引用格式：LIU Weili, ZHU Deli, LUO Huahao, et al. 3D Object Detection with Fusion Point Attention Mechanism in LiDAR Point Cloud[J]. Acta Photonica Sinica, 2023, 52(9):0912002
刘威莉,朱德利,骆华昊,等.LiDAR点云中融合点注意力机制的三维目标检测[J].光子学报,2023,52(9):0912002

LiDAR点云中融合点注意力机制的三维目标检测

刘威莉[1,2],朱德利[1,2],骆华昊[1,2],李益[3]
(1 重庆师范大学 计算机与信息科学学院,重庆 401331)
(2 重庆市数字农业服务工程技术研究中心,重庆 401331)
(3 重庆市畜牧科学院信息中心,重庆 401331)

摘　要:针对Pillar编码点云的三维目标检测算法中存在一定细粒度信息的丢失以及对点云特征提取能力不足等问题,基于PointPillars提出一种融合逐点空间注意力机制和跨阶段局部网络的三维目标检测算法。首先在支柱特征网络层中融入逐点空间注意力机制,增强网络对局部几何信息的提取并保留深层次信息,使得到的关键特征更适合检测任务;其次将对点云伪图像进行高维特征提取的降采样模块中的普通卷积替换为跨阶段局部网络,进一步提升网络的学习能力;最后算法在高速公路的应用场景下,以KITTI数据集中car类作为检测目标,与基准网络相比,在简单、中等和困难三种情况下的3D检测精度分别提高了2.23%、2.25%和2.30%。实验结果表明,所提算法在检测性能上有明显提升,同时检测速度达到实时检测水平,对自动驾驶技术的优化和完善具有一定的积极意义。

关键词:三维目标检测;点云;注意力机制;PointPillars;跨阶段局部网络
中图分类号:TP391　　**文献标识码**:A　　**doi**:10.3788/gzxb20235209.0912002

0 引言

随着深度学习的快速发展,目标检测已经在二维计算机视觉任务中取得了显著的成就,然而实际场景中存在的光照变化、天气条件、深度缺失等问题,仅仅依靠二维视觉感知无法解决[1]。由激光雷达获取的三维数据不依赖自然光等条件,弥补了二维视觉领域存在的一些缺陷。三维目标检测作为三维场景感知中一个重要领域被广泛研究[2],自动驾驶领域的三维目标检测是实现自动驾驶路径规划和安全避障的重要研究内容[3]。随着激光雷达技术的不断进步,以激光点云作为输入的深度学习检测器[4]也逐渐成熟,然而点云数据通常是稀疏和无序的,如何从不规则的点云中提取关键特征成为了三维目标检测任务中的一个关键性挑战[5]。

输入的LiDAR数据以点云的形式来表示,但由于点云非结构化和非固定大小的特征,使其不能直接被3D目标检测器处理,必须通过某种表达形式将其编码为更紧凑的结构,目前主要分为两种类型的表达形式:基于点(point-based)的方法和基于体素(voxel-based)的方法。QI C R等[6]首先提出pointnet网络直接对无序点云进行特征的学习,随后使用Max-pooling聚合为全局特征。沿用pointnet的思想,为了提取更有鉴别性的高维特征,pointnet++[7]采用球查询半径内的领域点,随后每个局部点通过集合抽象(Set Abstraction,SA)层进行多层次的特征提取。SHI S等[8]提出的PointRCNN算法是通过PointNet++进行特征的提取,基于提取到的特征进行前景和背景的分割,在每个前景点上进行3D框的预测,然后在提取到的目标的基础上进一步细化。3DSSD[9]网络则是利用SA层对输入的点云进行降采样后,利用Backbone等网络提取关键

基金项目:重庆市教育委员会科学技术研究项目(No. KJQN201800536),重庆市高校创新研究群体项目(No. CXQT20015)
第一作者:刘威莉,wl461007@163.com
通讯作者:朱德利,463453339@qq.com
收稿日期:2023-04-04;**录用日期**:2023-05-17

点的特征，并用其中的一部分来进行投票，投票结果进一步用SA层进行特征提取，最后利用该特征对检测框的种类和位置进行预测。此类方法在处理点云的过程中可以充分利用点云的几何特征，以此来获得更好的检测性能，但在特征提取过程中消耗了大量的时间和计算资源。在基于体素的方法中，考虑到点云的稀疏性（即大约90%的体素都没有点）。VoxelNet[10]算法将点云划分为等间距的规则体素，然后使用体素特征编码（Voxel Feature Encoder，VFE）层将体素内点的特征量化统一，再使用3D卷积神经网络对体素进行特征提取，最终使用RPN网络生成检测框[11]，但该算法由于体素数量庞大，特征提取速度极慢。YAN Y等[12]提出的SECOND网络将点云转化为规则的体素并使用3D稀疏卷积进行提取特征，相比于VoxelNet网络加快了对点云特征的提取速度。LANG A H等[13]提出的PointPillars网络则是将点云立柱化后转化伪图像，在保留三维特征的同时进一步采用二维卷积提取高维特征，极大地加快了算法的运行速度。基于体素的方法具有更高的检测效率，但是检测精度低于基于原始点云的方法。如何在保证检测效率的基础上，提高基于体素方法的检测精度成为近年来的研究热点。

基于Pillar编码点云的三维目标检测算法中存在一定细粒度信息的丢失，而这些丢失的局部几何信息对于检测目标是非常关键的[14]。王忠全[15]在PointPillars的2D CNN主干网络上增加了4个改进后的有效通道注意力（Efficient Channel Attention，ECA）模块提高三维目标检测精度。詹为钦等[16]引入2种注意力机制，实现对伪图中特征信息的放大和抑制。上述研究都是针对点云伪图像高维特征提取模块进行的改进，未考虑在点云柱内的特征学习机制。基于此，本文基于PointPillars提出了一种融合逐点空间注意力机制[17]和跨阶段局部网络（Cross Stage Partial Network，CSPNet）[18]的三维目标检测算法，以有效提高网络的特征提取能力，保留深层次点云特征，提升网络检测目标的准确率。

1 PointPliiars网络模型

PointPillars是一种单阶段的3D点云目标检测算法，使用原始点云作为输入，通过在鸟瞰图上划分栅格，实现立柱形式的体素的划分，随后经过降维处理生成伪图像，利用二维卷积对特征进行提取，极大地提高了检测效率。针对KITTI中car类目标，PointPillars很好地做到了在检测性能和效率之间的平衡。算法的整体框架如图1。

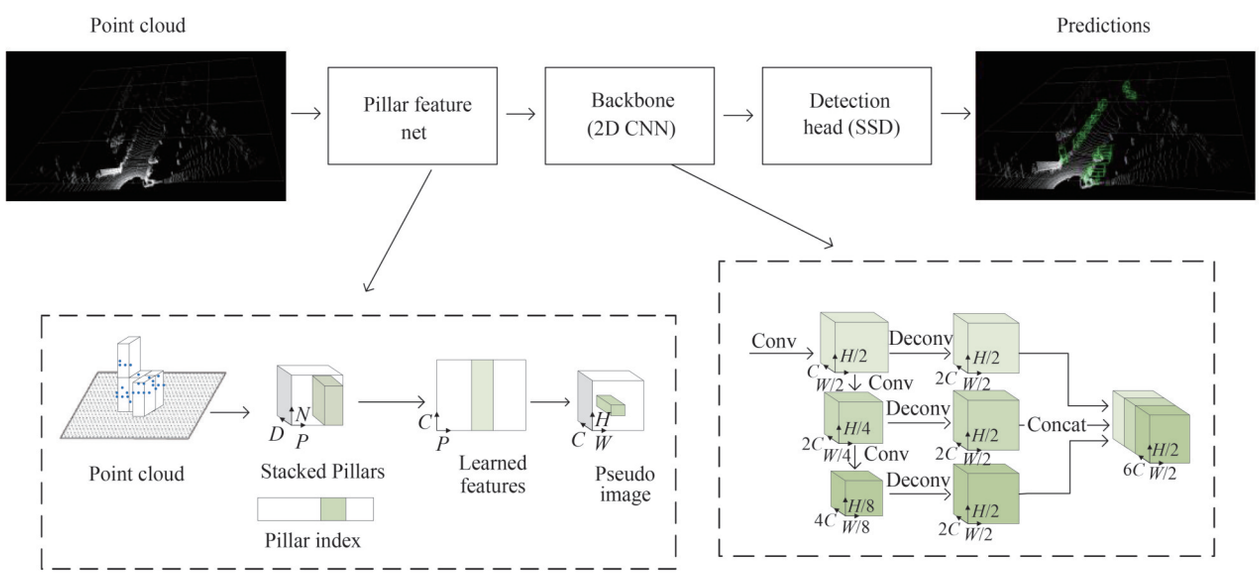


图1 PointPillars算法整体框架

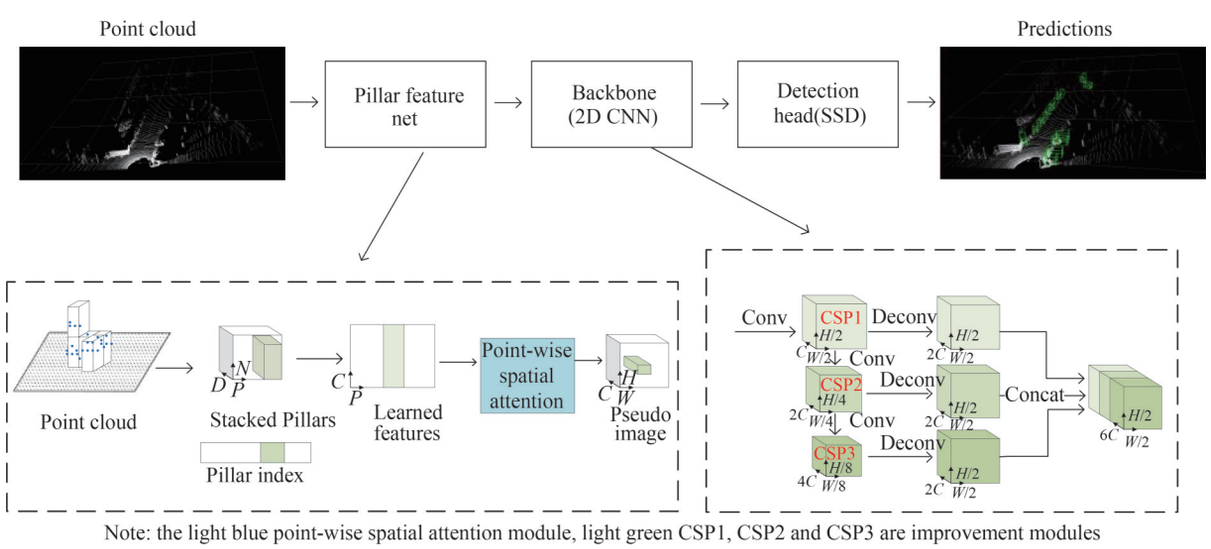
Fig. 1 Overall framework of PointPillars algorithm

该算法主要由三个核心模块组成，1）支柱特征网络（Pillar Feature Net，PFN）层：负责将三维点云转换为稀疏伪图像；2）2D BackBone主干网络层：通过二维卷积对点云伪图像进行特征提取；3）Detection Head（SSD）检测头层：预测目标类别以及三维边界框等信息。

算法具体流程为：首先在支柱特征网络层中将输入的点云数据划分为Pillars，每个Pillar中的点云由包含坐标、反射强度、几何中心和相对位置等信息的10维向量表示，之后用一个简化版的PointNet从D维原始数据中学习得到C维特征，得到一个(C, P, N)的张量，再使用maxpool操作提取每个pillar中最能代表该pillar的点，得到(C, P)维度数据，之后利用scatter算子，根据对应位置关系将数据映射到相应位置，实现三维数据向二维伪图像的转换；然后2D BackBone主干网络层对支柱特征网络生成的点云伪图像进行高维特征提取，包括两个分支，一支为自上而下的渐进式下采样分支，另一支为上采样分支，通过反卷积将多尺度的特征图上采样到统一大小，并进行拼接，得到最终的融合特征图；最后在检测头层采用了类似SSD的检测头来实现3D目标检测，回归3D框的中心、尺寸和朝向角。具体为使用2D联合交叉(IoU)将先验框与地面真值相匹配，在2D网络中进行目标检测，并通过回归的方式得到Z轴坐标和高度。

2 改进的PointPillars算法

原始PointPillars网络中Pillar编码的点云数据存在一定程度的信息丢失，没有考虑到点云空间分布的局部几何信息，对目标检测精度不高。本文在支柱特征网络层中融入逐点空间注意力机制，抑制点云支柱中的噪声，放大重要特征信息，提高对点云的特征提取能力；另外在降采样模块中使用可以分割梯度流的CSPNet替换原降采样中普通卷积块，使梯度流在不同的网络路径中传播，在减少计算量的同时提升网络的检测性能。改进之后的整体网络架构如图2。



Note: the light blue point-wise spatial attention module, light green CSP1, CSP2 and CSP3 are improvement modules

图2 改进后的PointPillars算法整体框架

Fig. 2 The overall framework of the improved PointPillars algorithm

2.1 融入逐点空间注意力机制

注意力机制是通过计算输入数据的权重，提高某个重要因素对结果的影响力，抑制不重要因素的影响[19]。逐点空间注意力遵循自注意力的基本结构，从局部点图的点空间捕捉更多形状相关特征和长距离相关性。此外，该机制还应用跳跃连接来加强输入和输出之间的关系，提高对特征的学习，加强高层的语义信息。逐点空间注意力模块整体结构如图3，使用两个多层感知机(Multilayer Perceptron，MLP)将局部特征F转换为特征X和Y，其中X, $Y\in\mathbb{R}^{C_1}$，与文献[20]中不同的是，用X和Y的转置来进行计算不同点云之间的关系，不需要对矩阵进行重构，保持了原来的空间分布。最后利用softmax对关系图进行归一化，达到大小为$N\times N$的空间注意图S(N表示点的个数)，表示为

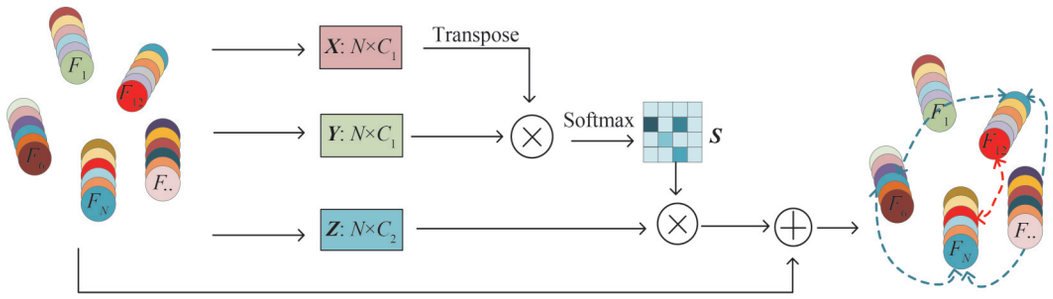
$$S_{ij}=\text{soft max}\left(\frac{\exp(X_i\cdot Y_j)}{\sum_{i=1}^{N}\exp(X_i\cdot Y_j)}\right) \tag{1}$$

式中，i和j分别表示点在X和Y中的位置，S_{ij}是i^{th}点对j^{th}点的影响，$\cdot$表示矩阵乘法。当两个点的特征具有相似的语义信息时，他们具有很强的相关性。同时局部特征F转化为新特征$Z\in\mathbb{R}^{C_2}$，通过MLP层，然后是S

和Z之间的矩阵乘法，并与特征F求和得到输出$F_{\text{final}}\in\mathbb{R}^{N\times C'}$，表示为

$$F_{\text{final}}=S\times Z+F \tag{2}$$

得到的F_{final}具有长距离相关性，并通过逐点空间注意力图S有选择性地聚合上下文，捕获全局相关性。



Note:the introduced point-wise spatial attention module, feature mapping is represented by the shape of its tensor, $N\times C_1$, where N denotes the number of points and C_1 denotes the feature size. ⊗ is the matrix multiplication, and ⊕is the point-wise sum. The pink, green and blue boxes denote the MLP layers. The features of the point cloud are fed into this module to learn the long-range correlation between points

图3 逐点空间注意力模块结构

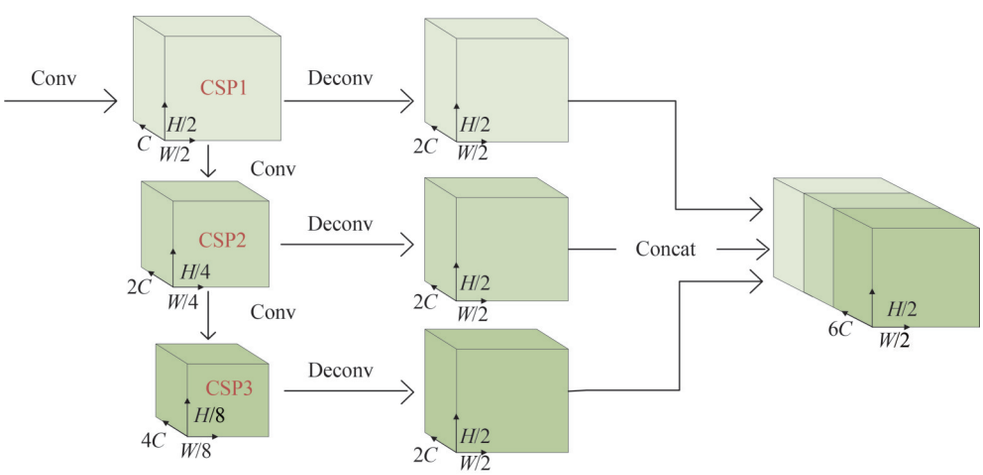
Fig. 3 Structure of point-wise spatial attention module

在PointPillars的支柱特征网络层中，由D维原始数据学习得到C维特征的感受野受限，过滤特征的各个单元不能利用其局部区域之外的上下文信息。为了解决这一问题，本文将重点放在全局空间关系上，捕获深层次信息。基于此融入一个逐点空间注意力模块，它通过在点集中建立特征之间的关联来捕捉全局依赖性。将Pillar编码后通过简化版PointNet提取特征的点云输入至逐点空间注意力模块中，增强了对点云特征提取的能力，同时可以有效避免冗余点云或噪声点对特征的影响，加强了对点云覆盖较少的特征描述，在一定程度上解决了基于Pillar编码点云的信息丢失问题，提高了三维目标检测的精度。

2.2 CSPNet改进的伪图像下采样

CSPNet是将上一层得到的特征图分割成两部分，然后通过跨阶段分层结构进行合并来实现的，主要概念是通过分割梯度流，使梯度流在不同的网络路径中传播[21]。这样的策略会大量减少计算量，加快模型的推理速度，有效增强网络的学习能力，提高模型检测精度。

PointPillars中点云经过体素特征编码后通过Scatter算子生成伪图像，随后对多尺度伪图像提取特征。针对网络对伪图像特征提取能力不足的问题，选择CSPNet作为对点云伪图像进行高维特征提取的下采样特征提取网络，进行特征融合以有效增强卷积神经网络的学习能力，提高模型的准确率。整体二维主干网络如图4，CSP1、CSP2、CSP3均为CSPNet网络结构。CSPNet和BottleNeck网络结构如图5，CSPNet由多



Note: light green CSP1, CSP2 and CSP3 are CSPNet improved pseudo-image downsampling modules

图4 二维主干网络结构

Fig. 4 2D backbone network structure

个1×1的卷积组成，首先通过将伪图像特征分成两部分，一部分用普通卷积提取特征信息，另一部分通过1×1的卷积和BottleNeck层。具体做法是先进行1×1卷积将通道数减小一半，再通过3×3卷积将通道数加倍，保证其输入与输出的通道数不发生改变，然后使用add进行特征融合，使得融合后的特征数不变。最后将两部分的特征图进行Concat拼接操作，使得融合前后的通道数不变，使用Silu激活函数，通道数等数据如表1。实验结果表明，CSPNet有效增强了网络的学习能力，并且提升了网络检测目标的准确率，此外，CSPNet网络的不同特征层的拼接重用还提高了模型对目标的泛化性。

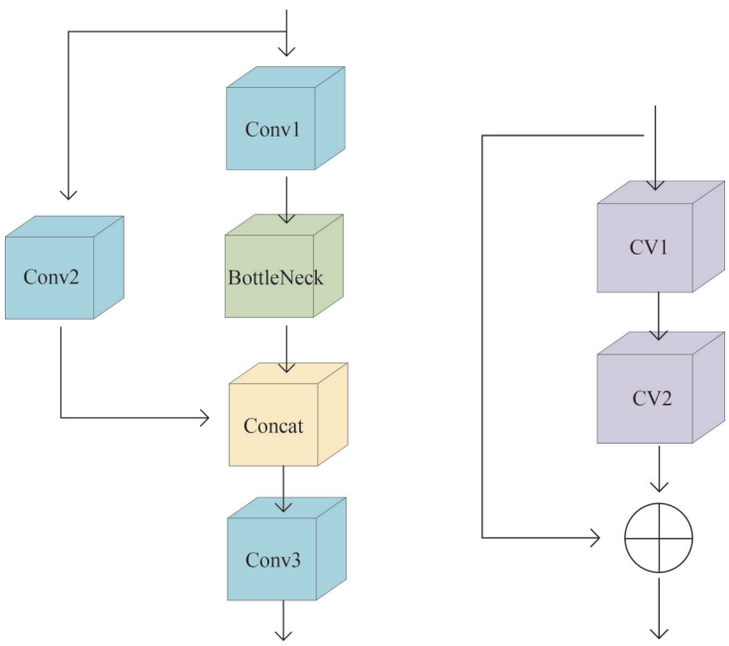


图5　CSPNet、BottleNeck网络结构

Fig. 5　CSPNet, BottleNeck network structure

表1　本文CSPNet网络结构

Table 1　CSPNet network structure of this paper

	Layer	Repeat	Kernel size	Stride	Output channels
	Conv2d	1	3×3	2	64
Stage1	Conv1	3	1×1	1	32
	Conv2		1×1		32
	BottleNeck		1×1	1	32
			3×3		32
	Conv3		1×1	1	64
Stage2	Conv2d	1	3×3	2	128
	Conv1	5	1×1	1	64
	Conv2		1×1		64
	BottleNeck		1×1	1	64
			3×3		64
	Conv3		1×1	1	128
Stage3	Conv2d	1	3×3	2	256
	Conv1	5	1×1	1	128
	Conv2		1×1		128
	BottleNeck		1×1	1	128
			3×3		128
	Conv3		1×1	1	256

3 实验结果与分析

3.1 实验数据

在KITTI数据集[22]上对算法进行验证，KITTI数据集是目前自动驾驶领域最重要的数据集之一[23]，数据集内包含市区、乡村和高速公路等真实驾驶场景的数据图像。KITTI共有7 481套训练样本和7 518个测试样本，其中每个样本场景中约包含16 384个点。采用Chen[24]等对训练数据划分的方式，将训练样本又分为训练集3 712套，验证集3 769套。主要类别有车辆、行人和自行车三类。因为车辆类别数量较多且是样本中最大的类别，并且拟应用的场景为高速公路，所以本文只在车辆类别上进行训练和测试。KITTI结果统计时，根据检测目标被遮挡情况、与当前视点距离以及框的高度等参量，将结果分为简单、中等和困难三种场景进行统计，具体数据划分如表2。

表2 三种场景下的数据划分

Table 2 Data division in three scenarios

	Easy	Moderate	Hard
Min height of bounding box	40 pixels	25 pixels	25 pixels
Max blocking level	Fully visible	Partially obscured	Hard to see
Maximum cut-off	15%	30%	50%

按照KITTI官方评价指标，以平均精度(Average Precision，AP)评价3D和BEV场景下的检测结果，作为检测性能的评估指标。采用的交并比(Intersection Over Union，IOU)阈值为0.7，使用40个召回位置计算平均精度，计算表达式为

$$\mathrm{AP}=\frac{1}{40}\sum_{R\in\left[0,\frac{1}{40},\frac{2}{40},\cdots,1\right]}P(R) \tag{3}$$

3.2 实验参数与环境

改进算法基于OpenPCDet框架实现，训练时超参数设置如下：采用Adam优化器训练160个epoch，batch size为4，学习率为0.003。实验使用的点云范围沿x、y、z轴，分别是$W=[0\ \mathrm{m},-39.68\ \mathrm{m}]$，$H=[-3\ \mathrm{m},69.12\ \mathrm{m}]$，$D=[39.68\ \mathrm{m},1\ \mathrm{m}]$。体素尺度设置为$v_{\mathrm{w}}=0.16$，$v_{\mathrm{h}}=0.16$，$v_{\mathrm{d}}=4$。将MAX_POINT_PER_VOXEL设置为32，作为每个体素中的最大点数，同时MAX_NUMBER_OF_VOXELLS训练时设置成16 000，测试时设置为40 000，作为最小批量中的最大非空体素数。具体实验环境配置如表3。

表3 实验环境配置

Table 3 Experimental environment configuration

Experimental environment	Configuration
Operating system	Ubantu 16.04
Processor	Intel Xeon Silver 411
Memory	64 GB
Video card	NVIDIA TITAN V
Deep learning framework	Pytorch 1.5
Development language	Python 3.7

3.3 实验结果分析

3.3.1 对比实验

为了评估改进的网络模型在KITTI测试集上的精度性能，选择F-PointNet、VoxelNet、SECOND、PointPillars、SegVoxelNet、TANet、PointRCNN、Part-A^2算法进行对比，表4为在KITTI测试集Car类下，本文算法与其他算法的平均精度对比。

本文算法在简单、中等、困难情况下的3D平均检测精度为88.52%、79.02%、76.22%，BEV平均检测精度为92.63%、88.53%、87.16%，均达到了最优。改进后的算法有较优的检测性能，尤其是在困难情况下的弱感知目标样本中有着较高的平均检测精度并取得了最为显著的精度提升幅度。同时，表5给出了本文算

表4　不同算法的AP对比(%)

Table 4　Comparison of AP for different methods(%)

Method/R_{40}	Car-3D(IoU=0.7)			Car-BEV(IoU=0.7)		
	Easy	Moderate	Hard	Easy	Moderate	Hard
F-PointNets[25]	82.19	69.79	60..59	91.17	84.67	74.77
VoxelNet[10]	87.93	75.37	73.21	89.35	79.26	77.39
SECOND[12]	83.34	72.55	65.82	89.39	83.77	78.59
PointPillars[13]	86.29	76.77	73.92	91.89	88.07	87.02
TANet[26]	84.39	75.94	68.82	75.70	59.44	52.53
SegVoxelNet[27]	86.04	76.12	70.76	91.62	86.37	83.04
PointRCNN[8]	86.96	75.64	70.70	92.13	87.39	82.72
Part-A^2[28]	87.81	78.49	73.51	91.70	87.79	84.61
Ours	**88.52**	**79.02**	**76.22**	**92.63**	**88.53**	**87.16**

法和现有的其他几种表现优异的三维点云目标检测算法的推理速度，对比可知，本文算法在有效提升基准网络检测精度的同时也保证了高效的推理速度，KITTI数据采集设备的64线激光雷达工作频率是10 Hz，即1 s处理获取10帧点云数据，本文提出的算法每秒处理的点云数据大于10帧，推理速度为0.037 2 frame·s^{-1}，满足实时性检测的要求[29]。

表5　不同算法的推理速度对比

Table 5　Inference speed comparison among different methods

Method	Reasoning speed/(frame·s^{-1})
F-PointNets[25]	0.169
VoxelNet[10]	0.033
SECOND[12]	0.380
3DSSD[9]	0.04
TANet[26]	0.035
SegVoxelNet[27]	0.04
PointRCNN[8]	0.067
Part-A^2[28]	0.08
SA-SSD[30]	0.04
Ours	**0.037 2**

3.3.2　消融实验

为了验证所提出的两个模块对网络性能的影响程度，通过消融实验来进行说明。以下所有模型都在KITTI数据集上进行训练并测试，表6和表7分别给出了KITTI验证集中消融实验的3D和BEV场景下的检测性能数据。消融实验是以单独模块，两个模块结合来展示改进点的贡献，PPPA为融合了逐点空间注意力模块，PPCSP为CSPNet改进的伪图像下采样模块。PPCSP+PPPA为融合了逐点空间注意力机制和CSPNet的三维目标检测算法。

表6　在KITTI测试集中消融实验的3D检测平均精度(%)

Table 6　Average precision of 3D detection for ablation experiments in the KITTI test set(%)

Method	Car-3D (IoU=0.7)		
	Easy	Moderate	Hard
PointPillars	86.29	76.77	73.92
PPPA	87.72	78.23	75.13
PPCSP	87.83	78.30	75.60
PPCSP+PPPA	88.52	79.02	76.22

表7 在KITTI测试集中消融实验的BEV场景下检测平均精度(%)

Table 7 Average precision of detection in the BEV scenario of the KITTI test focused ablation experiment(%)

Method	Car-BEV(IoU=0.7)		
	Easy	Moderate	Hard
PointPillars	91.89	88.07	87.02
PPPA	92.56	88.60	87.24
PPCSP	92.13	88.02	86.68
PPCSP+PPPA	92.63	88.53	87.16

消融实验结果表明:在PointPillars网络中加入逐点空间注意力模块,可以捕获全局相关性,有效抑制点云支柱中的噪声,放大重要特征信息,增强对点云的特征提取能力,提高检测精度;在网络中使用CSPNet对伪图像下采样进行改进,使梯度流在不同的路径中传播,增强了算法的特征提取能力;PPCSP+PPPA为加入两个模块后的检测结果,在简单、中等和困难级别下的3D检测精度分别为88.52%、79.02%和76.22%,与基准网络相比分别提升了2.23%,2.25%和2.30%,BEV场景下的平均检测精度为92.63%,88.53%,87.16%,均高于基准网络。结果表明改进后模型的检测性能得到了显著的提升,说明所使用的两种改进方法均具有有效性。

3.3.3 结果可视化

为了更加直观地验证改进网络对于车辆检测的有效性,图6给出了PointPillars和改进后网络在KITTI测试集中目标检测的可视化对比结果。

图6 PointPillars与本文算法的可视化结果对比

Fig.6 Comparison of the visualization results of PointPillars and the algorithm in this paper

图6分别给出了四个不同场景下的PointPillars目标样本(车辆)与本文算法目标样本的可视化结果对比图,其中用红色线圈标识目标车辆被误检的情况,用黄色线圈标识目标车辆被漏检的情况,从中可以看出改进算法在点云图中误检率和错检率更低。这是由于融合了逐点空间注意力机制和CSPNet网络的三维目标检测算法更加关注全局特征,减少了点云编码过程中造成的信息丢失,并改善了降采样模块特征提取能力不足的问题。因此改进后的PointPillars比改进前效果更好,在一定程度上消除了误检漏检的情况,提升了网络检测的性能。

4 结论

本文基于PointPillars提出一种融合逐点空间注意力机制和CSPNet网络的三维目标检测算法来实现对车辆的检测。首先在简化版PointNet提取点云特征后，融入逐点空间注意力机制进行有选择地聚合上下文信息，捕获全局相关性，进一步提高点云特征学习的能力。其次将点云伪图像进行高维特征提取的降采样模块中普通卷积替换为CSPNet网络，有效提高了卷积神经网络的特征提取能力，保留了深层次点云特征。

在高速公路的应用场景下，以KITTI中car类作为检测对象，在简单、中等和困难级别下的3D检测精度分别为88.52%、79.02%和76.22%，与基准网络相比分别提升了2.23%，2.25%和2.30%。另外，采用消融实验分析验证了所提模块的改进能够有效提高三维目标检测的性能。最后将算法与VoxelNet、SECOND、PointRCNN等经典三维目标检测算法的性能进行了对比，本文所提算法性能较优，同时检测速度也达到了实时检测水平，对自动驾驶技术的进一步优化和完善具有一定的积极意义。

参考文献

[1] LIN T Y, GOYAL P, GIRSHICK R, et al. Focal loss for dense object detection[C]. Proceedings of the IEEE International Conference on Computer Vision, 2017: 2980-2988.

[2] QIN Jianwei, WANG Chuanxu, FU Xiaoshan. AF-Center: multi-modal 3D object detection method with adaptive voxel-painting fusion and Gaussian center sample assignment[J]. Application Research of Computers, 2023, 40(2): 634-640.
秦建伟，王传旭，付小珊. AF-Center：基于自适应体素绘画融合和高斯中心样本分配的多模态三维目标检测[J].计算机应用研究，2023，40(2)：634-640.

[3] WANG Qinglin, LI Hui, XIE Lizhi, et al. Research on improving vehicle target detection algorithm based on lidar point cloud[J]. Electronic Measurement Technology, 2023, 46(1): 120-126.
王庆林，李辉，谢礼志，等. 基于激光雷达点云的车辆目标检测算法改进研究[J].电子测量技术，2023，46(1)：120-126.

[4] ZHAO L, HUJIE H L, YONGPENG A, et al. Deep learning based on semantic segmentation for three-dimensional object detection from point clouds[J]. Chinese Journal of Lasers, 2021, 48(17): 1710004.

[5] LI Qiao, LI Yaochen, ZHANG Yulong, et al. A Single-stage point cloud 3D object detection method using sparse 3D convolution[J]. Journal of Xi′an Jiaotong University, 2022, 56(9): 112-122.
李悄，李垚辰，张玉龙，等. 采用稀疏3D卷积的单阶段点云三维目标检测方法[J].西安交通大学学报，2022，56(9)：112-122.

[6] 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.

[7] QI C R, YI L, SU H, et al. PointNet++: deep hierarchical feature learning on point sets in a metric space[C]. Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017: 5100-5109.

[8] 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,2019: 770-779.

[9] YANG Z, SUN Y, LIU S, et al. 3dssd: point-based 3d single stage object detector[C].Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,2020: 11040-11048.

[10] 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.

[11] LI Ruilong, WU Chuan, ZHU Ming. 3D object detection in voxelized point cloud scene[J]. Chinese Journal of Liquid Crystals and Displays, 2022, 37(10): 1355-1363.
李瑞龙，吴川，朱明. 体素化点云场景下的三维目标检测[J].液晶与显示，2022，37(10)：1355-1363.

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

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

[14] ZHOU S, TIAN Z, CHU X, et al. FastPillars: a deployment-friendly Pillar-based 3D detector[J]. Arxiv Preprint Arxiv: 2302.02367, 2023.

[15] WANG Zhongquan. Research on object detection and path planning based on deep learning[D].Harbin: Harbin Institute of Technology, 2021.
王忠全. 基于深度学习的目标检测及路径规划研究[D]. 哈尔滨：哈尔滨工业大学，2021.

[16] ZHAN Weiqin, NI Rongrong, YANG Biao. An attention-based PointPillars+3D object detection[J]. Journal of Jiangsu University, 2020, 41(3): 268-273.
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.

[17] FENG M, ZHANG L, LIN X, et al. Point attention network for semantic segmentation of 3D point clouds[J]. Pattern

Recognition, 2020, 107: 107446.
[18] WANG C Y, LIAO H Y M, WU Y H, et al. CSPNet: A new backbone that can enhance learning capability of CNN[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020: 390-391.
[19] MNIH V, HEESS N, GRAVES A, et al. Recurrent models of visual attention[C].Proceedings of the 27th International Conference on Neural Information Processing Systems, 2014: 2204-2212.
[20] FU J, LIU J, TIAN H, et al. Dual attention network for scene segmentation[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019: 3146-3154.
[21] QIN Tong, LI Xiaoming. Longitudinal tear detection based on improved YOLOv4 for conveyor belt[J]. Computer Systems&Applications, 2023,32(3): 186-194.
秦彤, 李晓明. 改进YOLOv4的输送带纵向撕裂检测[J].计算机系统应用, 2023, 32(3): 186-194.
[22] GEIGER A, LENZ P, STILLER C, et al. Vision meets robotics: the kitti dataset[J]. The International Journal of Robotics Research, 2013, 32(11): 1231-1237.
[23] HAN Zhuang, WANG Tao, XU Kun, et al. Unsupervised monocular depth estimation with full-scale feature fusion[J]. Journal of Qiqihar University, 2023, 39(2): 25-30, 43.
韩壮, 王涛, 许锟, 等. 基于全尺度特征融合的无监督单目深度估计[J].齐齐哈尔大学学报, 2023, 39(2): 25-30, 43.
[24] 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.
[25] 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.
[26] 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, 2020, 34(7): 11677-11684.
[27] YI H, SHI S, DING M, et al. Segvoxelnet: exploring semantic context and depth-aware features for 3d vehicle detection from point cloud[C].2020 IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2020: 2274-2280.
[28] SHI S, WANG Z, SHI J, et al. From points to parts: 3D object detection from point cloud with part-aware and part-aggregation network[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 43(8): 2647-2664.
[29] WANG Shun. Research on 3D target detection algorithm based on lidar point cloud[D]. Wuhan: Jianghan University, 2022.
王顺. 基于激光雷达点云的三维目标检测算法研究[D]. 武汉: 江汉大学, 2022.
[30] HE C, ZENG H, HUANG J, et al. Structure aware single-stage 3d object detection from point cloud[C].Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 11873-11882.

3D Object Detection with Fusion Point Attention Mechanism in LiDAR Point Cloud

LIU Weili[1,2], ZHU Deli[1,2], LUO Huahao[1,2], LI Yi[3]
(1 *School of Computer and Information Science, Chongqing Normal University, Chongqing* 401331, *China*)
(2 *Chongqing Digital Agricultural Service Engineering Technology Research Center, Chongqing* 401331, *China*)
(3 *Information Center of Chongqing Academy of Animal Husbandry, Chongqing* 401331, *China*)

Abstract: With the rapid development of computer vision, object detection has made remarkable achievements in 2D vision tasks, but it still can not solve the problems such as light changes and lack of depth that occur in actual scenes. The 3D data acquired by LiDAR makes up for some defects existing in the 2D vision field, so 3D object detection is widely studied as an important field in 3D scene perception. 3D object detection in the field of autonomous driving is an important part of intelligent transportation, and the 3D object detection algorithm based on LiDAR point cloud provides an important perception means for it. Perception is a key component of autonomous driving, ensuring the intelligence and safety of driving. 3D object detection refers to the detection of physical objects from sensor data, predicting and estimating the category, bounding box, and spatial position of the target. However, due to the unstructured and non-fixed size characteristics of point clouds, they can not be directly processed by 3D object detectors and must be encoded into a more compact structure through some form of expression. There are currently two main types of expressions: point-based and voxel-based methods. Voxel-based methods have higher detection

efficiency, but their detection accuracy is lower than that of methods based on raw point clouds. Therefore, how to improve the detection accuracy of voxel-based methods while ensuring detection efficiency has become a research hotspot in recent years.

In view of the problems of loss of fine-grained information and insufficient ability to extract point cloud features in the 3D object detection algorithm for Pillar-encoded point clouds, this paper proposes a 3D object detection algorithm based on PointPillars that integrates point-wise spatial attention mechanism and CSPNet. Firstly, the point-wise spatial attention mechanism is integrated into the pillar feature network layer, which can enhance the network's ability to extract local geometric information and retain deep-level information, making the obtained key features more suitable for detection tasks. Point-wise spatial attention follows the basic structure of self-attention, which can effectively avoid the impact of redundant point clouds or noise points on features, strengthen the description of features with less coverage of point clouds, and to a certain extent solve the problem of information loss based on Pillar-encoded point clouds. Secondly, replacing the ordinary convolution in the downsampling module that extracts high-dimensional features from pseudo-images of point clouds with CSPNet can achieve gradient flow segmentation, further enhance the network's learning ability while reducing computational complexity, and improve model detection accuracy.

Finally, the algorithm in this paper improves the 3D detection accuracy by 2.23%, 2.25%, and 2.30% in easy, medium, and hard cases, respectively, compared with the benchmark network under the application scenario of highway with car class in KITTI as the detection target. The experimental results show that the algorithm in this paper has significantly improved the detection performance, while the detection speed reaches the real-time detection level, which has some positive significance for the optimization and improvement of autonomous driving technology, and has great potential in the application scenario of highways.

Key words: 3D object detection; Point cloud; Attention mechanism; PointPillars; Cross stage partial network

OCIS Codes: 120.1880; 100.2000; 100.4996; 100.1830

Foundation item: Science and Technology Research Project of Chongqing Municipal Education Commission (No. KJQN201800536), Chongqing University Innovation Research Group Project (No. CXQT20015)