计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.190900172

基于两级网络的三维目标检测算法

沈　琦[1]　陈逸伦[2]　刘　枢[3]　刘利刚[1]
1 中国科学技术大学数学科学学院　合肥 230026
2 香港中文大学计算机科学与工程学院　香港 999077
3 深圳市腾讯计算机系统有限公司　广东 深圳 518057
(sq000333@mail.ustc.edu.cn)

摘　要　文中提出了一种基于激光雷达点云的三维目标检测算法 VoxelRCNN(Voxelization Region-Based Convolutional Neural Networks),该算法基于 VoxelNet 三维目标检测网络算法,将 RCNN 算法的思想从二维目标检测运用到三维目标检测中。VoxelRCNN 算法由两级构成,第一级的目标是用区域提案网络提取候选区域框信息,第二级的目标是对第一级提取的目标检测框进行更精细的修正,以得到更精确的目标检测结果。第一级网络对整个场景的点云进行体素化,对每个体素块提取特征作为卷积神经网络的输入,经过卷积神经网络计算得到最后的特征图,根据特征图对包围盒信息进行回归学习。第二级网络依据第一级提取的候选区域信息以及特征信息,通过池化得到等大特征信息,再次回归学习包围盒信息。在 KITTI 数据集上的实验结果表明,提出的网络结构是有意义的。

关键词:三维目标检测;体素化;卷积神经网络;区域提案网络;KITTI 数据集
中图法分类号　TP391.41

3D Object Detection Algorithm Based on Two-stage Network

SHEN Qi[1],CHEN Yi-lun[2],LIU Shu[3] and LIU Li-gang[1]
1 School of Mathematical Sciences,University of Science and Technology of China,Hefei 230026,China
2 Department of Computer Science,and Engineering,The Chinese University of Hong Kong,Hong Kong 999077,China
3 Tencent Holdings Ltd.,Shenzhen,Guangdong 518057,China

Abstract　This paper proposes a 3D object detection algorithm,named VoxelRCNN,on the basis of LIDAR point cloud. This algorithm is based on VoxelNet 3D object detection network algorithm,and the idea of RCNN algorithm is applied to 3D object detection from 2D object detection. The VoxelRCNN algorithm is composed of two stages. Stage-1 aims to extract the information of candidate region box with the regional proposal network,and stage-2 aims to refine the object detection box extracted in stage-1,to obtain more accurate detection results. The stage-1 network voxelizes the point cloud of the whole scene,extracts the features of each voxel block as the input of the convolutional neural network,and obtains the final characteristic map through the convolutional neural network calculation. Then,the enveloping box information is learnt by regression according to the feature map. In stage-2,on the basis of the candidate region information and feature information extracted in stage-1,equivalent feature information is obtained by pooling,and returning to learning bounding box information again. Experimental results on KITTI dataset show that the proposed network structure performs well.

Keywords　3D object detection,Voxelization,Convolutional neural network,Region proposal network,Kitti dataset

1　引言

随着硬件技术的发展,自动驾驶得到越来越多的关注,车辆目标检测是自动驾驶算法中非常重要的一个环节,成为了炙手可热的研究方向。

目前,深度学习在二维图像目标检测工作中取得了显著的进展,譬如 RCNN[1],YOLO[2],SSD[3] 等。但是,在自动驾驶和机器人等应用场景中,二维场景下的目标检测对于三维真实世界的场景描述不够,存在遮挡、阴影等问题,因此三维目标检测不可或缺。

与二维目标检测相比,三维目标检测更具挑战性,其不仅关注目标在图像平面的位置,还关注其在真实三维空间中的位姿。在三维车辆目标检测中,最常用的传感器是激光雷达传感器,尽管激光雷达传感器的精度高,但其存在点云不规则的问题。目前,基于点云较成熟的三维检测方法有两种,一种是将点云投射到正视图或鸟瞰图中,即 MV3D[4] 和 PIXOR[5]

到稿日期:2019-09-25　返修日期:2019-11-25　本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
通信作者:刘利刚(lgliu@mail.ustc.edu.cn)

等，另一种是对点云构造新的结构，即 VoxelNet[6] 和 PointNet[7] 等，本文采用对点云进行体素化的方法。三维目标检测的另一个难点在于高自由度，三维的包围盒(Bounding Box)是在真实三维世界中包围目标物体的最小长方体，理论上其有 9 个自由度，而自动驾驶场景下的物体都是水平放置于地面的，可以省略 3 个自由度，即三维车辆目标检测是 6 个自由度的目标检测问题。

本文提出了一种新的直接作用于激光雷达点云两级的三维目标检测算法 VoxelRCNN，其结构如图 1 所示。该算法分为两级，第一级的目标是利用卷积神经网络(Convolutional Neural Network，CNN)和区域提案网络(Region Proposal Network，RPN)[8]提取候选区域框信息。RPN 是目标检测中的一种高效算法，可以有效预测具有广泛规模和纵横比的区域方案。本文在基于卷积特征的基础上，通过添加一些额外的卷积层来构造 RPN，这些卷积层同时回归规则网格上每个位置的区域边界和对象性分数，以提取初步的候选区域框信息。

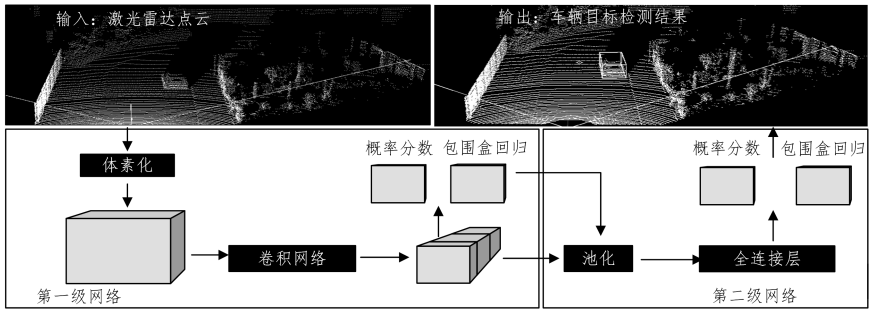


图 1 VoxelRCNN 网络结构

Fig. 1 VoxelRCNN network structure

由于第一级的核心目标是对巨大量级的样本中的无意义负样本进行抑制，并完成一个粗糙的定位，因此我们需要第二级网络来对第一级提取的目标检测框进行更精细的修正，以获得更精准的检测结果。我们对第一级得到的目标检测结果进行池化操作(RoI Align)，将每一个第一级得到的目标检测框转化为固定大小的特征，然后接上全连接层，对包围盒修正量进行进一步的学习。实验结果表明，本文方法相较于目前的三维车辆检测方法是有意义的。基于三维检测 KITTI 评估集的常用评估标准，VoxelRCNN 的第一级结果相较于 VoxelNet 提升了 8.39%，第二级网络修正后的结果相较于第一级结果又提升了 1.37%。

2 相关工作

2.1 二维图像目标检测

2014 年至今，涌现了两类经典的二维图像检测算法：1)以 RCNN 为代表的结合区域提案(Region Proposal)和卷积神经网络的目标检测算法，如 RCNN[1]，SPP-NET[9]，Fast RCNN[10]，Faster RCNN[8]，Mask RCNN[11]；2)以 YOLO 为代表的将定位分类问题转换为回归问题的目标检测算法，如 YOLO[2]，SSD[3]。第一类算法融合了区域提案和卷积神经网络，使端到端的网络进行目标检测实际可行，速度与精度都有了很大提高，然而区域提案的提取和分类的计算量仍然较大，此类算法仍无法达到实时目标检测的要求。第二类基于位置回归的方法直接在图像的多个位置上回归出目标位置和类别，加快了检测速度，使目标检测在实际应用中成为可能。

2.2 三维目标检测

2014 年以前，三维目标检测主要是基于人工设计特征进行模板匹配的算法(如 DOT[12]，GRM[13])，即利用滑动窗口方式进行搜索，利用离散采样模板来估计姿态，但传统方法的速度和精度都较差。2014 年以后，随着二维图像目标检测算法的快速发展，三维目标检测也快速发展，基于学习的方法成为主流。下文根据检测设备的不同(单目相机、双目相机和多线激光雷达)，对三维目标检测算法进行分类介绍。

2.2.1 基于图像的三维目标检测

由于激光雷达传感器的价格较高，有不少工作仅利用单目或双目相机图像来恢复物体的位姿信息。Deep3Dbox[14]利用深度神经网络回归出相对稳定的三维目标特征，再利用卷积网络得到二维包围盒信息，以此作为物体三维位姿的几何约束。Mono3D[15]经过语义信息、上下文信息和定位信息(鸟瞰图与正视图的先验信息)等多种信息编码后，利用回归得到三维包围盒。Multi-Level3D[16]利用单目图像多层融合来预测三维包围盒信息。Deep MANTA[17]利用三维对象与其对应模型间的相似性来构建三维空间坐标与二维图像点之间的关系，从而利用卷积神经网络回归进行包围盒预测。但由于深度信息的不完整性，以上方法只能生成粗略的检测结果。

2.2.2 基于激光雷达点云的三维目标检测

基于激光雷达点云的三维目标检测工作众多，根据对点云的处理，可以分为 3 类：1)不做任何处理，利用原始点云对卷积网络进行检测，譬如 3D FCN[18]和 Vote3Deep[19]等，由于激光雷达点云稀疏且不规则，这类方法通常耗时且学习效果差。2)PIXOR[5]和 Complex-YOLO[20]等将点云投影到鸟瞰图或前视图上，利用卷积神经网络学习点云的分类特征和包围盒回归特征。尽管这样的映射过程会造成一部分信息损失，但针对车辆检测这类场景，这类工作能够很好地解决遮挡问题，并且可以将三维网络转化为二维网络，大大降低了空间复杂度和时间复杂度。3)有些研究提出用特殊的网络结构来表示点云，如 VoxelNet[6]将点云划分为体素结构单元，在非空结构单元利用网络提取特征；PointNet[7]和 PointNet++[21]等

方法基于原始点云学习更有效的空间几何表示，根据学习提取点云特征。

2.2.3　基于图像和点云的三维目标检测

基于图像和点云的目标检测大多是利用点云和图像两类信息进行并行处理与融合，利用融合结果对目标进行预测。MV3D[4]将三维点云投影到俯视图和鸟瞰图，利用鸟瞰图通过卷积网络生成低精度的目标区域，再利用该区域与俯视图、鸟瞰图和单目图像的特征，构建一个融合网络进行训练。AVOD[22]输入二维图像及鸟瞰图，利用 FPN 网络得到二者全分辨率的特征图，再通过裁剪修整提取两个特征图对应的区域进行融合，筛选后进行三维物体检测。ContFuse[23]将图像特征从前视图转化到俯视图，再与点云数据对应的俯视图特征进行数据融合，最后进行三维检测。这类方法由于存在融合操作，导致运行速度较慢，无法满足实时性。另外一类方法是利用二维图像来驱动三维点云，即利用串行的方法来实现目标检测。F-PointNet[24]通过图像到点云的待检测物体的定位过程，实现了 PointNet[7]的有效使用。

3　网络结构

3.1　第一级网络

第一级检测结构以一个原始点云为输入，将其转化为体素特征，接着用卷积神经网络对特征进行操作，最后用 RPN 来生成第一级目标检测结果。

3.1.1　体素化

由于激光雷达传感器在采集点云时存在遮挡、阴影等因素，因此激光雷达点云通常是稀疏的，且点密度是高度可变的，我们需要对点云进行体素化操作。本文的体素化操作相较于 VoxelNet[6]中的体素化操作进行了大幅度简化，将取样多个点计算 VFE 特征简化为仅取样单点以空间坐标为特征。这不仅节省了大量的计算量，且对车辆检测结果无影响。

本文将点云所在三维空间细分为等大小的体素。假设这片点云沿着 Z,Y,X 方向的尺寸大小为 D,H,W，相对应的每个体素小方块的尺寸为 v_D,v_H,v_W，则得到体素网格的大小为 $D'=\frac{D}{v_D},H'=\frac{H}{v_H},W'=\frac{W}{v_W}$（我们假设 D,H,W 为 v_D,v_H,v_W 的倍数）。这样体素化分组后，由于点云具有点密度可变性，每个体素将包含可变数量的点。为了减少体素间点的不平衡，减少采样偏差，也为了节省计算量，我们对每个包含点的体素随机选取一个点，并以该点的空间坐标为该体素的体素特征。尽管点云包含 10 万个点，但超过 90%的体素为空，因此可以将获得的体素特征表示为一个稀疏的 5 维张量，大小为 $C\times D'\times H'\times W'\times 4$，其中 C 为批尺寸(Batch Size)，4 为体素特征大小，即点的空间坐标。

3.1.2　区域提案网络

目前，区域提案网络为目标检测算法中的重要组成部分。本文也利用 RPN 网络进行目标检测。

RPN 的输入是由卷积中间层提供的特征图，该网络结构如图 2 所示。网络的输入为体素化得到的 5 维张量，先将其转化为 4 维张量，再通过一个核为 3、步幅为 2、过滤器数目为 128 的卷积网络。而后，网络经过 3 个完全卷积网络块，每个网络块由 5 层构成，每个网络块的第一层通过一个核为 3、步幅为 2 的卷积层，将特征图缩小 1/4，其余 4 层通过一个核为 3、步幅为 1 的卷积层。此外，在每一个卷积层之后均添加 BN 和 Relu 操作。本文将 3 个网络块的最后一层进行逆卷积操作，将其转化为相同大小的特征图并进行组合。最后，将组合得到的特征映射到所需的学习目标，即概率分数和包围盒回归。

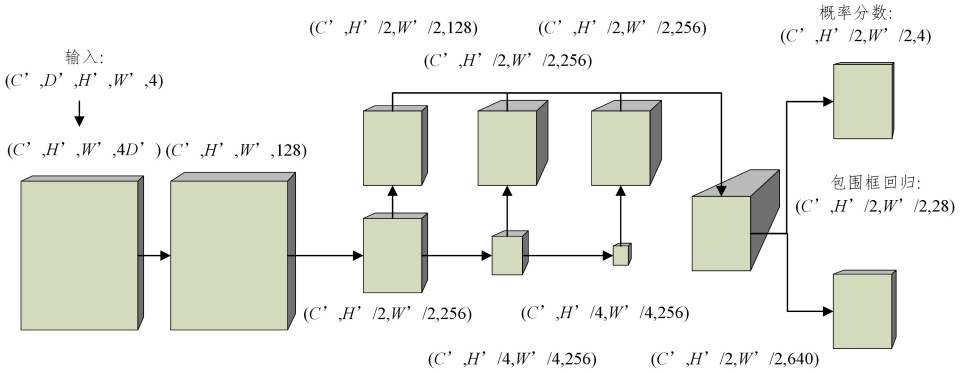


图 2　区域提案网络结构

Fig. 2　Regional proposal network structure

3.1.3　损失函数

本文构造锚盒(Anchor Box)，将其参数化为 $(x_c^a,y_c^a,z_c^a,l^a,w^a,h^a,\theta^a)$，其中 x_c^a,y_c^a,z_c^a 表示包围盒中心点的空间坐标，l^a,w^a,h^a 表示包围盒的长宽高，θ^a 表示绕 Z 轴旋转的角度。与常见的锚盒构造原理相同，本文线性选取 $W'/2$ 个 $x\in[x_{\min},x_{\max}]$，$H'/2$ 个 $y\in[y_{\min},y_{\max}]$(与特征图大小相同)，$\theta=[0°,45°,90°,135°]$将其组合即得锚盒。

同样地，本文将三维真实车辆包围盒参数化为 $(x_c^g,y_c^g,z_c^g,l^g,w^g,h^g,\theta^g)$。为了使计算得到的正锚盒能够恢复成真实包围盒，本文定义了残差向量 $\boldsymbol{u}^*$，其包含了 7 项目标回归量：

$$\Delta x=(x_c^g-x_c^a)/d^a,\Delta y=(y_c^g-y_c^a)/d^a$$

$$\Delta z=(z_c^g-z_c^a)/h^a,\Delta\theta=G(\theta^g-\theta^a)$$

$$\Delta l=\log(l^g/l^a),\Delta w=\log(w^g/w^a),\Delta h=\log(h^g/h^a)$$

其中，$d^a=\sqrt{(l^a)^2+(w^a)^2}$ 为锚盒的对角线长度，G 将任意大小的角度映射到$[-\pi/2,\pi/2]$。将损失函数定义为：

$$L=\sum_i L_{\text{reg}}(u_i,u_i^*)/N_{\text{pos}}+\alpha_1\sum_i L_{\text{cls}}(p_i^{\text{pos}},1)/N_{\text{pos}}+$$

$$\beta_1 Ohem(\sum_j L_{cls}(p_j^{neg},0)/N_{neg})$$

其中，p_i^{pos} 和 p_j^{neg} 表示正锚盒和负锚盒的概率分数，u_i 和 u_i^* 表示网络包围盒回归结果、真实样本和正锚盒的残差结果，N_{pos} 和 N_{neg} 表示正、负样本的数量，α_1 和 β_1 为正、负样本系数。回归损失函数 L_{reg} 采用稳定的光滑的 L_1 函数，即：

$$L_{reg}(x)=\begin{cases}0.5x^2 & ,\text{if } |x|<1\\ |x|-0.5 & ,\text{otherwise}\end{cases}$$

这是一种目标检测常用的 L_1 损失函数，与 L_2 损失相比，对离群点不会太敏感。

由于在检测数据集中总是包含大量简单样本和少量困难样本，选择困难样本能让训练更加快速有效。因此，在损失函数定义中，本文运用 Ohem(Online Hard Example Mining)算法来平衡样本集，该算法的核心思想是根据输入样本的损失函数值筛选出困难的样本，即具有多样性和高损失的样本，然后得到样本应用到训练中。

3.1.4 数据扩充

在第一级网络训练中遇到的主要问题是样本真实值太少，这严重限制了网络的收敛速度和最终性能。为了解决该问题，我们对数据集进行扩充。首先对点云添加噪声，使用均匀分布的 $\Delta\theta\in[-\pi/2,\pi/2]$ 绕 Z 轴随机旋转，以及基于平均值为 0、标准偏差为 1.0 的高斯分布的随机线性变换。此外，我们运用全局缩放和旋转点云来扩充数据集，即将点坐标与从均匀分布[0.95,1.05]中提取的随机变量相乘，将绕 Z 轴的旋转角度与从均匀分布$[-\pi/4,\pi/4]$中提取的随机变量相加。

3.2 第二级网络

第一级网络的核心目标为筛选合适的样本，并获得一个粗糙的目标候选区域。第二级网络的核心目标为对第一级网络得到的目标检测结果的包围盒进行修正，以进一步提升检测结果。

3.2.1 网络结构

依据第一级网络得到的概率分数和包围盒回归值，我们可以得到感兴趣的目标区域。

基于第一级网络中得到的特征与目标区域，我们进行池化(RoI Align)操作。该操作利用最大池化将感兴趣的目标区域的特征转化为一个固定大小 $w_a\times h_a$ 的特征。具体操作如下：将候选目标区域分割为 $w_a\times h_a$ 个单元，对每个单元选取 4 个顶点位置，利用双线性内插的方法计算 4 个顶点在原目标区域的位置，将这 4 个位置对应的特征值进行最大池化操作，以得到该单元的值。在池化层后连接全连接层对概率分数和包围盒回归进行修正学习。

3.2.2 损失函数

类似第一级网络，本文将第一级网络得到的区域参数化为$(x_c^a,y_c^a,z_c^a,l^a,w^a,h^a,\theta^a)$，将三维真实车辆包围盒参数化为$(x_c^g,y_c^g,z_c^g,l^g,w^g,h^g,\theta^g)$。为了让检测结果能够更加接近真实包围盒结果，我们采用回归项：

$$\Delta x=(x_c^g-x_c^a)/l^a,\Delta y=(y_c^g-y_c^a)/w^a$$

$$\Delta z=(z_c^g-z_c^a)/h^a,\Delta\theta=G(\theta^g-\theta^a)$$

$$\Delta l=\log(l^g/l^a),\Delta w=\log(w^g/w^a),\Delta h=\log(h^g/h^a)$$

损失函数的定义如下：

$$L=\sum_i L_{reg}(u_i,u_i^*)/N_{pos}+\alpha_2\sum_i L_{cls}(p_i^{pos},1)/N_{pos}+\beta_2\sum_j L_{cls}(p_j^{neg},0)/N_{neg}$$

我们采用 smooth_{L_1} 函数为回归损失函数 L_{reg}，p_i^{pos} 和 p_j^{neg} 表示正锚盒和负锚盒的概率分数，u_i 和 u_i^* 表示网络包围盒回归结果、真实样本和正锚盒的残差结果，N_{pos} 和 N_{neg} 表示正、负样本的数量，α_2 和 β_2 为正、负样本系数。

4 实验结果与分析

本文基于激光雷达点云的 KITTI 数据集进行了实验分析，本文方法的测试结果如图 3 所示。

注：为了更好地可视化，第一列将三维检测框映射到图像上；第二列和第三列为在点云三维视角和鸟瞰视角的三维检测结果

图 3 基于 KITTI 数据集的可视化结果

Fig. 3 Visualization results based on KITTI data set

4.1 网络参数

在体素化时，本文考虑点云沿着 Z,Y,X 在$[-3,1]\times[-40,40]\times[0,70.4]$范围内(单位为米)的点，将其余点移除。我们选择每个体素的大小为 $v_D=0.4\text{m},v_H=0.2\text{m},v_W=0.2\text{m}$，体素网格的大小即为 $D'=10,H'=400,W'=352$。

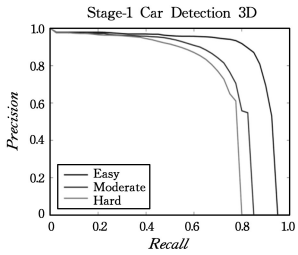
在构造锚盒时，对锚盒的一些变量选择固定数值，$l^a=3.9,w^a=1.6,h^a=1.56,z_c^a=-1.78$。对构造锚盒的判定规则如下：若该锚盒在鸟瞰视角下与某个真实包围盒的交并比(Intersection over Union，IoU)大于 0.6，则其被认为是正锚盒；若该锚盒与任何一个真实包围盒的交并比都小于 0.45 时，则其被认为是负锚盒；对于其余的锚盒我们不做训练使用。

训练时，第一级网络训练为 100 个周期，批量大小分别为 3，1，10，初始学习率为 0.00125，并且网络分别至第 70 个和第 95 个周期学习率衰减为原来的 1/10。

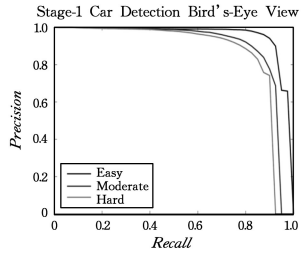
第二级网络训练为 70 个周期，批量大小为 12，α_2 为 1，β_2 为 2，w_a 为 4，h_a 为 8，初始学习率为 0.00125，并且网络分别至第 40 个、第 55 个和第 65 个周期学习率衰减为原来的 1/10。

4.2 结果分析

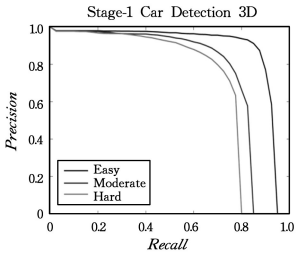
我们在 KITTI 数据集上进行训练，并且对三维和鸟瞰视角进行检测。车辆检测可以划分为 3 个难度：简单、中等、困难。KITTI 数据集包含 7481 个训练数据和 7518 个测试数据。我们依据文献[7]提出的方法将 7481 个训练数据划分为 3712 个训练样本点云以及 3769 个评估样本点云。我们采用平均精度(Average Precision，AP)作为评价指标，并且设置交并比阈值为 0.7。基于评估样本，对鸟瞰图和三维视角进行评估，结果如图 4、表 1 和表 2 所示。



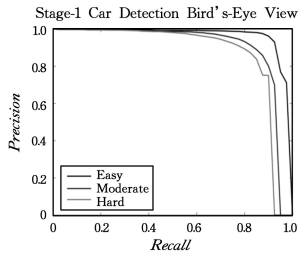
(a)第一级网络三维检测结果



(b)第一组网络鸟瞰检测结果



(c)第二级网络三维检测结果



(d)第二级网络鸟瞰检测结果

图 4　基于平均精度度量所得的结果

Fig. 4　Results based on average metrics

表 1　鸟瞰图检测的平均精度

Table 1　*AP* of Bird's-eye view detection

(单位：%)

Method	Easy	Moderate	Hard
VoxelNet[6]	89.35	79.26	77.39
MV3D[4]	86.55	78.10	76.67
F-PointNet[24]	88.16	84.02	76.44
VoxelRCNN Stage-1	89.91	87.33	86.25
VoxelRCNN Stage-2	90.06	87.71	86.43

表 2　三维检测性能的平均精度

Table 2　*AP* of 3D detection

(单位：%)

Method	Easy	Moderate	Hard
VoxelNet[6]	81.98	65.46	62.85
MV3D[4]	71.29	62.68	56.56
F-PointNet[24]	83.76	70.92	63.65
VoxelRCNN Stage-1	85.28	73.85	67.31
VoxelRCNN Stage-2	86.45	75.22	67.75

从表 1、表 2 中可以看到，在鸟瞰图检测上，中等难度和困难难度的检测结果提升得较多。在三维检测中，整体的检测结果都有提升。特别地，在三维检测的中等难度的情况下，VoxelRCNN 第一级结果相较于 VoxelNet 提升了 8.39%，第二级结果相较于第一级结果又提升了 1.37%。第一级检测结果提升 8.39%的主要原因是采用数据扩充和损失函数平衡了正负样本。加入数据扩充方法和损失函数平衡方法后，对比检测结果如表 3 所列。

表 3　第一级网络检测性能的平均精度

Table 3　*AP* of stage-1 network detection

(单位：%)

	No Ohem	Ohem
Not Expand Data	65.99	72.60
Expand Data	71.09	73.85

基于 KITTI 数据集中等难度交并比阈值为 0.7 的评估标准，本文对传感器不同范围内的检测结果进行评估，结果如表 4 所列。可以看到，检测目标距离传感器越近，检测结果越好，距离传感器越远，检测结果越差。相较于第一级网络，第二级网络在不同距离范围内的检测结果的平均精度均有提高，对中远距离的精度提高更为显著。

表 4　不同范围检测下的平均精度

Table 4　*AP* under different ranges

(单位：%)

	All	0～30m	30～50m	50～60m
VoxelRCNN Stage-1	73.85	87.21	46.02	4.62
VoxelRCNN Stage-2	75.22	87.57	48.03	7.27

结束语　本文提出了基于激光雷达点云的两级三维目标检测算法 VoxelRCNN。基于 VoxelNet 三维目标检测网络算法，将 RCNN 算法的思想从二维目标检测运用到三维目标检测中，得到了较好的实验结果。基于 KITTI 数据集中等难度交并比阈值为 0.7 的平均精度，VoxelRCNN 第一级结果相较于 VoxelNet 提升了 8.39%，第二级结果相较于第一级结果又提升了 1.37%。

参 考 文 献

[1] GIRSHICK R, DONAHUE J, DARRELL T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2014: 580-587.

[2] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: Unified, real-time object detection[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 779-788.

[3] LIU W, ANGUELOV D, ERHAN D, et al. Ssd: Single shot multibox detector[C]// European Conference on Computer Vision. Springer, Cham, 2016: 21-37.

[4] CHEN X, MA H, WAN J, et al. Multi-view 3d object detection network for autonomous driving[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 1907-1915.

[5] YANG B, LUO W, URTASUN R. Pixor: Real-time 3d object detection from point clouds[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 7652-7660.

[6] ZHOU Y, TUZEL O. Voxelnet: End-to-end learning for point cloud based 3d object detection[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 4490-4499.

[7] QI C R, SU H, MO K, et al. Pointnet: Deep learning on point sets for 3d classification and segmentation[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 652-660.

[8] REN S, HE K, GIRSHICK R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks[C]// Advances in Neural Information Processing Systems. 2015: 91-99.

[9] HE K, ZHANG X, REN S, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(9): 1904-1916.

[10] GIRSHICK R. Fast r-cnn[C]// Proceedings of the IEEE International Conference on Computer Vision. 2015: 1440-1448.

[11] HE K, GKIOXARI G, DOLLÁR P, et al. Mask r-cnn[C]// Proceedings of the IEEE international conference on computer vision. 2017: 2961-2969.

[12] HINTERSTOISSER S, LEPETIT V, ILIC S, et al. Dominant orientation templates for real-time detection of texture-less objects[C]// 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE, 2010: 2257-2264.

[13] HINTERSTOISSER S, CAGNIART C, ILIC S, et al. Gradient response maps for real-time detection of textureless objects[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011, 34(5): 876-888.

[14] MOUSAVIAN A, ANGUELOV D, FLYNN J, et al. 3d bounding box estimation using deep learning and geometry[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 7074-7082.

[15] CHEN X, KUNDU K, ZHANG Z, et al. Monocular 3d object detection for autonomous driving[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 2147-2156.

[16] XU B, CHEN Z. Multi-level fusion based 3d object detection from monocular images[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 2345-2353.

[17] CHABOT F, CHAOUCH M, RABARISOA J, et al. Deep manta: A coarse-to-fine many-task network for joint 2d and 3d vehicle analysis from monocular image[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 2040-2049.

[18] LI B. 3d fully convolutional network for vehicle detection in point cloud[C]// 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2017: 1513-1518.

[19] ENGELCKE M, RAO D, WANG D Z, 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). IEEE, 2017: 1355-1361.

[20] SIMON M, MILZ S, AMENDE K, et al. Complex-YOLO: An Euler-Region-Proposal for Real-Time 3D Object Detection on Point Clouds[C]// European Conference on Computer Vision. Springer, Cham, 2018: 197-209.

[21] QI C R, YI L, SU H, et al. Pointnet++: Deep hierarchical feature learning on point sets in a metric space[C]// Advances in Neural Information Processing Systems. 2017: 5099-5108.

[22] 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). IEEE, 2018: 1-8.

[23] LIANG M, YANG B, WANG S, et al. Deep continuous fusion for multi-sensor 3d object detection[C]// Proceedings of the European Conference on Computer Vision (ECCV). 2018: 641-656.

[24] QI C R, LIU W, WU C, et al. Frustum pointnets for 3d object detection from rgb-d data[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 918-927.

SHEN Qi, born in 1996, postgraduate. Her main research interests include object detection and so on.

LIU Li-gang, born in 1975, Ph.D, professor, Ph.D supervisor, is a member of China Computer Federation. His main research interests include computer graphics and so on.