其中共享网络由一个仅包含一个隐藏层的多层感知器 (MLP) 组成. 为了减少参数量运算开销, 将隐藏层的激活大小设置为$R^{(C/r)\times 1\times 1}$, 其中 $r$ 是缩小率. 将共享网络应用于每个特征后, 使用逐元素求和运算输出合并后特征向量.

通道注意力[17] 的计算公式为:

$$\begin{aligned} M_c(F) &= \sigma\{MLP[MaxPool(F)+AvgPool(F)]\} \\ &= \sigma\{W_1[W_0(F_{\text{cmax}})]+W_1[W_0(F_{cavg})]\} \end{aligned} \tag{8}$$

式中, $\sigma$表示 Sigmoid 函数; $W_0 \in R^{(C/r)\times C}$, $W_1 \in R^{C\times(C/r)}$, 且两个输入均共享 MLP 权重参数$W_0$和$W_1$.

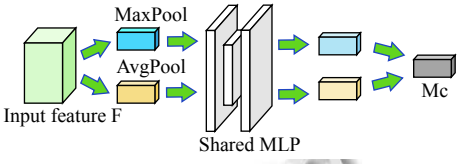


图 2 通道注意力模型

(2) 空间注意力. 与通道注意力不同, 空间注意力着重关注输入数据“在哪”的位置信息部分,这是对通道注意力的补充. 如图 3 所示, 模型同样使用两个池化层操作来聚合通道信息, 生成两个 2D 特征分别为: 平均池化特征$F^s_{\text{avg}} \in R^{1\times H\times W}$和最大池化特征$F^s_{\text{max}} \in R^{1\times H\times W}$, 然后将这两个特征连接起来, 通过标准卷积生成 2D 空间注意力图.

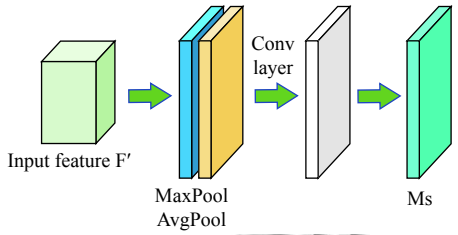


图 3 空间注意力模型

空间注意力[17] 的计算公式为:

$$\begin{aligned} M_s(F) &= \sigma\{[MaxPool(F);AvgPool(F)]f^{7\times 7}\} \\ &= \sigma\{[F_{\text{cmax}});(F_{\text{cavg}})]f^{7\times 7}\} \end{aligned} \tag{9}$$

式中, $f^{7\times 7}$表示卷积计算, 其中 7×7 为卷积核大小.

本文提出的注意力模块整体结构如图 3 所示. 棕黄色立方块表示为通道注意力模型与空间注意力模型的串联结构, 其中通道注意力模型在前, 空间注意力模型在后. 棕黄色立方块分布在整个注意力模块的首、中、尾部分, 其它部分嵌套着普通的二维卷积网络 (蓝色立方块). 实验结果表明, 这种“立体式”的布局结构使得整个模型对空间结构的感知能力大大加强, 同时对最终检测结果的提升也产生了重要的影响.

### 3.4 检测模块

检测模块采用普通卷积网络和 PSWarp[13] 变换机制对车辆目标进行检测. 为了解决最终车辆目标预测

的边界框和对应的置信度之间不匹配的问题, 本文采用了一种变形操作, 即 PSWarp 变换机制. PSWarp 可以看成比 PSRoIAlign 更加有效的变体, 它通过将特征图进行空间变换使得预测的边界框与分类置信度相匹配. PSWarp 是由一个特征图采样器组成, 它输入分类图和采样网格, 生成从网格点采样的输出图, 最后, 通过取 $K$ 个采样分类图中的平均值来计算分类置信度 $C$. 假定 $p$ 为一个车辆目标预测检测出的边界框, 则$\{(u^k, v^k) = S_p^k : k = 1,2,\cdots,k\}$为相对应的采样点, 则该预测边界框最终置信度计算公式为:

$$C_p = \frac{1}{k}\sum_{k=1}^{k}\sum_{\substack{i\in\{\lfloor u^k\rfloor\lfloor u^k+1\rfloor\}\\ j\in\{\lfloor v^k\rfloor,\lfloor v^k+1\rfloor\}}} x_{ij}^k \times b(i,j,u^k,v^k) \tag{10}$$

式中, $b$ 为双线性采样, 计算公式为:

$$b(i,j,u,v) = \max(1-|i-u|,0)\times\max(1-|j-u|,0) \tag{11}$$

PSWarp 与 PSRoIAlign 以及其它 RoI 的方法相比,有效地减轻了使用 NMS 从密集的特征图中生成 RoI 的需求, 达到节约时间成本提高效率的目的.

### 3.5 损失函数

本文参考文献 [1–3] 设置损失函数来优化主干网络, 在回归分支和分类分支上分别使用$L_{\text{loc}}$和$L_{\text{cls}}$损失, 其中$L_{\text{loc}}$使用的是$Smooth\text{-}l_1$[16] 范数, $L_{\text{cls}}$是$Focal\ loss$[17] 损失函数.

为了最小化 (12) 式中损失的加权和, 通过使用梯度下降方法联合优化检测和辅助任务:

$$L = L_{\text{cls}} + wL_{\text{loc}} + \mu L_{\text{seg}} + \lambda L_{\text{ctr}} \tag{12}$$

式中, $w$设置为 2; $\mu$和$\lambda$是使辅助任务与检测任务平衡的超参数分别为 0.9 和 2.

## 4 实验

### 4.1 数据集

KITTI 数据集是目前国际上最大的自动驾驶场景下的计算机视觉算法评测数据集, 它包含 7481 个训练样本和 7518 个测试样本, 在训练过程中又将训练集分为 3712 个训练样本和 3769 个验证样本, 本文主要是对应用最广的车辆类别进行实验, 并使用平均精度 (AP) 和 (IoU) 阈值 0.7 作为评估指标. 基准测试会根据目标的大小、遮挡情况以及截断程度区分出 3 个检测难度级别: 容易 (easy), 中等 (moderate) 和困难 (hard).

### 4.2 实验细节

实验环境为 Ubuntu 16.04 操作系统, Python 3.6, PyTorch 1.15. CPU 型号为 Inter Xeon Silver 4214, 显卡是 GeForce RTX 2080 Ti. 在训练中, 正锚和负锚使用的匹配阈值分别为 0.6 和 0.45; 用于检测车辆的锚点的尺寸为 1.6 m (宽), 3.9 m (长) 和 1.56 m (高), 所有不包含点的锚都忽略; 使用 SGD 优化器对整个网络进行了 90 个周期的训练. 批次大小, 学习率和权重衰减分别为 2、0.01 和 0.003; 在推理阶段, 使用 0.3 阈值过滤掉低置信度边界框, NMS 的 IoU 阈值为 0.1.

### 4.3 数据增强

数据增强通过增加训练数据量以及增加一些干扰噪声数据来提高模型的泛化能力和鲁棒性. 具体来说, 首先收集所有真实目标的三维边界框以及边界框内的点云作为样本池, 对于每个样本, 采取随机抽取的方法, 从样本池中随机抽取不超过 10 个样本, 并将它们放入到当前点云数据中参与训练, 在每个样本放入数据后, 还需要对其进行碰撞测试, 避免违反物理规律. 接着对真实目标边界框的数量进行扩充, 新增的每个真值框都进行随机旋转和平移. 旋转从 [−π/15, π/15] 开始, 均匀的增加旋转噪声. 真值框的 X, Y, Z 坐标按 $N$(0, 0, 0.25) 的正态分布进行平移, 增加平移噪声. 除此之外, 还将对整个点云进行随机翻转、全局旋转和全局缩放. 全局旋转从 [−π/15, π/15] 开始, 均匀的增加全局旋转噪声. 缩放因子则从 [0.95, 1.05] 开始改变.

### 4.4 实验结果及分析

为了验证基于注意力机制的三维点云车辆目标检测算法的有效性, 本文在 KITTI 3D 目标检测基准上对本文所提出的单阶段 3D 车辆检测模型进行评估. 实验通过将相关数据提交给 KITTI 线上测试服务器, 生成的实验检测结果与新近主流的 3D 车辆检测算法相比较 (KITTI 数据集默认按中等难度的检测精度进行排名):

从表 1 可以看出, 本文所提出的单阶段车辆检测算法甚至比新近一些主流的两阶段车辆检测算法 (如 F-PointNet, TANet, 3D IoU-Net) 表现更好, 并且它在 3 个检测难度级别中均取得较好的结果. 其中在中等和困难检测级别中取得最优成绩, 在简单检测级别中取得的检测结果也与第一名相差不大. 以上可证明本文所提算法的有效性.

值得关注的是, 当引入注意力机制后, 模型整体的

方向性预估变得更加优秀. 与表 1 中简单检测级别中检测精度最高的模型 EBM3DOD 相比, 本文提出的算法模型在 3 个检测难度级别中方向性预估性能均表现的更好. 其中与当前主流的单阶段车辆检测算法 SA-SSD 相比, 性能提升尤为明显. 检测结果如表 2、图 4、图 5 和图 6 所示:

表 1 3D 车辆目检测精度对比

| 算法 | 车辆的mAP (%) | | |
|---|---|---|---|
| | Easy | Moderate | Hard |
| VoxelNet[4] | 77.47 | 65.11 | 57.73 |
| F-PointNet[5] | 82.05 | 68.46 | 62.42 |
| SECOND[1] | 83.13 | 73.66 | 66.20 |
| PointPillars[3] | 79.05 | 74.99 | 68.30 |
| TANet[18] | 84.39 | 75.94 | 68.82 |
| CLOCs_SecCas[19] | 86.38 | 78.45 | 72.45 |
| SERCNN[20] | 87.74 | 78.96 | 74.30 |
| 3D IoU-Net[21] | 87.96 | 79.03 | 72.78 |
| EPNet[22] | 89.81 | 79.28 | 74.59 |
| Point-GNN[23] | 88.33 | 79.47 | 72.29 |
| 3DSSD[24] | 88.36 | 79.57 | 74.55 |
| SA-SSD[13] | 88.75 | 79.79 | 74.16 |
| EBM3DOD[25] | **91.05** | 80.12 | 72.78 |
| CLA-SSD[26] | 89.59 | 80.28 | 72.87 |
| 本文 | 89.58 | **80.30** | **75.02** |

表 2 3D 车辆目标方向性预估对比

| 算法 | 方向的mAP (%) | | |
|---|---|---|---|
| | Easy | Moderate | Hard |
| SA-SSD[13] | 39.40 | 38.30 | 37.07 |
| EBM3DOD[25] | 96.39 | 92.88 | 87.58 |
| 本文 | **96.56** | **93.18** | **90.13** |

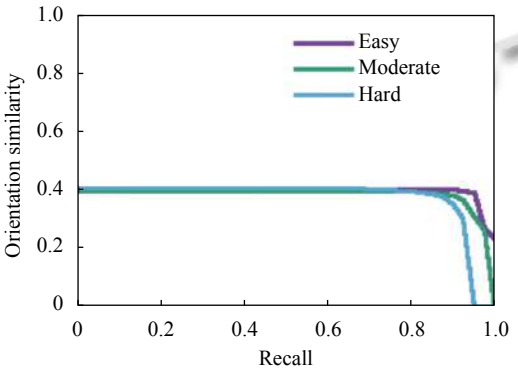


图 4 SA-SSD 车辆方向估计检测结果

## 5 结论与展望

提出了一种基于注意力机制的单阶段检测模型,用来检测真实生活场景下的车辆目标. 通过引入注意力机制, 使得模型对空间的感知能力更强, 从而使得车辆目标方向性预估更准确. 再将实验检测结果与新近一些优秀模型的检测结果相比较可以发现, 基于注意力机制的检测模型在综合预测结果方面也具有较大优势. 但对于自动驾驶领域的 3D 目标检测研究, 无论是检测速度还是检测精度都还有进一步提升空间. 后续的研究将不仅限于对车辆目标的检测, 可考虑在基于注意力机制的检测模型上对行人、骑行者等小目标进行检测, 通过进一步在模型结构设计及目标检测过程的机理上深入挖掘, 以达到更好的效果.

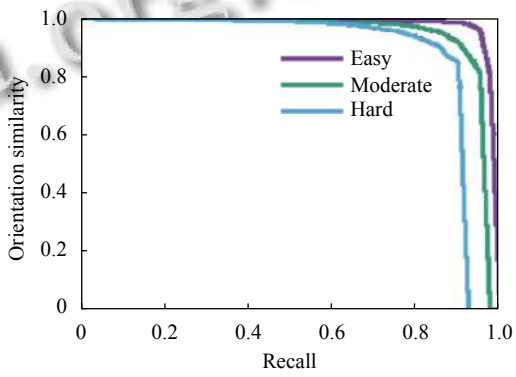


图 5 EBM3DOD 车辆方向估计检测结果

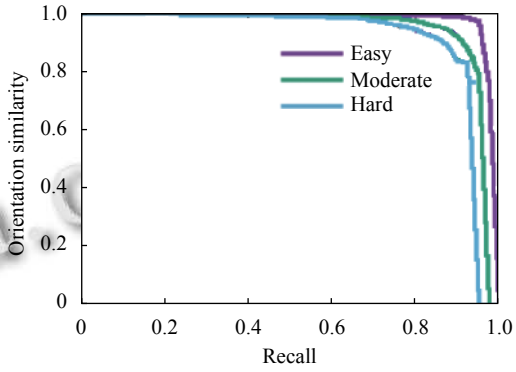


图 6 本文车辆方向估计检测结果

## 参考文献

1 Yan Y, Mao YX, Li B. Second: Sparsely embedded convolutional detection. Sensors, 2018, 18(10): 3337. [doi: 10.3390/s18103337]

2 Lang AH, Vora S, Caesar H, *et al*. PointPillars: Fast encoders for object detection from point clouds. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 12689–12697.

3 Simon M, Milz S, Amende K, *et al*. Complex-YOLO: An euler-region-proposal for real-time 3D object detection on

point clouds. Proceedings of the European Conference on Computer Vision (ECCV). Munich: Springer, 2019. 197–209.
4 Zhou Y, Tuzel O. VoxelNet: End-to-end learning for point cloud based 3D object detection. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 4490–4499.
5 Qi CR, Liu W, Wu CX, *et al*. Frustum PointNets for 3D object detection from RGB-D data. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 918–927.
6 Shi SS, Wang XG, Li HS. PointRCNN: 3D object proposal generation and detection from point cloud. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach: IEEE, 2019. 770–779.
7 Yang ZT, Sun YN, Liu S, *et al*. STD: Sparse-to-dense 3d object detector for point cloud. Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul: IEEE, 2019. 1951–1960.
8 Shi SS, Wang Z, Wang XG, *et al*. Part-$A^2$ Net: 3D part-aware and aggregation neural network for object detection from point cloud. arXiv: 1907.03670, 2019.
9 Geiger A, Lenz P, Stiller C, *et al*. Vision meets robotics: The KITTI dataset. The International Journal of Robotics Research, 2013, 32(11): 1231–1237. [doi: 10.1177/0278364913491297]
10 Li B. 3D fully convolutional network for vehicle detection in point cloud. 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Vancouver: IEEE, 2017. 1513–1518.
11 Yang B, Luo WJ, Urtasun R. Pixor: Real-time 3D object detection from point clouds. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 7652–7660.
12 Wu BC, Wan A, Yue XY, *et al*. SqueezeSeg: Convolutional neural nets with recurrent CRF for real-time road-object segmentation from 3D LiDAR point cloud. 2018 IEEE International Conference on Robotics and Automation (ICRA). Brisbane: IEEE, 2018. 1887–1893.
13 He CH, Zeng H, Huang JQ, *et al*. Structure aware single-stage 3D object detection from point cloud. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle: IEEE, 2020. 11870–11879.
14 Charles RQ, Su H, Kaichun M, *et al*. PointNet: Deep learning on point sets for 3D classification and segmentation. Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu: IEEE, 2017. 77–85.
15 Lin TY, Goyal P, Girshick R, *et al*. Focal loss for dense object detection. Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV). Venice: IEEE, 2017. 2999–3007.
16 Liu W, Anguelov D, Erhan D, *et al*. SSD: Single shot MultiBox detector. 14th European Conference on Computer Vision. Amsterdam: Springer, 2016. 21–37.
17 Woo S, Park J, Lee JY, *et al*. CBAM: Convolutional block attention module. Proceedings of the 15th European Conference on Computer Vision. Munich: Springer, 2018. 3–19.
18 Liu Z, Zhao X, Huang TT, *et al*. TANet: Robust 3D object detection from point clouds with triple attention. Proceedings of the 34th AAAI Conference on Artificial Intelligence. Palo Alto: AAAI, 2020. 11677–11684.
19 Pang S, Morris D, Radha H. CLOCs: Camera-LiDAR object candidates fusion for 3D object detection. 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Las Vegas: IEEE, 2020. 10386–10393.
20 Zhou DF, Fang J, Song XB, *et al*. Joint 3D instance segmentation and object detection for autonomous driving. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle: IEEE, 2020. 1836–1846.
21 Li JL, Luo SJ, Zhu ZQ, *et al*. 3D IoU-Net: IoU guided 3D object detector for point clouds. arXiv: 2004.04962, 2020.
22 Huang TT, Liu Z, Chen XW, *et al*. EPNet: Enhancing point features with image semantics for 3D object detection. 16th European Conference on Computer Vision. Glasgow: Springer, 2020. 35–52.
23 Shi WJ, Rajkumar R. Point-GNN: Graph neural network for 3D object detection in a point cloud. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle: IEEE, 2020. 1708–1716.
24 Yang ZT, Sun YN, Liu S, *et al*. 3DSSD: Point-based 3D single stage object detector. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle: IEEE, 2020. 11037–11045.
25 Gustafsson FK, Danelljan M, Schön TB. Accurate 3D object detection using energy-based models. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Nashville: IEEE, 2021. 2849–2858.
26 Zheng W, Tang WL, Chen SJ, *et al*. CIA-SSD: Confident IoU-aware single-stage object detector from point cloud. Proceedings of the 35th AAAI Conference on Artificial Intelligence. Palo Alto: AAAI, 2021. 3555–3562.