Artificial Intelligence and Robotics Research 人工智能与机器人研究, 2023, 12(4), 319-327
Published Online November 2023 in Hans. https://www.hanspub.org/journal/airr
https://doi.org/10.12677/airr.2023.124035

基于注意力机制的改进PointPillars三维目标检测

司文悦[1,2]，高　佼[1,2]，李国栋[1,2]，张吉卫[1,2]

[1]山东交通学院，轨道交通学院，山东　济南
[2]山东省交通运输行业轨道交通安全技术与装备重点实验室，山东　济南

收稿日期：2023年9月30日；录用日期：2023年11月10日；发布日期：2023年11月22日

摘　要

针对传统三维点云目标检测算法对小目标检测精度低的弱点，提出一种基于空间注意力机制的改进PointPillars方法。首先，在pillar特征网络中增加点云特征表示来丰富特征编码，提高每个点的表征能力，其次，在伪图像上通过空间注意力机制重新计算编码后空间点的特征权重，增强算法特征提取能力，提高检测性能，最后，利用公开数据集KITTI对改进算法进行验证。实验结果表明，该方法能够准确地检测出小尺寸行人和骑行者目标，同时在大尺寸汽车目标检测上保持稳定性能。此外，在中等检测难度条件下，三维模式、鸟瞰图模式和平均方向相似度模式三个类别平均精度均值(mAP)分别达到了62.07%、68.85%和70.02%，较改进前算法均有较大提升。

关键词

三维点云，目标检测，注意力机制，PointPillars

Improved PointPillars 3D Object Detection Based on Attention Mechanism

Wenyue Si[1,2], Jiao Gao[1,2], Guodong Li[1,2], Jiwei Zhang[1,2]

[1]School of Rail Transit, Shandong Jiaotong University, Jinan Shandong
[2]Shandong Province Key Laboratory of Rail Transit Safe Technology and Equipment in Transportation Industry, Shandong Jiaotong University, Jinan Shandong

Received: Sep. 30th, 2023; accepted: Nov. 10th, 2023; published: Nov. 22nd, 2023

Abstract

Aiming at the weaknesses of traditional 3D point cloud object detection algorithms with low de-

文章引用: 司文悦, 高佼, 李国栋, 张吉卫. 基于注意力机制的改进 PointPillars 三维目标检测[J]. 人工智能与机器人研究, 2023, 12(4): 319-327. DOI: 10.12677/airr.2023.124035

tection accuracy for small objects, an improved PointPillars method based on spatial attention mechanism is proposed. Firstly, the point cloud feature representation is added to the pillar feature network to enrich the feature encoding and improve the representation ability of each point, secondly, the feature weights of the encoded spatial points are recalculated on the pseudo-image by the spatial attention mechanism, which enhances the algorithm's feature extraction ability and improves the detection performance, and lastly, the improved algorithm is validated by using the publicly available dataset KITTI. The experimental results show that the method is able to accurately detect small-size pedestrian and cyclist object, while maintaining stable performance on large-size car object detection. In addition, the mean average precision (mAP) of the three categories of 3D mode, bird's-eye view mode, and average orientation similarity mode reached 62.07%, 68.85%, and 70.02%, respectively, under the medium detection difficulty condition, which are all greatly improved over the pre-improvement algorithm.

Keywords

3D Point Cloud, Object Detection, Attention Mechanism, PointPillars

1. 引言

三维激光雷达点云目标检测是自动驾驶[1]、智能机器人以及智能交通等领域的一项关键技术，受到了国内外学者的普遍关注。不同于传统图像数据，三维点云数据具有直接空间信息，可以为实现三维场景理解和环境感知提供更加真实和全面的视角。然而，点云数据具有无序性、稀疏性、置换不变性等特殊性质，同时含有较多噪点，为在遮挡程度较高的环境中实现高效且准确的目标检测带来极大困难。

近年来，为了克服点云数据的缺点，国内外学者提出了多种不同的三维点云目标检测方法。Charles等[2]提出的 PointNet 算法直接逐点处理点云，完成特征学习，该方法能够保留点云的原始信息，但无法解决点云无序性和稀疏性带来的特征提取困难及计算效率低下问题；Zhou 等[3]提出的 VoxelNet 算法将空间划分为体素，对每个体素应用 PointNet，然后使用 3D 卷积中间层来巩固垂直轴，最后应用 2D 卷积检测体系结构，虽然这种方法在一定程度上克服了点云数据的不规则性，但由于体素化过程会导致模型推理时间延长，无法实时部署；Yan 等[4]提出的 SECOND 算法有效提高了 VoxelNet 的推理速度，但其中的 3D 卷积运算成为影响算法实时性的瓶颈。为了提高目标检测算法的运行速度，Lang 等[5]提出的 PointPillars 将点云数据划分为一系列垂直排列的支柱，然后通过 2D 卷积在支柱表示上进行特征提取，该方法可以更有效地处理大规模点云数据并保留重要信息，并且在速度和准确性之间取得了完美的平衡，但存在诸如小尺寸目标检测效果差、遮挡漏检等问题需要解决。

为了弥补经典 PointPillars 算法的缺点，本研究在经典 PointPillars 算法的基础上，在支柱特征网络中增加点云特征表示来丰富特征编码，以提高每个点的表征能力。随后，在伪图像上应用空间注意力机制来重新计算编码后空间点的特征权重，从而增强算法的特征提取能力，进一步提高目标检测性能。最后的实验结果表明，这一改进方法能够精确地检测小尺寸行人和骑行者目标，精确预测目标的朝向，同时在大尺寸汽车目标检测方面保持稳定的性能。

2. PointPillars 算法原理

相较于其他三维目标检测算法，PointPillars 算法采用点云立柱化的表征方式，将三维点云转化成二维伪图像，然后使用二维的 Backbone 进行特征提取[6]，模型主要包括三大模块：1) 支柱特征网络模块，完成点云到伪图像的转换；2) Backbone 主干特征提取模块，使用二维卷积神经网络(2DCNN)提取不同支柱之间的空间和语义特征；3) SSD 检测头模块，完成包围框的回归和目标的分类。具体的检测过程如下文所述。

图 1 所示为支柱特征网络模块结构图。首先，将 X-Y 平面上的输入激光雷达点云等间隔离散化，构成一组支柱(pillars)；其次，通过支柱堆叠创建维度为(D, P, N)的密集张量，其中 D 表示点的特征维度数，P 表示非空支柱的数量，N 表示每个支柱包含点的个数；再次，使用 PointNet [2]简化版网络结构，对于柱中每个点，经过线性层，批归一化(BatchNorm)层和 ReLU 激活层，输出维度为(C, P, N)的张量；最后，对多通道特征图进行最大池化(MaxPool)创建维度为(C, P)的输出张量，并通过编码将它们分散回原始支柱位置，获得维度为(C, H, W)的伪图像，其中 C 表示伪图像通道数，H 和 W 表示伪图像的高度和宽度。

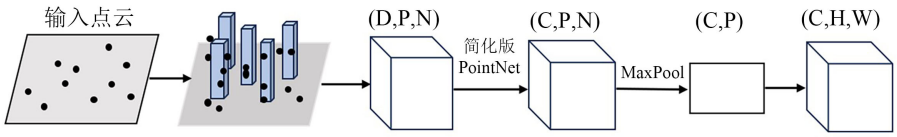


Figure 1. Pillar feature network structure diagram
图 1. 支柱特征网络结构图

图 2 所示为 Backbone 主干特征提取模块结构图。该模块包含下采样和上采样两个子网络，下采样网络由三个自上而下的卷积核为 3 × 3 的二维卷积(Conv)模块组成，三个模块层数依次为 4，6，6，输出特征维度分别为 64，128，256，在每层卷积之后均依次经过 BatchNorm 层和 ReLU 激活层；上采样网络由三个二维反卷积(Deconv)模块组成，用于将三个下采样块输出的不同维度特征上采样为 128 维特征，特征经过 BatchNorm 层和 ReLU 激活层后通过特征拼接 Concat 模块连接在一起，最终输出 384 维特征。

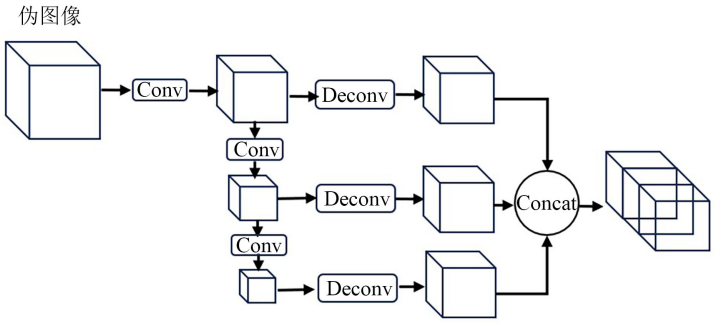


Figure 2. Backbone feature extraction network structure diagram
图 2. 主干特征提取网络结构图

三维点云经支柱特征网络和主干特征提取网络，顺利完成目标特征提取。将提取的特征图输入 SSD [7]检测头模块，进行包围框(bounding box)回归和目标分类，得到物体的位置和种类[6]。

3. PointPillars 算法改进

在传统 PointPillars 算法中，点云支柱特征编码部分容易导致特征丢失，同时主干特征提取网络对伪图像的特征提取不够充分，因此存在目标检测精度低的问题。为了提高检测性能，在原支柱特征网络部

分添加点云编码特征，对特征编码进行丰富；此外，在主干特征提取之前添加空间注意力机制(Spatial Attention, SAt)，对伪图像空间点细微特征进行提取。

改进后网络结构如图 3 所示。其中 Backbone 主干特征提取保持原始 PointPillars 中的结构，检测头部分使用二维卷积网络训练汽车、行人、骑行者三个类别，先验框与真实框匹配过程不考虑高度信息，使用 2D (IoU) [8]匹配方式。

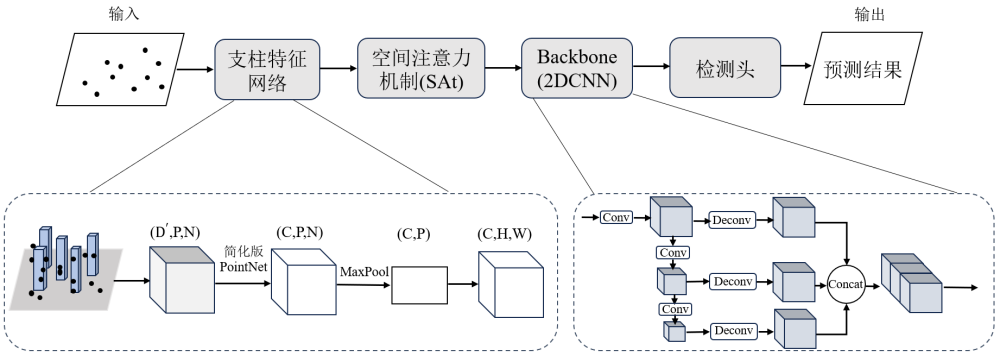


Figure 3. Improved network structure diagram
图 3. 改进后网络结构图

3.1. 支柱特征网络

原 PointPillars 模型在通过支柱特征网络模块对点云进行编码时，每个支柱中激光雷达点 s 维度数 D = 9，点 s 编码如式(1)所示：

$$s = \left[x, y, z, r, x_c, y_c, z_c, x_p, y_p \right] \tag{1}$$

式中，$[x, y, z]$是激光雷达坐标系中点的空间坐标，r 是反射强度，$[x_c, y_c, z_c]$表示点 s 的$[x, y, z]$到当前支柱中所有点空间坐标的算术平均值的距离，$[x_p, y_p]$表示点 s 的(x, y)距当前支柱中心(x, y)的偏移量。

可以看出，原 PointPillars 算法仅考虑了点的空间坐标与支柱中所有点的空间坐标均值的偏差，没有考虑点的反射率强度偏差特征。在某些场景中，目标可能与背景具有相似的几何特征，但反射率强度特征有明显差异，通过分析反射率强度偏差，可以更准确地将目标与周围环境分离开来。其次，在光照条件变化或目标具有复杂反射特性的情况下，有效地捕获每个点的相对亮度或暗度信息，可以更好地区分不同表面特性的目标。因此，在点云特征表示方面，在点 s 上增加当前点的反射率到支柱中所有点反射率均值的偏差特征r_c，使特征维度数 D' = 10，增强后的特征编码如式(2)所示：

$$s' = \left[x, y, z, r, x_c, y_c, z_c, r_c, x_p, y_p \right] \tag{2}$$

3.2. 注意力模块

注意力模块的主要功能是通过添加注意力机制来增加数据的表征能力，使网络学习特征中的重要信息并抑制不重要信息[9]，常见的有通道注意力机制 SE [10]、ECA [11]，空间注意力机制 FPN [12]、PAN [13]、混合注意力机制 CBAM [14]等[15]。为了使模型动态地学习点云数据的空间关联性，使模型可以更好地理解支柱编码生成的伪图像中不同点之间的相对位置和分布，受注意力机制 CBAM 的启发，在支柱特征网络生成的伪图像上使用如图 4 所示的空间注意力机制(SAt)，更准确地捕获不同目标的几何形状和空间特征。空间注意力模块如式(3)所示：

$$F_{\mathrm{SAt}} = \sigma\left(\mathrm{f}^{3\times 3}\left(\left[\mathrm{AvgPool}(F); \mathrm{MaxPool}(F) \right] \right) \right) \tag{3}$$

式中，F_{SAt} 为 SAt 的输出数据，F 为输入数据，[• ； •]表示沿通道维度的拼接操作，MaxPool(F)和AvgPool(F)分别表示通道维度的最大池化和平均池化；$f^{3\times3}$是 3 × 3 卷积，其输入和输出通道数分别为 2 和 1。最大池化和平均池化操作在特征提取方面呈现出多尺度的特性，能够捕获不同尺度下的特征信息差异，通过将最大池化和平均池化的结果进行拼接并应用 3 × 3 卷积，模型可以同时考虑不同尺度的特征信息，在汽车、行人和骑行者等不同目标的检测方面，这种多尺度信息整合有助于提高模型不同尺度尤其是小尺度目标的检测性能。σ 表示 sigmoid 激活函数，用于对注意力权重进行归一化，确保每个点的注意力权重在 0 到 1 之间，让模型可以对每个点进行不同程度的关注，将更多的注意力集中在可能包含目标的区域，从而提高检测效率。

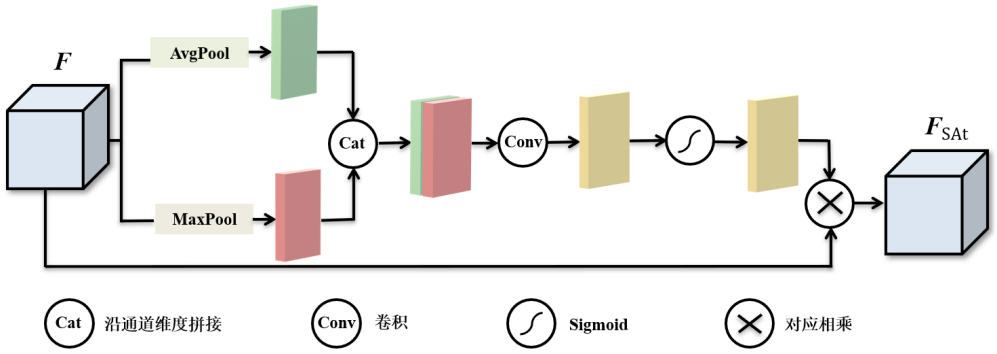


Figure 4. Spatial attention mechanism SAt structure
图 4. 空间注意力机制 SAt 结构图

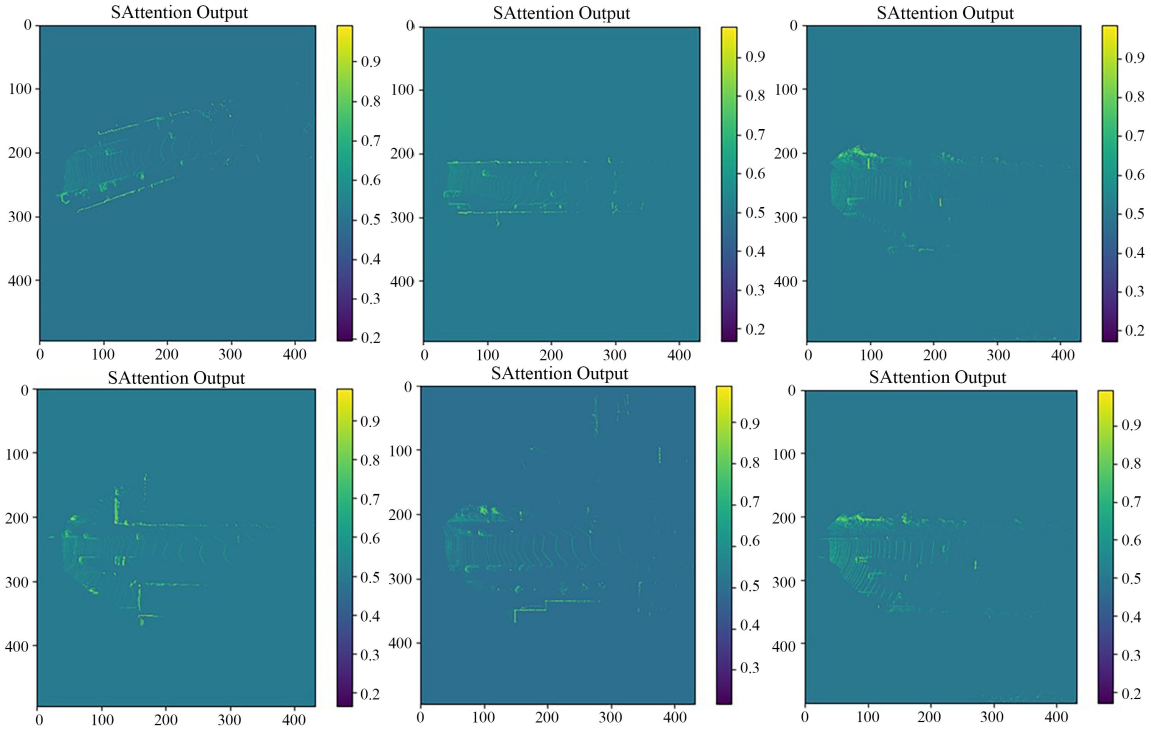


Figure 5. Plot of weights for different samples
图 5. 不同样本权值图

该注意力机制可以以较低的模型复杂度学习特征图中每一个空间点的权重，通过权重的计算和特征

的更新，增强对点云数据空间关系的建模能力，提高模型性能。图 5 所示为使用改进 PointPillars 网络模型对数据集中不同环境下的不同测试样本引入 SAt 的可视化验证。图像亮度对应利用 Sigmoid 函数对注意力权重进行归一化后，获得的 0 到 1 之间的权重值。

从图 5 可以看出：含有目标的区域点权重值较高，不含目标的区域点权重值相对较低，意味着 SAt 有效起到了特征筛选作用。

4. 实验与结果分析

使用官方 KITTI [16]数据集进行实验，实验环境为 Linux 操作系统，Python3.7，Pytorch1.10，Cuda11.1，主机配置有 Intel(R) Silver 4210R CPU，GPU 为 NVIDIA RTX A5000，显存为 24 G。本小节详细介绍 KITTI 数据集、网络参数设置、损失函数以及结果分析。

4.1. KITTI 数据集

官方 KITTI 数据集根据目标遮挡程度、目标大小、点云噪声、天气状况等条件将目标检测分为简单、中等、困难三个级别，数据集中训练样本、测试样本总个数分别为 7481、7518。本文在激光雷达点云上进行训练，在实验过程中，使用 3712 个训练样本和 3769 个验证样本[9]，同时为了增强点云数据多样性，采取随机镜像翻转、全局旋转和缩放等点云增强操作。

4.2. 网络参数设置

网络在训练过程中，使用的所有权重都是均匀分布且随机初始化。支柱特征网络中支柱长宽为 0.16 m、高度为 4 m，每个支柱最多包含 32 个点，训练阶段可以使用的最大体素数 16,000，测试阶段可以使用的最大体素数 40,000。优化器采用 Adam、权重衰减系数为 0.01，动量值为 0.9，初始学习率为 0.003，学习率衰减值为 0.1，最大迭代次数为 90。

4.3. 损失函数

使用 PointPillars 中引入的损失函数，定位损失使用 SmoothL1 损失；为了区分目标方向，减小目标朝向反向误差，在离散方向上使用 SECOND 网络中的 Softmax 分类损失；目标类别分类损失使用 Focal Loss 损失。

4.4. 结果分析

4.4.1. 实验定量分析

使用平均精度(AP)指标在 KITTI 数据集三个级别难易程度上对改进算法和原有 PointPillars、VoxelNet、Second 等算法的目标检测性能进行评估与对比分析。汽车的 IOU 阈值为 0.7，行人以及骑行者的 IOU 阈值为 0.5。

Table 1. KITTI dataset 3D model evaluation results
表 1. KITTI 数据集三维模式评估果

方法	汽车行人骑行者									mAP
	简单中等困难			简单中等困难			简单中等困难			
VoxelNet [3]	77.47	65.11	57.73	39.48	33.69	31.50	61.22	48.36	44.37	49.05
Second [4]	83.13	73.66	66.20	51.07	42.56	37.29	70.51	53.85	46.90	56.69
PointPillars [5]	79.05	74.99	68.30	52.08	43.53	41.49	75.48	59.07	52.92	59.20
改进 PointPillars	85.59	75.97	71.99	53.50	47.15	43.23	80.71	63.11	60.14	**62.07**

Table 2. KITTI dataset BEVmodel evaluation results
表 2. KITTI 数据集鸟瞰图模式评估结果

方法	汽车行人骑行者									mAP
	简单中等困难			简单中等困难			简单中等困难			
VoxelNet [3]	89.35	79.26	77.39	46.13	40.74	38.11	66.70	54.76	50.55	58.25
Second [4]	88.07	79.37	77.95	55.10	46.27	44.76	73.67	56.04	48.78	60.56
PointPillars [5]	88.35	86.10	79.83	58.66	50.23	47.19	79.14	62.25	56.00	66.19
改进 PointPillars	89.85	87.19	83.66	58.62	53.46	48.73	82.86	65.90	62.45	**68.85**

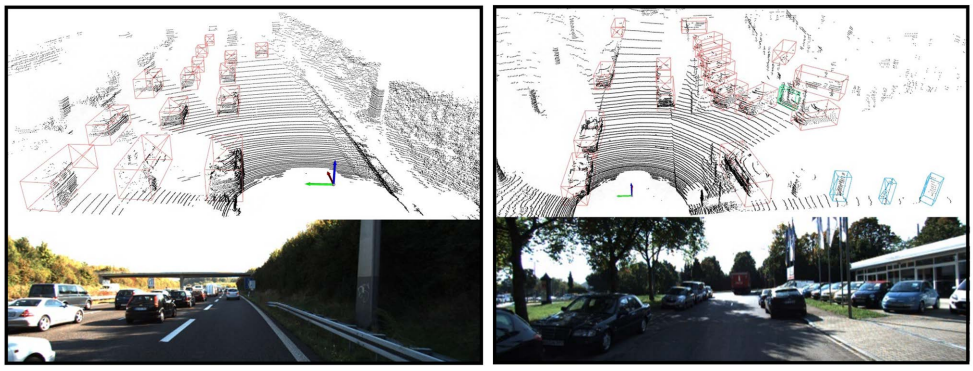
Table 3. KITTI dataset AOS model evaluation results
表 3. KITTI 数据集 AOS 模式评估结果

方法	汽车行人骑行者									mAP
	简单中等困难			简单中等困难			简单中等困难			
Second [4]	87.84	81.31	71.95	55.56	43.51	38.78	80.97	57.20	55.14	54.53
PointPillars [5]	90.19	88.76	86.83	58.05	49.66	47.88	82.43	68.16	61.96	68.86
改进 PointPillars	90.62	89.91	87.20	50.08	45.74	43.25	85.75	75.42	71.72	**70.02**

表 1~表 3 分别表示 KITTI 数据集三维模式、鸟瞰图(BEV)模式以及平均方向相似度(AOS)模式的评估与对比结果。其中，mAP 为在中等难度条件下，对三个类的平均精度(AP)进行平均计算得到。

4.4.2. 实验定性分析

改进后的 PointPillars 算法在 KITTI 数据集的部分检测结果如图 6 所示。图中上半部分为在点云场景中的三维检测效果图，下半部分为对应的相机中的真实场景图。检测结果用不同颜色的框表示不同的类别，红色框代表汽车、蓝色框代表行人、绿色框代表骑行者，目标朝向由三维框的交叉线表示，含有交叉线的方向为目标的前向。



(a) (b)

Figure 6. A classic example of spatial data sets
图 6. 空间数据集经典例子

可以看出：对于远距离场景图 6(a)以及较为复杂的场景图 6(b)，改进算法依然可以实现较准确的检测；从目标的朝向角度分析，本文算法能准确预测出目标朝向。

此外，改进算法还可以较好的减小误检率，如图 7 所示，图 7(a)为原始 PointPillars 算法的检测结果，

图 7(b)为本文算法的检测结果，图 7(c)为对应的相机中的真实场景图。改进算法没有将相机图像中紫色标注框的电线杆误检成行人。

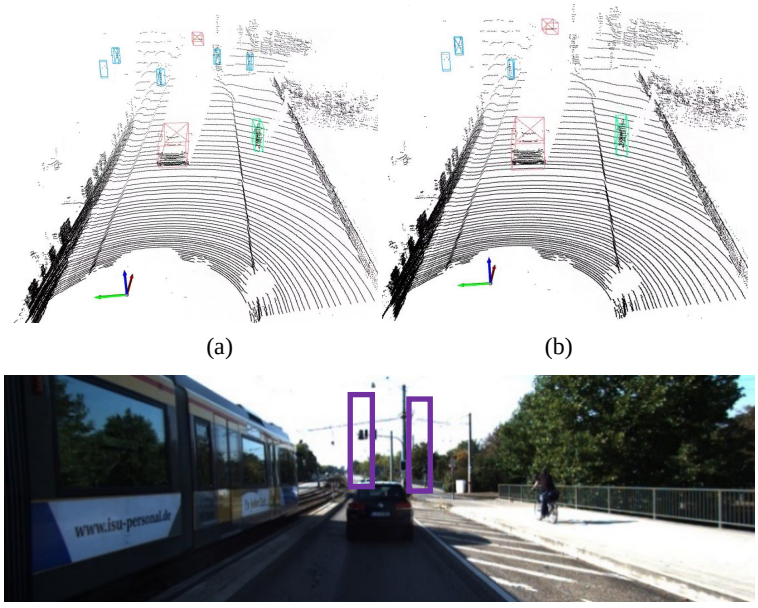


Figure 7. Comparison of detection results between this algorithm and the original PointPillars algorithm
图 7. 本文算法与原 PointPillars 算法检测结果对比

最后，改进算法可以较好的减小漏检率，如图 8 所示，漏检车辆用紫色边框标出，图 8(a)为原始 PointPillars 算法的检测结果，图 8(b)为本文算法的检测结果，图 8(c)为对应的相机中的真实场景图。改进算法可以检测出被前方车辆遮挡相机图像中紫色标注框的后方车辆。

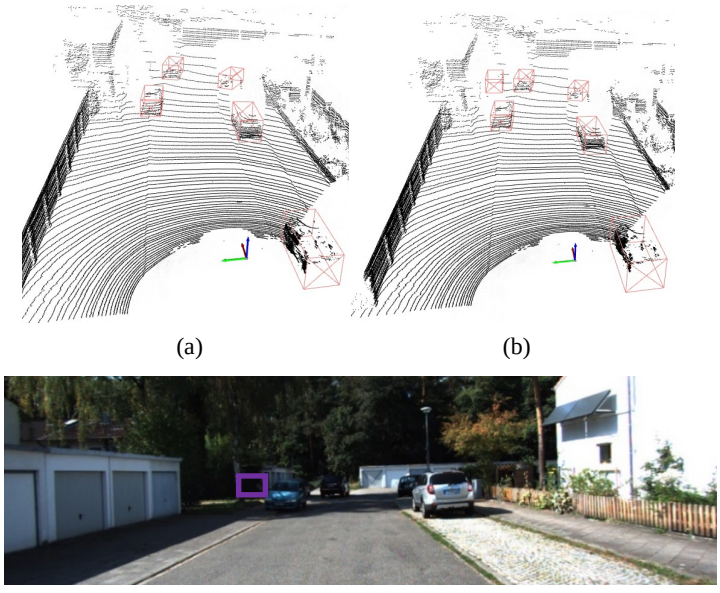


Figure 8. Comparison of detection results between this algorithm and the original PointPillars algorithm
图 8. 本文算法与原 PointPillars 算法检测结果对比

5. 结论

提出一种基于经典 PointPillars 三维激光雷达目标检测改进算法，解决了小尺度目标检测效果差的难题。通过添加当前点的反射率与当前支柱所有点反射率均值的偏差特征，增强点的表征能力，通过引入空间注意力机制，提高算法对伪图像的特征提取能力，提高算法的目标检测性能。在 KITTI 数据集上的实验结果表明：改进算法相较于原始算法在三种模式下的检测精度都有所提高，对于目标朝向角度误差，在小目标骑行者类别上精度提高效果尤其显著；在保证精度的同时，本文算法也能够有效降低目标误检以及目标被部分遮挡造成的漏检概率。

参考文献

[1] Alaba, S.Y. and Ball, J.E. (2022) A Survey on Deep-Learning-Based LiDAR 3D Object Detection for Autonomous Driving. *Sensors*, **22**, Article 9577. https://doi.org/10.3390/s22249577

[2] Qi, C.R., Su, H., Mo, K., *et al*. (2017) PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation. Computer Vision and Pattern Recognition. https://arxiv.org/abs/1612.00593

[3] Zhou, Y. and Tuzel, O. (2017) VoxelNet: End-to-End Learning for Point Cloud Based 3D Object Detection. 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-23 June 2018, 4490-4499. https://doi.org/10.1109/CVPR.2018.00472

[4] Yan, Y., Mao, Y. and Li, B. (2018) Second: Sparsely Embedded Convolutional Detection. *Sensors*, **18**, Article 3337. https://doi.org/10.3390/s18103337

[5] Lang, A.H., Vora, S., Caesar, H., *et al*. (2018) PointPillars: Fast Encoders for Object Detection from Point Clouds. 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Long Beach, 15-20 June 2019, 12689-12697. https://doi.org/10.1109/CVPR.2019.01298

[6] 陈德江, 余文俊, 高永彬. 基于改进 PointPillars 的激光雷达三维目标检测[J]. 激光与光电子学进展, 2023, 60(10): 447-453.

[7] Liu, W., Anguelov, D., Erhan, D., *et al*. (2016) SSD: Single Shot Multibox Detector. *Computer Vision-ECCV* 2016: 14*th European Conference*, Amsterdam, 11-14 October 2016, 21-37. https://doi.org/10.1007/978-3-319-46448-0_2

[8] Everingham, M.R., Eslami, S.M.A., Gool, L.J., *et al*. (2015) The Pascal Visual Object Classes Challenge. *International Journal of Computer Vision*, **111**, 98-136. https://doi.org/10.1007/s11263-014-0733-5

[9] 詹为钦, 倪蓉蓉, 杨彪. 基于注意力机制的 PointPillars+三维目标检测[J]. 江苏大学学报(自然科学版), 2020, 41(3): 268-273.

[10] Hu, J., Shen, L. and Sun, G. (2018) Squeeze-and-Excitation Networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-23 June 2018, 7132-7141. https://doi.org/10.1109/CVPR.2018.00745

[11] Wang, Q., Wu, B., Zhu, P., *et al*. (2020) ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, 13-19 June 2020, 11534-11542. https://doi.org/10.1109/CVPR42600.2020.01155

[12] Lin, T.Y., Dollár, P., Girshick, R., *et al*. (2017) Feature Pyramid Networks for Object Detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, 21-26 July 2017, 2117-2125. https://doi.org/10.1109/CVPR.2017.106

[13] Wang, W., Xie, E., Song, X., *et al*. (2019) Efficient and Accurate Arbitrary-Shaped Text Detection with Pixel Aggregation Network. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Seoul, 27 October 2019-2 November 2019, 8440-8449. https://doi.org/10.1109/ICCV.2019.00853

[14] Woo, S., Park, J., Lee, J.Y., *et al*. (2018) Cbam: Convolutional Block Attention Module. In: Ferrari, V., Hebert, M., Sminchisescu, C. and Weiss, Y., Eds., *Computer Vision-ECCV* 2018, *Lecture Notes in Computer Science*, Vol. 11211, Springer, Cham, 3-19. https://doi.org/10.1007/978-3-030-01234-2_1

[15] Tao, Z. and Su, J. (2022) Research on Object Detection Algorithm of 3D Point Cloud PointPillar Based on Attention Mechanism. 2022 *China Automation Congress* (*CAC*), Xiamen, 25-27 November 2022, 4382-4385. https://doi.org/10.1109/CAC57257.2022.10055052

[16] Geiger, A., Lenz, P. and Urtasun, R. (2012) Are We Ready for Autonomous Driving? The KITTI Vision Benchmark Suite. 2012 *IEEE Conference on Computer Vision and Pattern Recognition*, Providence, 16-21 June 2012, 3354-3361 https://doi.org/10.1109/CVPR.2012.6248074