基于改进 Frustum PointNet 的 3D 目标检测

刘训华[1,2*]，孙韶媛[1,2]，顾立鹏[1,2]，李想[1,2]

[1]东华大学信息科学与技术学院，上海 201620；

[2]东华大学数字化纺织服装技术教育部工程研究中心，上海 201620

摘要 提出对图像和激光雷达点云数据进行 3D 目标检测的改进 F-PointNet(Frustum PointNet)。首先利用图像的 2D 目标检测模型提取目标 2D 区域，并将其映射到点云数据中，得到该目标的点云候选区域，然后预测候选区域的 3D 目标掩模，最后利用掩模对 3D 目标进行检测。当预测掩模时，提出的宽阈值掩模处理可以用来减少原始网络的信息损失；增加注意力机制可以获取需要被关注的点和通道层；使用 Focal Loss 可以解决目标与背景不平衡的问题。通过多次对比实验，证明宽阈值掩模处理可以提高 3D 目标检测的准确率，同时注意力机制和 Focal Loss 可以提高预测的准确率。

关键词 机器视觉；激光雷达；点云数据；3D 目标检测；宽阈值掩模处理

中图分类号 TN958.98 **文献标志码** A **doi**：10.3788/LOP57.201508

3D Object Detection Based on Improved Frustum PointNet

Liu Xunhua[1,2*], Sun Shaoyuan[1,2], Gu Lipeng[1,2], Li Xiang[1,2]

[1] *College of Information Science and Technology, Donghua University, Shanghai* 201620, *China*;

[2] *Engineering Research Center of Digitized Textile & Fashion Technology, Ministry of Education, Donghua University, Shanghai* 201620, *China*

Abstract An improved F-PointNet (Frustum PointNet) for 3D target detection on image and lidar point cloud data is proposed. First, the 2D target detection model of the image is used to extract 2D region of the target, and it is mapped to the point cloud data to obtain the candidate region of the target. Then, the 3D target mask of the candidate region is predicted. Finally, the 3D target is detected by using mask. When the mask is predicted, the proposed wide-threshold mask processing is used to reduce the information loss of the original network, the attention mechanism is added to obtain the points and channel layers that require attention, the Focal Loss can solve the imbalance between the target and the background problem. Through multiple comparison experiments, it is proved that wide-threshold mask processing can improve the accuracy of 3D target detection, and the attention mechanism and Focal Loss can improve the accuracy of prediction.

Key words machine vision; lidar; point cloud data; 3D object detection; wide-threshold mask processing

OCIS codes 150.6910; 150.4232; 100.4996

1 引 言

随着自动驾驶技术的发展，3D 目标检测变得越来越重要。3D 目标检测是在自动驾驶环境中感知动态障碍物，如车辆、行人和自行车[1]，实现难度大。基于图像的 3D 目标检测，其准确率较低，近年来 $In_{0.53}Ga_{0.47}As$(InGaAs)单光子探测阵列的深入研究推动了激光雷达的发展[2]，因此基于激光雷达以及基于图像与激光雷达融合的方案越来越多且准确率较高。融合的方案能够结合图像与激光雷达的优势，有效提高 3D 目标检测的准确率，但是如何融合才能减少较少的信息损失且同时剔除大量的冗余信息，这是提高 3D 目标检测效果的关键。

基于 RGB(Red，Green，Blue)图像和激光点云融

收稿日期：2019-12-24；**修回日期**：2020-02-19；**录用日期**：2020-03-09

基金项目：上海市科委基础研究项目(15JC1400600)

* **E-mail**：XunHua_LIU@163.com

合数据的3D目标检测方法主要有多视图融合检测框架和PointNet融合检测框架[3]。Chen等[4]提出的MV3D(Multi-View 3D)检测网络是典型的多视图融合检测结构,使用激光点云数据的前视图和鸟瞰图来表示三维点云信息,并与RGB图像融合以预测定向3D边界框,但是该网络使用的是转换成视图的点云数据,不是原始点云数据,这会造成信息损失。Qi等[5]提出的F-PointNet(Frustum PointNet)是典型的PointNet融合检测结构,通过Mask R-CNN(Region Convolutional Neural Networks)[6]结合深度信息以得到视锥目标区域,然后使用PointNet++[7]对得到的视锥区域进行目标的3D边界框回归估计,该网络是在原始点云上进行特征提取的,但是边界框预测直接使用了掩模结果,这依旧会存在信息损失的问题。万鹏[8]在F-PointNet的基础上修改参数初始化的方式、增加正则化以及减少T-Net(Transform Network)卷积核数,有效提升了检测精度,但是并未解决原始网络信息损失的问题。

经过掩模预测后,发现大部分非目标点的预测概率接近0,剩余部分目标点的预测概率既有大于0.5,也有小于0.5。因此本文对F-PointNet中掩模预测部分的网络结构进行改进,使用宽阈值掩模处理放宽掩模判定的边界,解决原始网络在该处损失信息的问题。同时增加注意力机制[9]并提取全局特征以进一步增加目标信息,并且更换损失函数为适合目标点较少、背景点较多的Focal Loss[10],通过多个对比实验,证明改进的F-PointNet可以提高3D目标检测的效果。

2 基本原理

2.1 整体网络框架

原有F-PointNet的基础架构由锥形点云候选区域提取部分、3D目标掩模预测部分以及3D目标边界框预测部分组成,对3D目标掩模预测部分的网络结构进行改进,改进的F-PointNet如图1所示。从图1可以看到,在3D目标掩模预测部分加入宽阈值掩模处理、注意力机制以及全局特征,并且使用Focal Loss损失函数计算掩模损失。

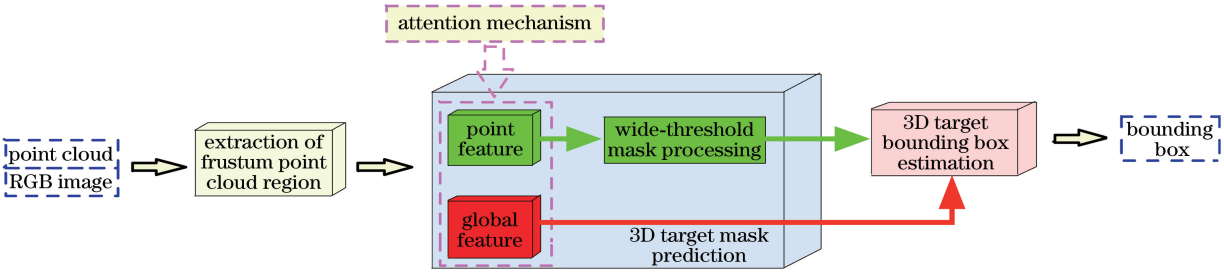

图1 改进的F-PointNet结构

Fig. 1 Improved F-PointNet structure

2.2 锥形点云候选区域的提取

锥形点云候选区域提取部分的整体网络结构如图2所示,其中k为类别数目,n为点云数,c为特征通道数。从图2可以看到,提取部分先采用基于FPN(Feature Pyramid Networks)[11]的2D目标检测模型,以获取RGB图像中的2D目标边界框并对其进行分类。然后使用已知的相机投影矩阵校准2D图像与3D点云数据,配准结果如图3所示。接着采用视锥的形式将2D边界框映射到点云数据中对应的位置,再收集视锥候选区域中的所有点以形成视锥体点云,进而初步得到3D目标视锥候选区域,如图4所示。最后,将3D目标视锥候选区域从摄像机坐标系转换至视锥坐标系中表示,即将视锥朝向调整到中心视角,视锥体的中心轴正交于图像平面,从而得到最终的3D目标视锥候选区域,如图5所示。

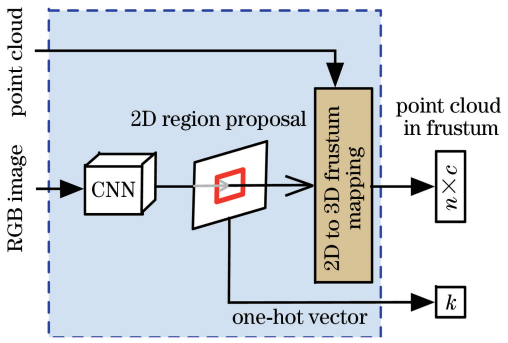

图2 锥形点云候选区域提取的网络结构

Fig. 2 Network structure for extracting candidate regions of frustum point cloud

2.3 3D目标掩模预测

3D目标掩模预测部分是将候选区域的点云数据(每个候选区域只包含一个感兴趣目标)输入到3D目标掩模预测网络中,以预测候选点云区域中每

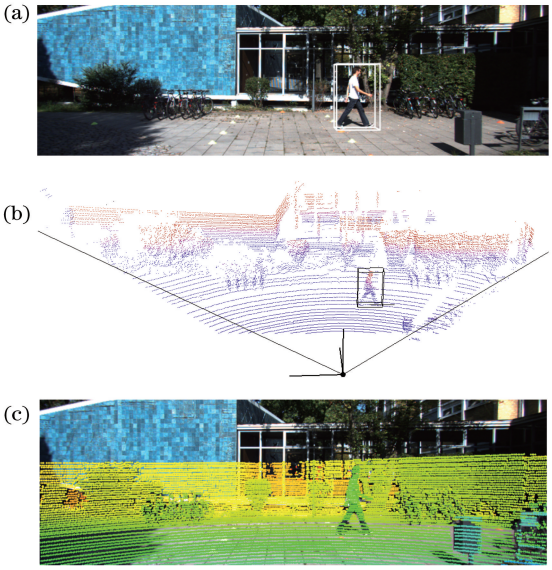

图 3 2D 图像与 3D 点云的配准结果。(a) RGB 图像；(b)3D 点云数据；(c)图(a)与图(b)的配准效果

Fig. 3 Registration results of 2D images and 3D point clouds. (a) RGB image; (b) 3D point cloud data; (c) registration effect of Fig. (a) and Fig. (b)

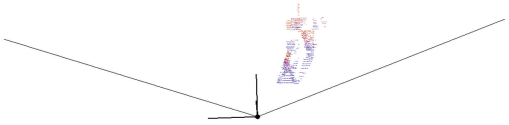

图 4 初步获取的 3D 目标视锥候选区域

Fig. 4 3D target frustum candidate region initially obtained

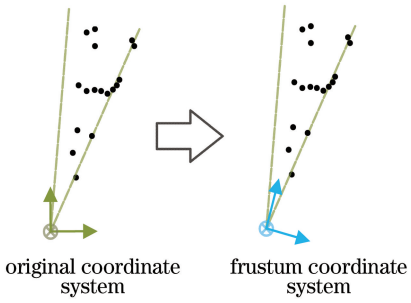

图 5 视锥朝向调整示意图

Fig. 5 Schematic of viewing frustum orientation adjustment

个点的概率得分，进而得到该候选区域的目标掩模预测结果，最后根据预测的掩模结果，在 3D 目标视锥候选区域点云中提取目标实例点云。3D 目标掩模预测网络如图 6 所示，其中 MLP 为多层感知机。从图 6 可以看到，3D 目标掩模预测网络主要由 PointNet＋＋构成。与原始网络不同，该网络是在此基础上使用宽阈值掩模处理，增加注意力机制且提取全局特征，并更换损失函数为 Focal Loss。宽阈值掩模处理能够丰富下一步 3D 目标边界框预测的输入信息，从而提高 3D 目标边界框预测的准确率。注意力机制能够找到点云数据中需要被关注的空间点和特征通道，结合全局特征可以有效增加目标信息。使用 Focal Loss 能够解决点云数据中目标与背景类别不平衡的问题，二者相结合有利于 3D 目标掩模的预测。

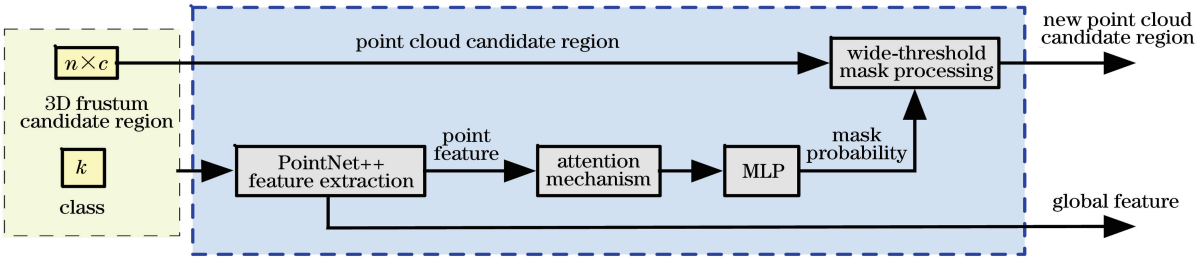

图 6 3D 目标掩模预测网络

Fig. 6 3D target mask prediction network

当预测 3D 目标边界框时，需要利用掩模结果来提取目标实例点云，这可能会出现由于目标掩模预测结果不准确而缺失有用信息的情况。因此采用宽阈值掩模处理放宽判定目标掩模的边界，从原来大于 1/2 为掩模，放宽为大于 $(1-x_{\text{margin}})/2$，其中 x_{margin} 为原阈值被放宽的程度，取值范围为 0～1。该处理过程不仅能够将大部分预测概率接近于 0 的非目标点剔除，而且能够有效保留预测概率小于 0.5 的目标点。通过多次实验，实验选取最佳的 $x_{\text{margin}}=0.2$，既减少在 3D 目标边界框预测输入的冗余信息，又增加有用信息，有利于获得更好的检测结果。

受到基于二维图像的混合域软注意力机制的启发，实验对 3D 点云数据的注意力机制分为空间域和通道域。空间域注意力机制能够找到 3D 点云数据中需要关注的点，通道域能够找到特征图中需要关注的通道层，实现方式均采用基础的注意力模型。图 7 为通道/空间域注意力机制实现流程，FC 为全连接层，其中输入为上一层的特征，输出为经过注意力机制处理后的特征。

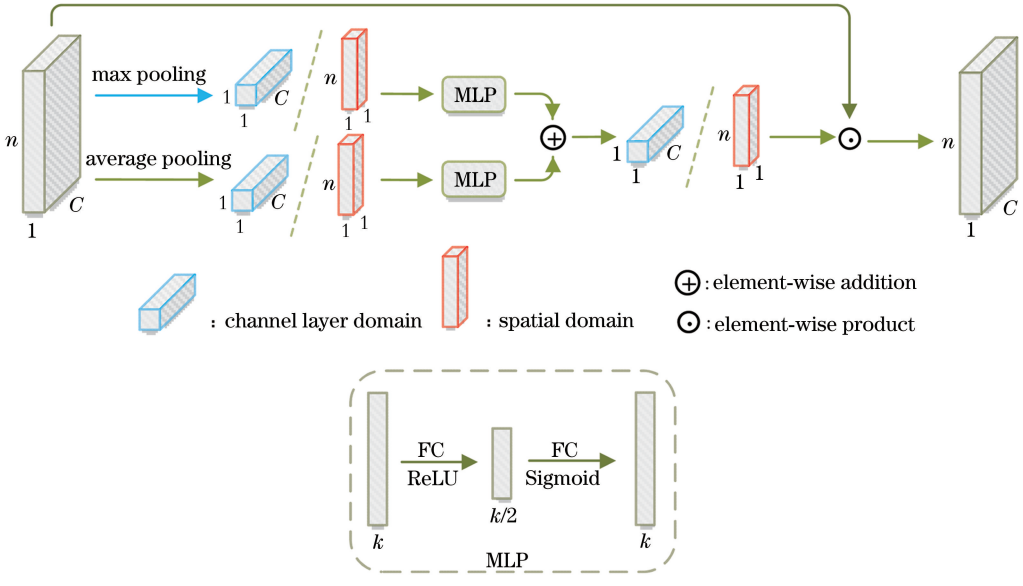

图 7 注意力机制实现流程

Fig. 7 Attention mechanism implementation process

原始掩模预测网络使用的是普通的交叉熵损失函数，但是由于点云数据中目标前景点与背景点的数量比例十分不平衡，拥有大量的背景无关点。普通的交叉熵损失函数无法挖掘前景点的重要性，因此实验中使用 Focal Loss 损失函数，其能够降低大量背景点在训练中所占的权重，使得训练更关注前景点即目标物体，掩模预测损失计算公式为

$$L_{\text{mask}}=L_{\text{FL}}=\begin{cases}-\alpha(1-p)^{\gamma}\ln(p), & \text{if } y=1\\ -(1-\alpha)p^{\gamma}\ln(1-p), & \text{otherwise}\end{cases},\tag{1}$$

式中：p 为预测概率；α 为权重因子，能够平衡正负样本的重要性；γ 为调制因子，能够调节样本权重降低的速率；y 为真值的类别，对于 $y=1$，预测概率为 p。实验过程中，α 取 0.25，γ 取 2。

2.4 3D 目标边界框预测

3D 目标边界框预测部分是对 2.3 节得到的目标实例点云和全局特征进行 3D 目标检测，网络结构如图 8 所示。首先，将目标实例点云从视锥坐标系转换至 3D 目标掩模中心坐标系中表示。然而由于点云数据的采集特性，即使是真实的目标实例点云也无法形成完整的物体，故 3D 目标掩模中心并不是目标边界框的中心，即真实完整目标的中心，因此使用与 STN（Spatial Transformer Network）[12] 相似的 T-Net 来估计目标边界框的中心，然后转换至目标边界框中心坐标系中表示，坐标系转换结果如图 9 所示。最后使用 PointNet++ 以及多层感知机，并采用残差回归的方式来实现目标实例点云的 3D 边界框回归估计，其中回归参数包括 3D 目标边界框的中心坐标、长宽高以及目标的方位角。

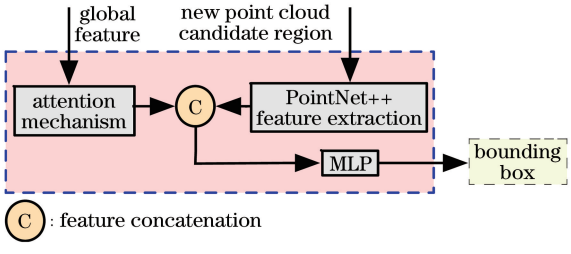

图 8 3D 目标边界框预测网络

Fig. 8 3D target bounding box prediction network

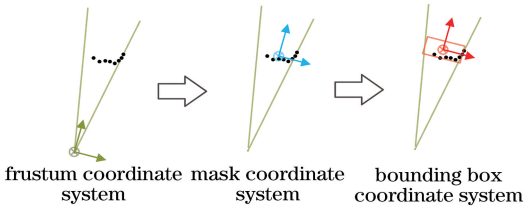

图 9 目标实例点云的坐标转换

Fig. 9 Coordinate transformation of target instance point cloud

2.5 损失函数

改进的 F-PointNet 使用与原始网络一样的多任务损失函数 $L_{\text{multi-task}}$，其包含 3D 目标掩模预测损失函数 L_{mask} 和 3D 目标边界框预测损失函数 L_{bbox}。3D 目标边界框预测损失函数与原始网络相同，包含 T-Net 产生的损失函数 $L_{\text{c1-reg}}$、目标边界框中心预测的损失函数 $L_{\text{c2-reg}}$、目标方位角的分类损失函数 $L_{\text{h-cls}}$ 和回归损失函数 $L_{\text{h-reg}}$、目标边界框大小的分类

损失函数 $L_{s\text{-}cls}$ 和回归损失函数 $L_{s\text{-}reg}$、目标边界框中 8 个顶点的损失函数 L_{corner}。但是 3D 目标掩模预测损失函数与原始网络不同，使用 2.3 节(1)式的 Focal Loss。$L_{multi\text{-}task}$ 计算公式为

$$L_{multi\text{-}task}=L_{mask}+L_{bbox}=L_{mask}+\lambda(L_{c1\text{-}reg}+L_{c2\text{-}reg}+L_{h\text{-}cls}+L_{h\text{-}reg}+L_{s\text{-}cls}+L_{s\text{-}reg}+\gamma L_{corner})。\tag{2}$$

$L_{h\text{-}cls}$ 和 $L_{s\text{-}cls}$ 使用 Softmax 交叉熵损失函数，L_{mask} 使用 Focal Loss，$L_{c1\text{-}reg}$、$L_{c2\text{-}reg}$、$L_{h\text{-}reg}$ 和 $L_{s\text{-}reg}$ 使用 L1 范数损失函数损失，λ 和 γ 沿用原始网络中的设置，L_{corner} 是指预测的边界框与真值之间对应 8 个顶点的距离之和。

3 分析与讨论

3.1 实验配置与数据

所提网络使用 TensorFlow 框架来实现，数据使用公开的 KITTI 数据集[13]，实验配置如表 1 所示。

表 1 实验配置

Table 1 Experimental configuration

Item	CPU	Computing memory	GPU	System	CUDA
Content	Intel i5-6600	8 GB	NVIDIA GTX 1070	Ubuntu 16.04	CUDA 9.0

3.2 实验步骤

首先在 ImageNet 分类和 COCO 目标检测数据集上对 2D 目标检测模型的权重进行预训练，在 KITTI 2D 目标检测数据集上对预训练后的数据集进行微调。然后使用训练好的 2D 目标检测器得到 2D 目标，再提取 KITTI 数据集中的 3D 点云候选数据并进行存储。最后，使用存储的候选区域点云数据进行 3D 目标掩模预测和 3D 目标边界框预测的联合训练。模型训练过程中，先进行宽阈值掩模实验，实验时设定多个阈值并从中选择平均准确率(AP)最高的阈值，然后在该阈值下对各处理部分进行对比，最后将最终改进模型的结果与其他模型进行比较。学习率选择 0.001，其随训练衰减，优化器选择 Adam 算法，Batch 大小为 12，训练执行 100 轮。

3.3 实验结果及分析

可视化的候选区域提取结果如图 4 所示，图 10 为可视化的 3D 目标检测结果，其中图 10(a)为 2D 目标检测结果，图 10(b)为 3D 目标检测结果，该图仅给出估计的边界框，不包含真值及其类别。为了确定最佳掩模判定的阈值而进行多次实验，设置 x_{margin} 值分别为 0、0.1、0.2 和 0.3，AP 值如表 2 所示，easy 为目标全部可见，moderate 为目标部分遮挡，hard 为目标难以看见。从表 2 可以看到，宽阈值掩模处理确实可以提高 3D 目标检测的结果，且当 $x_{margin}=0.2$ 时，在汽车、行人和骑自行车的人中 AP 值最大。

表 2 各阈值下 3D 目标检测的 AP 值

Table 2 AP values of 3D target detection under each threshold unit: %

x_{margin}	Car			Pedestrian			Cyclist		
	Easy	Moderate	Hard	Easy	Moderate	Hard	Easy	Moderate	Hard
0	82.05	68.46	62.42	65.94	58.35	50.87	74.10	55.54	52.09
0.1	82.39	69.53	62.52	61.90	55.20	49.02	73.45	55.46	52.26
0.2	82.79	**70.85**	**63.49**	**67.05**	**59.16**	**51.82**	**76.04**	**57.09**	**53.33**
0.3	**83.19**	70.59	63.13	65.06	57.53	50.59	73.55	55.76	52.73

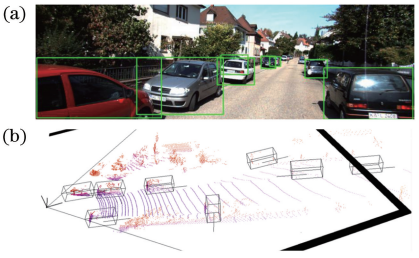

图 10 可视化的 3D 目标边界框预测结果。(a) 2D 目标检测结果；(b) 3D 目标检测结果

Fig. 10 Visual 3D target bounding box prediction results. (a) 2D target detection result; (b) 3D target detection result

为了验证各个处理部分对原始网络的影响程度，仅对汽车类别进行对比实验，结果如表 3 所示。从表 3 可以看到，宽阈值掩模处理可以有效提高 3D 目标检测的 AP 值，注意力机制和 Focal Loss 对于模型略有帮助。最后，将所提模型与其他模型进行比较，结果如表 4 所示。从表 4 可以看到，与原始 F-PointNet 相比，改进的 F-PointNet 可以提升整体的检测精度，同时与 UberATG-ContFuse[14] 相比，仅 hard 结果略低，与 MLOD (Multi-view Labelling Object Detector)[15] 相比整体的准确率较高。

表 3 各处理部分对 AP 值的影响

Table 3 Influence of each processing part on AP values

Part			AP /%		
Wide-threshold mask ($x_{\text{margin}}=0.2$)	Attention mechanism	Focal Loss	Easy	Moderate	Hard
—	—	—	82.05	68.46	62.42
√	—	—	82.79	70.85	63.49
—	√	—	81.89	69.23	62.54
—	—	√	82.73	69.89	63.27
√	√	√	**83.04**	**71.25**	**63.82**

表 4 不同模型的 AP 值对比

Table 4 Comparison of AP values of different models

Method	AP /%		
	Easy	Moderate	Hard
MV3D[4]	71.29	62.28	56.56
F-PointNet[5]	82.05	68.46	62.42
UberATG-ContFuse[14]	82.54	66.22	**64.04**
MLOD[15]	72.24	64.20	57.20
Proposed	**83.04**	**71.25**	63.82

4 结 论

为了提高 3D 目标检测的准确率，对 F-PointNet 的掩模预测部分进行改进，使用宽阈值掩模处理，增加注意力机制并更换损失函数为 Focal Loss。实验结果表明，宽阈值掩模处理在减少点云冗余信息的同时，能够充分利用其有用信息，最终与基于其他网络的 3D 目标检测相比，基于改进的 F-PointNet 的 3D 目标检测可以获得较优的结果。但针对无人驾驶领域的 3D 目标检测研究，准确率还有提升的空间，后续的研究将进一步在网络结构设计及目标检测过程的机理上深入挖掘。

参 考 文 献

[1] Zhang Y, Ren G Q, Cheng Z Y, et al. Application research of there-dimensional LiDAR in unmanned vehicle environment perception [J]. Laser & Optoelectronics Progress, 2019, 56(13): 130001.
张银, 任国全, 程子阳, 等. 三维激光雷达在无人车环境感知中的应用研究[J]. 激光与光电子学进展, 2019, 56(13): 130001.

[2] Liu K B, Yang X H, He T T, et al. Indium phosphide-based near-infrared single photon avalanche photodiode detector arrays [J]. Laser & Optoelectronics Progress, 2019, 56(22): 220001.
刘凯宝, 杨晓红, 何婷婷, 等. InP 基近红外单光子雪崩光电探测器阵列[J]. 激光与光电子学进展, 2019, 56(22): 220001.

[3] Ji Y M, Chen Z Y, Tian P H, et al. A survey of 3D target detection methods in unmanned driving [J]. Journal of Nanjing University of Posts and Telecommunications (Natural Science Edition), 2019, 39(4): 72-79.
季一木, 陈治宇, 田鹏浩, 等. 无人驾驶中 3D 目标检测方法研究综述[J]. 南京邮电大学学报(自然科学版), 2019, 39(4): 72-79.

[4] Chen X Z, Ma H M, Wan J, et al. Multi-view 3D object detection network for autonomous driving[C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE, 2017: 6526-6534.

[5] 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, 2018: 918-927.

[6] He K M, Gkioxari G, Dollar P, et al. Mask R-CNN [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 42(2): 386-397.

[7] Qi C R, Yi L, Su H, et al. PointNet++: deep hierarchical feature learning on point sets in a metric space[EB/OL]. (2017-07-07)[2019-12-23]. https://arxiv.org/abs/1706.02413.

[8] Wan P. Object detection of 3D point clouds based on F-PointNet [J]. Journal of Shandong University (Engineering Science), 2019, 49(5): 98-104.
万鹏. 基于 F-PointNet 的 3D 点云数据目标检测[J]. 山东大学学报(工学版), 2019, 49(5): 98-104.

[9] Woo S, Park J, Lee J Y, et al. CBAM: convolutional block attention module[M]//Ferrari V, Hebert M, Sminchisescu C, et al. Computer vision-ECCV 2018. Lecture notes in computer science. Cham: Springer, 2018, 11211: 3-19.

[10] 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.

[11] Lin T Y, Dollár P, Girshick R, et al. Feature pyramid networks for object detection[C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE, 2017: 936-944.

[12] Jaderberg M, Simonyan K, Zisserman A, et al. Spatial transformer networks[EB/OL]. (2016-02-04)[2019-12-23]. https://arxiv.org/abs/1506.02025.

[13] Geiger A, Lenz P, Urtasun R. Are we ready for autonomous driving? The KITTI vision benchmark suite[C]//2012 IEEE Conference on Computer Vision and Pattern Recognition, June 16-21, 2012, Providence, RI, USA. New York: IEEE, 2012: 3354-3361.

[14] Liang M, Yang B, Wang S L, et al. Deep continuous fusion for multi-sensor 3D object detection[M]// Ferrari V, Hebert M, Sminchisescu C, et al. Computer vision-ECCV 2018. Lecture notes in computer science. Cham: Springer, 2018, 11220: 663-678.

[15] Deng J, Czarnecki K. MLOD: a multi-view 3D object detection based on robust feature fusion method[C]// 2019 IEEE Intelligent Transportation Systems Conference (ITSC), October 27-30, 2019, Auckland, New Zealand. New York: IEEE, 2019: 279-284.