基于改进 PointPillars 的激光点云三维目标检测

王家琦[1] 吴叶兰[1] 郝凤桐[1] 张峻景[1]
WANG Jiaqi WU Yelan HAO Fengtong ZHANG Junjing

摘 要 针对复杂场景下三维目标检测算法对小目标物体识别精度不高、容易出现错检漏检问题，提出一种基于改进 PointPillars 的三维目标检测算法，利用锥形点云获取物体的边界信息，抑制环境噪声造成的干扰；设计一种空间自注意力模块，捕获点云支柱间的全局上下文信息和空间信息，扩大点云特征感知范围，提升小目标物体识别在复杂场景下的鲁棒性；改进主干网络的下采样模块，采用 ConvNeXt v2 模块增强网络的特征提取能力。在 KITTI 数据集上的测试结果表明，相比 PointPillars 原始网络，改进算法在汽车、行人、骑行者类别上的平均检测精度分别提升了 3.73%、5.89%、5.7%，证明了所提出方法的有效性。

关键词 三维目标检测；PointPillars；注意力机制；点云支柱；ConvNeXt v2

doi：10.3969/j.issn.1672-9528.2024.02.013

0 引言

激光点云三维目标检测是自动驾驶领域的研究热点，能辅助车辆感知周围环境，对环境中的三维目标进行准确检测和定位[1]，被广泛应用于路径规划和安全避障等任务中。激光雷达点云数据对光照变化、天气条件等因素具有鲁棒性，可快速、精确地获取环境信息，然而其数据本身具有稀疏、不规则的特点，在处理诸如骑行者、行人等小目标时面临挑战。小目标往往因体积较小、易被遮挡，在点云数据中的特征表示受限，导致检测效果不理想，容易发生误检或漏检现象[2]。

激光点云三维目标检测通常分为基于原始点云和基于体素两种方法。PointPillars[3] 网络是一种基于体素的方法，该算法将点云立柱化后转换成伪图像，在保留三维特征的同时用二维卷积提取高维特征，具有较高的运行效率，但对尺寸较小或距离较远的目标识别效果不佳。为提升对小目标的识别效率，有学者对 PointPillars 进行了改进。詹为钦等人[4] 在 PointPillars 基础上引入空间和通道注意力模块，放大伪图像的重要特征，抑制无关特征，减少了对小目标物体的误检，但该方法没有考虑支柱内点云的相互关联。陈德江等人[5] 使用 Swin transformer 改进 PointPillars 的二维卷积降采样结构，增强网络的特征提取能力，但其仅针对目标朝向的精度有所提升，并未涉及三维检测框的准确率。

针对 PointPillars 现有改进方法的不足，本文提出一种基于改进 PointPillars 的激光点云三维目标检测算法，利用锥形点云获取物体的边界信息，抑制环境噪声干扰；设计一种空间自注意力模块，实现支柱间全局特征的充分聚合与提取；改进 PointPillars 主干网络下采样策略，增强对深层特征的提取能力，提高模型检测性能。

1 PointPillars 网络模型

PointPillars 采用体素化方法将原始点云转换成结构化伪图像，以实现目标检测，由点特征网络（pillar feature net，PFN）、主干网络（backbone）和检测头（detection head，DH）三部分组成。

算法基本原理是：首先将原始点云离散成均匀网格形成点柱组 P，构建密集张量 (D, P, N)，对数据稀疏的点柱用零填充，对数据过多的点柱进行随机抽样；再利用 PFN 将 D 维特征压缩至 C 维，得到一个 (C, P, N) 张量，通过池化操作产生特征张量 (C, P)，并映射回点柱形成伪图像 (C, H, W)；接着利用主干网络的下采样和上采样提取伪图像的高维特征，输出融合不同尺度的特征图；最后采用单次多边框检测头（single shot multibox detector，SSD）[6] 优化检测框，将二维检测结果映射回三维空间，回归 3D 检测框的中心、尺寸和朝向角，完成目标检测。

PointPillars 算法高效简洁，适用于实时系统，在自动驾驶中易于部署。但其存在小目标物体识别效果不佳的问题，主要原因是对点云特征的学习限制在柱体局部空间内，无法利用相邻柱体的全局上下文信息，因此影响小目标的检测精度；同时，主干网络使用传统的 2D CNN 提取特征，可能会忽略一些重要特征和上下文信息，导致特征提取能力较弱。

1. 北京工商大学计算机与人工智能学院 北京 100048

2 改进的 PointPillars 网络模型

为解决 PointPillars 复杂场景下小目标物体识别效果不佳的问题，本文进行三点改进：（1）提取锥形点云缩小目标搜索范围，抑制背景噪声干扰；（2）构建一种空间自注意力模块，利用自注意力和空间注意力强化全局上下文信息和空间位置信息的关联性，突出关键特征；（3）改进主干网络下采样模块，采用 ConvNeXt v2[7] 模块增强网络的特征提取能力。改进后的模型架构如图 1 所示。

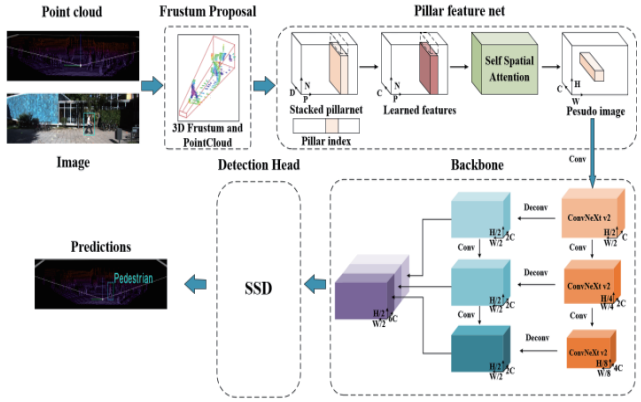

图 1 改进的 PointPillars 结构

2.1 锥形点云提取

锥形点云是结合激光点云和 RGB 图像获得的，首先对激光点云数据采用统计滤波 [8] 的方法去除离群点，再利用 Mask R-CNN 得到 RGB 图像中的 2D 目标检测框，通过相机投影矩阵配准 RGB 图像和激光点云数据，将 2D 检测框映射到激光点云的对应位置，形成锥形点云。为去除锥体内点云的背景噪声，采用高斯掩膜 [9] 方法，将高斯距离 L 作为一个附加特征增加到点云的特征向量中，从而提供更加准确的物体边界信息，有效抑制背景噪声干扰。高斯距离 L 的计算公式为：

$$L(\bar{x},\bar{y})=\exp\left(-\frac{(\bar{x}-\bar{x}_0)^2}{2w^2}-\frac{(\bar{y}-\bar{y}_0)^2}{2h^2}\right) \tag{1}$$

式中：$\bar{x}$ 和 $\bar{y}$ 为 RGB 图像上的点云投影坐标，$\bar{x}_0$ 和 $\bar{y}_0$ 是 2D 检测框的中心坐标，w 和 h 为检测框的宽度和高度。

2.2 空间自注意力机制

在 Pointpillars 原始支柱特征网络中，点云特征提取被限制在划分好的网格中，无法有效提取邻域点云的全局上下文信息，导致小目标物体特征的提取较为困难。本文设计一种空间自注意力模块，采用自注意力和空间注意力捕捉不同支柱间的全局上下文信息和空间信息，扩大点云特征感知范围，实现全局特征的充分聚合与提取，整体结构如图 2 所示。

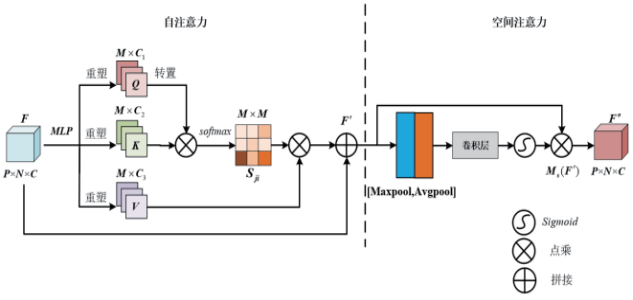

图 2 空间自注意力模块

2.2.1 自注意力模块

自注意力模块通过计算相关点云间的相似度自适应地生成和分配注意力权重，实现对点云支柱间全局上下文特征的有效提取与聚合，增强网络对关键点云特征的感知能力。自注意力模块使用多层感知机（multilayer perceptron，MLP）对点云特征 $\boldsymbol{F}$ 进行映射，重构后得到大小为 $R^{M\times C}$ 的特征 $\boldsymbol{Q}$、$\boldsymbol{K}$ 和 $\boldsymbol{V}$。对 $\boldsymbol{Q}^T$ 与 $\boldsymbol{K}$ 执行点积操作，利用 *softmax* 层计算全局自注意力权重矩阵 $\boldsymbol{S}_{ij}$。将 $\boldsymbol{S}_{ij}$ 与 $\boldsymbol{V}$ 相乘，并采用跳跃连接结构与特征 F 逐元素求和，得到全局自注意力输出特征 $\boldsymbol{F}'$。自注意力模块的计算公式为：

$$\boldsymbol{S}_{ij}=softmax\left(\frac{\exp(\boldsymbol{Q}_i\cdot\boldsymbol{K}_j)}{\sum_{i=1}^{N}\exp(\boldsymbol{Q}_i\cdot\boldsymbol{K}_j)}\right) \tag{2}$$

$$\boldsymbol{F}'=\alpha\sum_{i=1}^{N}(\boldsymbol{S}_{ij}\boldsymbol{V})+\boldsymbol{F}_j \tag{3}$$

式中：$\boldsymbol{Q}_i$ 为 $\boldsymbol{Q}$ 在位置 i 的特征，$\boldsymbol{K}_j$ 为 $\boldsymbol{K}$ 在位置 j 的特征，α 为可学习的尺度参数。

2.2.2 空间注意力模块

空间注意力模块是通过两个池化操作对空间特征进行强化，以获取特征间的空间关联性，为点云特征补充关键空间位置信息，增强模型对空间特征的理解能力。空间注意力模块以全局自注意力输出特征 $\boldsymbol{F}'$ 作为输入，利用最大池化和平均池化在通道维度对数据执行下采样操作，将得到的两个特征图拼接后，利用二维卷积提取空间特征，利用 sigmoid 激活函数对权重归一化得到空间注意力特征图 $\boldsymbol{M}_s(\boldsymbol{F}')$，与输出特征 $\boldsymbol{F}'$ 逐元素相乘得到最终输出特征 $\boldsymbol{F}''$，计算公式如下：

$$\boldsymbol{M}_s(\boldsymbol{F}')=\sigma(conv([AvgPool(\boldsymbol{F}');MaxPool(\boldsymbol{F}')])) \tag{4}$$

$$\boldsymbol{F}''=\boldsymbol{M}_s(\boldsymbol{F}')\otimes\boldsymbol{F}' \tag{5}$$

式中：σ 为 sigmoid 激活函数，*conv* 为 7×7 卷积操作，*AvgPool* 为平均池化操作，*MaxPool* 为最大池化操作，$\otimes$ 为逐元素乘法。

2.3 改进的主干网络

PointPillars 的主干网络是利用传统的 CNN 网络进行特征提取，特征提取不够充分，因此本文对主干网络的下采样

网络进行改进，采用 ConvNeXt v2 网络结构以获得更丰富的特征信息，网络结构如图 3 所示。

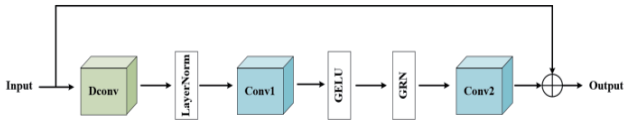

图 3 ConvNeXt v2 网络结构

ConvNeXt v2 借鉴 Swin Transformer 的设计思路，采用深度可分离卷积（DConv）来增强全局信息感知能力；引入全局响应归一化（GRN）来提高通道之间的对比度和选择性，从而增强网络的特征提取性能；采用反向瓶颈结构来减少信息丢失，提供更丰富的上下文信息。经过 ConvNeXt v2 改进的下采样特征提取网络结构如表 1 所示。

表 1 改进的下采样特征提取网络

	Layer	Repeat	Kernel size	Stride	Output channels
	Conv2d	1	3×3	1	64
Stage1	dwconv	3	7×7	3	64
	pwconv1		1×1	1	256
	pwconv2		1×1	1	64
	Conv2d	1	3×3	2	128
Stage2	dwconv	9	7×7	3	128
	pwconv1		1×1	1	512
	pwconv2		1×1	1	128
	Conv2d	1	3×3	4	256
Stage3	dwconv	3	7×7	3	256
	pwconv1		1×1	1	1024
	pwconv2		1×1	1	256

3 实验与分析

3.1 实验数据集

实验采用 KITTI 公开数据集，包含 7481 个训练样本、7518 个测试样本，分为汽车（car）、行人（pedestrian）和骑行者（cyclist）三种类别。遵循 PointPillars 算法将训练样本划分为 3712 帧的训练集和 3769 帧的验证集，其中训练集用于模型训练，验证集用于模型性能的评估。按照 KITTI 官方评价指标，用平均精度（average precision，AP）评价 3D 场景下的检测结果。

3.2 实验环境与设置

本文实验基于 OpenPCDet 目标检测框架实现，运行平台为 Inteli7-11700H 处理器，GPU 为 NVIDIA RTX 2060，6 GB 显存，系统环境为 Ubuntu 18.04。采用 Adam 优化器最小化损失函数，训练的最大迭代次数设置为 160，网络的批处理大小设为 2，初始学习率为 0.000 2，动量优化系数为 0.8，权重衰减率为 0.000 1。

3.3 对比实验

为评估本文改进算法在 KITTI 测试集上的性能，选择 F-PointNets、VoxelNet、SECOND、TANet、PointRCNN、PointPillars 算法进行对比，表 2 为在 KITTI 测试集 car、pedestrian 和 cyclist 类下本文算法与其他算法的平均精度对比。

表 2 不同算法的 AP 对比（%）

Algorithm	car			pedestrian			cyclist		
	Easy	Mod.	Hard	Easy	Mod.	Hard	Easy	Mod.	Hard
F-PointNets[10]	83.76	70.92	63.65	70.00	61.32	53.59	77.15	56.49	53.37
VoxelNet[11]	85.1	72.54	70.38	63.65	59.36	54.71	79.36	60.39	53.3
SECOND[12]	85.78	75.79	74.39	51.84	45.86	39.48	82.64	65.1	58.34
TANet[13]	85.34	74.92	72.48	66.64	59.29	54.06	86.88	66.96	63.24
PointRCNN[14]	86.18	75.94	75.27	58.53	51.2	47.44	86.15	67.04	61.5
PointPillars	86.1	76.74	73.78	63.57	56.52	51.17	85.96	66.19	59.29
本文算法	**89.26**	**79.61**	**78.94**	**68.93**	**63.56**	**56.44**	**88.63**	**73.28**	**66.65**

由表 2 数据可知，本文算法在 KITTI 数据集下的 AP 均优于其他主流 3D 目标检测算法，尤其对 pedestrian 和 cyclist 小目标物体的提升更为显著，这是由于本文算法在 PointPillars 的基础上采用了锥形点云提取、空间自注意力模块以及 ConvNeXt v2 模块，有效提升了模型对小目标的检测准确率。

3.4 可视化分析

为了直观表现目标检测效果，图 4 给出了本文算法与 PointPillars 在 2 个不同场景下的可视化效果。每个场景分为三层，第一层为 PointPillars 算法的检测结果，第二层为本文算法的检测结果，第三层为本文算法所得预测框在 RGB 图像上的投影。图中绿色检测框表示汽车（car）的真实框，蓝色检测框表示行人（pedestrian）真实框，黄色检测框表示骑行者（cyclist）真实框，红色检测框为预测框，红色圆圈中对不存在的目标进行了预测，代表错检，黄色圆圈中没有检测到存在的目标，代表漏检。可以看出，PointPillars 在场景 1 和场景 2 中出现了大量错检、漏检，检测框与真实框重合率较低，本文算法在场景 1 仅存在较少的错检，但通过改进增强了复杂场景下小目标物体的检测能力，有效减少了对小目标物体的错检、漏检情况，提升了整体网络的目标检测性能。

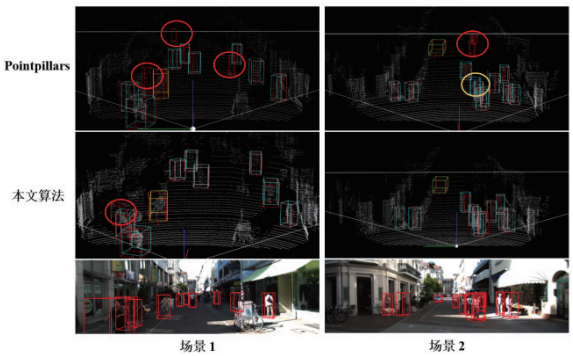

图 4 可视化效果对比

3.5 消融实验

为验证本文所提改进点对模型性能的影响程度，设计了消融实验。表 3 给出了 KITTI 验证集中 3D 场景下消融实验的 3D 平均检测精度，其中本文算法是融合了锥形点云、空间注意力和 ConvNeXt v2 三个模块的改进算法。实验结果表明，单独增加三种改进策略后检测精度均有相应提高，表明三个改进策略均是有效的；与基准网络相比，融合三种改进点的网络模型在 car、pedestrian 和 cyclist 类别上分别提升了 3.37%、5.89%、5.7%，对 pedestrian 和 cyclist 这两类小目标的检测精度有显著提升。

表 3 消融实验的 3D 平均检测精度（%）

Algorithm	car	pedestrian	cyclist
PointPillars	78.87	57.08	70.48
PointPillars + 锥形点云	81.15	60.81	74.04
PointPillars + 空间自注意力	80.52	61.97	72.61
PointPillars + 改进的主干网络	81.43	59.27	73.46
本文算法	**82.60**	**62.97**	**76.18**

4 结论

为解决小目标物体识别精度不高的问题，本文提出一种基于改进 PointPillars 的三维目标检测算法。利用锥形点云获取更为准确的物体边界信息，抑制背景噪声干扰；在点云编码部分加入空间自注意力，捕获点云支柱间的全局上下文信息和重要空间特征，扩大点云特征的感知范围，增强对小目标物体识别的鲁棒性；在主干网络的下采样部分加入 ConvNeXt v2，增强对于高维特征的提取能力，促进多尺度特征聚合，提升 2D CNN 对小目标物体特征的提取能力。实验结果表明，本文算法在检测精度方面优于主流算法，提升了小目标物体的检测准确率，在未来的研究中将考虑采用轻量级网络降低网络的复杂度，进一步提升算法的实时性。

参考文献：

[1] 郭毅锋，吴帝浩，魏青民 . 基于深度学习的点云三维目标检测方法综述 [J]. 计算机应用研究 ,2023,40(1):20-27.

[2] 周燕，蒲磊，林良熙，等 . 激光点云的三维目标检测研究进展 [J]. 计算机科学与探索 ,2022,16(12):2695-2717.

[3]LANG A H,VORA S,CAESAR H,et al. PointPillars: fast encoders for object detection from point clouds[C]// Proceedings of the 2019 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway:IEEE, 2019:12697-12705.

[4] 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 ,2020,41(3):268-273.

[5] 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.

[6]LIU W,ANGUELOV D,ERHAN D,et al. SSD: single shot MultiBox detector[M]//LEIBE B, MATAS J, SEBE N, et al. Computer vision-ECCV 2016. Lecture notes in computer science,Springer, 2016,9905: 21-37.

[7]WOO S,DEBNATH S,HU RH,et al. ConvNeXt V2: Co-designing and scaling ConvNets with masked autoencoders[C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway:IEEE, 2023:16133-16142.

[8]ZHOU L, SUN G X, LI Y,et al. Point cloud denoising review: from classical to deep learning-based approaches[J]. Gpaphical models, 2022, 121(5):101140.

[9]PAIGWAR A,SIERRA-GONZALEZ D,ERKENT Ö,et al. Frustum-PointPillars: a multi-stage approach for 3D object detection using RGB camera and LiDAR[C]// 2021 IEEE/CVF International Conference on Computer Vision Workshops （ICCVW）. Piscataway:IEEE, 2021: 2926-2933.

[10]QI C R,LIU W,WU C,et al. Frustum pointNets for 3D object detection from RGB-D data[C]// Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway:IEEE, 2018: 918-927.

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

[12]YAN Y,MAO Y,LI B. SECOND: sparsely embedded convolutional detection[J]. Sensors,2018,18(10): 3337-3353.

[13]LIU Z,ZHAO X,HUANG T,et al. Tanet: robust 3D object detection from point clouds with triple attention[C]// Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto: AAAI, 2020: 11677-11684.

[14]SHI S,WANG X,LI H. Pointrcnn：3D object proposal generation and detection from point cloud[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway:IEEE, 2019: 770-779.

【作者简介】

王家琦（1999—），男，北京人，硕士研究生，研究方向：3D 目标检测。

吴叶兰（1970—），女，湖北荆州人，硕士，副教授，研究方向：智能信息处理。

郝凤桐（1997—），男，北京人，硕士研究生，研究方向：激光 SLAM。

张峻景（1998—），男，北京人，硕士研究生，研究方向：移动机器人控制技术。

（收稿日期：2024-01-16）