文章编号　1004-924X(2023)19-2910-11

基于特征融合的改进型PointPillar点云目标检测

张　勇，石志广*，沈　奇，张　焱，张　宇
（国防科技大学 电子科学学院 ATR重点实验室，湖南 长沙 410073）

摘要：针对PointPillar在自动驾驶道路场景下对点云稀疏小目标检测效果差的问题，通过引入一种多尺度特征融合策略和注意力机制，提出一种点云目标检测网络Pillar-FFNet。针对网络中的特征提取问题，设计了一种基于残差结构的主干网络；针对馈入检测头的特征图没有充分利用高层特征的语义信息和低层特征的空间信息的问题，设计了一种简单有效的多尺度特征融合策略；针对主干网络提取的特征图中信息冗余的问题，提出了一种卷积注意力机制。为验证所提算法的性能，在KITTI和DAIR-V2X-I数据集上进行实验。实验结果表明，所提出的算法在KITTI数据集上与PointPillar相比，汽车、行人和骑行者的平均精度最大提高分别为0.84%，2.13%和4.02%；在DAIR-V2X-I数据集上与PointPillar相比，汽车、行人和骑行者的平均精度最大提高分别为0.33%，2.09%和4.71%，由此证明了所提方法对点云稀疏小目标检测的有效性。

关　键　词：小目标检测；点云稀疏；PointPillar；残差结构；多尺度特征融合；卷积注意力
中图分类号：TP391.1　　**文献标识码**：A　　**doi**：10.37188/OPE.20233119.2910

Improved PointPillar point cloud object detection based on feature fusion

ZHANG Yong, SHI Zhiguang*, SHEN Qi, ZHANG Yan, ZHANG Yu

(*National Key Laboratory of Science and Technology on Automatic Target Recognition, College of Electronic Science and Technology, National University of Defense Technology, Changsha* 410073, *China*)

* *Corresponding author, E-mail: szgstone75@sina.com*

Abstract: A point cloud object detection network, Pillar-FFNet, is proposed by introducing a multiscale feature fusion strategy and an attention mechanism to address the ineffectiveness of PointPillar in detecting small sparse objects in point clouds in autonomous driving road scenarios. First, a backbone network based on a residual structure is designed for feature extraction in the network. Second, a simple and effective multiscale feature fusion strategy is designed to address the problem that the feature maps fed into the detection head do not make full use of the semantic information of high-level features and the spatial information of low-level features. Finally, a convolutional attention mechanism is proposed to treat information redundancy in the feature maps extracted using the backbone network. To validate the performance of the proposed algorithm, experiments are conducted on the KITTI and DAIR-V2X-I datasets. The results

收稿日期：2023-03-30；**修订日期**：2023-05-04.
基金项目：国家自然科学基金资助项目(No. 62075239)

show that the proposed algorithm achieves maximum average accuracy improvements of 0.84%, 2.13%, and 4.02% for cars, pedestrians, and cyclists, respectively, on the KITTI dataset and maximum average accuracy improvements of 0.33%, 2.09%, and 4.71% for cars, pedestrians, and cyclists, respectively, on the DAIR-V2X-I dataset compared with the PointPillar results. Experimental results demonstrate the effectiveness of the proposed method for the detection of sparse small objects in point clouds.

Key words: small object detection; point cloud sparse; PointPillar; residual structure; multi-scale feature fusion; convolutional attention

1　引　言

近年来，随着深度学习的发展，点云处理技术取得了重大突破[1]。点云目标检测是点云处理的基本任务之一[2]，当前可以分为基于原始点云的检测、基于体素的检测、基于数据降维的检测、基于点云和体素的混合检测4大类。基于原始点云的检测通过在原始点上进行数据处理、分析，从而判断目标类别并回归目标边界框。此类方法的优点是充分利用点云信息，提取的点特征能够有效表征目标，检测效果好，但内存占用高、计算量大[3-5]。基于体素的检测将点云在三维空间中划分为大小一致、规则的体素，再通过提取体素特征进行目标检测。此类方法的三维空间特征表征能力有限，相对于基于原始点云的方法计算成本更小，但三维卷积仍需耗费大量显存和算力[6-7]。基于数据降维的检测将点云转换为二维图像，利用成熟的图像目标检测算法进行检测。该类方法相对于上述三类方法计算成本小、易部署，但数据降维过程中信息会丢失，检测效果相对较差[8-10]。基于点云和体素混合的检测同时利用点云与体素进行检测，综合了基于原始点云和基于体素两种方法的优点。该类方法在充分保留三维空间结构的前提下减少了计算量，但仍保留了三维卷积，因此其计算成本仍高于基于数据降维的检测[11-12]。

PointPillar是从数据降维的角度提出的一种点云目标检测网络，具备良好的工程实用性：第一，它将三维数据转换为二维数据来处理，减少了数据量；第二，它利用二维卷积代替三维卷积这类显存占用率高、计算量大且难以部署的算子来提取特征，减少了计算量，提升了算法的可部署性。但该网络的检测精度低于同级别的其他类别的检测方法，其主要原因有三点：第一，该网络的检测性能受柱体尺寸的影响，柱体尺寸越大，生成的伪图像分辨率越小，运行速度高，但检测效果差；柱体尺寸越小，生成的伪图像分辨率越大，运行速度低，但检测效果好；第二，伪图像通过特征编码网络生成，生成的图像质量直接影响检测结果；第三，用于检测的特征包含大量冗余信息且缺少小目标特征。

本文针对PointPillar对小目标检测效果差的问题，设计了一个点云目标检测网络。具体地，设计了一个以残差结构为基础模块的主干网络，用来提升对伪图像的特征提取能力；设计了一个基于多层特征融合策略的检测头，用来提升小目标的检测性能；设计了一个卷积注意力模块，用于抑制特征图中的冗余信息。最后，在KITTI和DAIR-V2X-I数据集上验证了提出算法的有效性。

2　PointPillar网络

PointPillar的思想是将点云转换为二维伪图像，在图像上提取特征并完成目标分类和边界框回归，其网络结构如图1所示，主要分为三部分：特征编码网络(Feature Encode Network, FEN)、主干网络(Backbone Network, BN)和检测头(Detection Head, DH)。PointPillar首先将点云编码为柱体，利用FEN提取柱特征，为减少三维数据带来的计算量，将柱特征转化为二维伪图像，再利用二维卷积BN提取图像特征，最后经过DH获取检测结果。FEN由两个多层感知机(Multilayer Perceptron, MLP)构成，首先将点云在[X-Y]平面划分柱体，随后将柱体馈入MLP提取柱特征，输入为[N, 4]的柱向量，输出为[N, 64]的柱向量，为避免三维信息带来的计算量，将柱特征在俯视图视角下转换为二维伪图像。二维卷积主干网络由16层卷积直接堆叠构

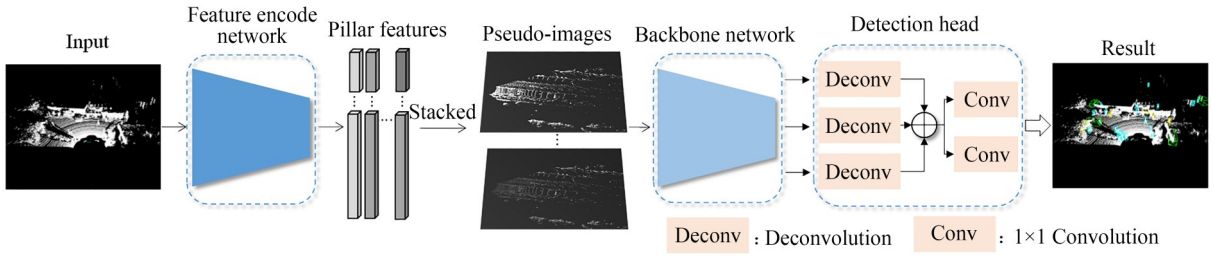


图 1　PointPillar 网络结构

Fig. 1　Structure of PointPillar network

成，用于提取伪图像特征。其输入是尺寸为[C, H, W]的伪图像，输出是尺寸分别为[C, $H/2$, $W/2$]，[$2C$, $H/4$, $W/4$]，[$4C$, $H/8$, $W/8$]3 组不同的特征图 F_1, F_2, F_3。检测头由三个反卷积和两个 1×1 卷积构成，用于在伪图像上检测目标。PointPillar 的检测头与 SSD 检测头结构[13]类似，同样利用了多尺度特征图。3 个反卷积的输入分别为 F_1, F_2 和 F_3，输出均为尺寸为[$2C$, $H/2$, $W/2$]的特征图。3 个反卷积的输出特征经级联后生成尺寸为[$6C$, $H/2$, $W/2$]的特征图，将级联后的特征图输入两个 1×1 卷积分别用于边界框的回归和分类，最后得到检测结果。

PointPillar 通过将三维点云转换为二维伪图像，大大减少了后续需要处理的数据量；在伪图像上利用基于图像的目标检测算法检测目标，避免使用 3D 卷积，使得算法轻量、高效且易部署。但它对小目标的检测效果差，主要原因有两点：首先，主干网络输出的特征图均为高层特征，不利于小目标检测；其次，检测头将主干网络输出的特征图直接进行反卷积和级联操作，导致特征图中包含大量噪声且语义信息和空间信息利用不充分。针对上述问题，本文提出了一种基于特征融合策略的点云目标检测算法 Pillar-FFNet。

3　Pillar-FFNet 结构

为了提高对小目标的检测效果，本文将 PointPillar 作为基础网络，提出了一种点云目标检测网络 Pillar-FFNet，其结构如图 2 所示。Pillar-FFNet 由特征编码网络、主干网络和检测头组成，特征编码网络与 PointPillar 中的一致。设计了一个基于残差的主干网络，输出四组不同尺

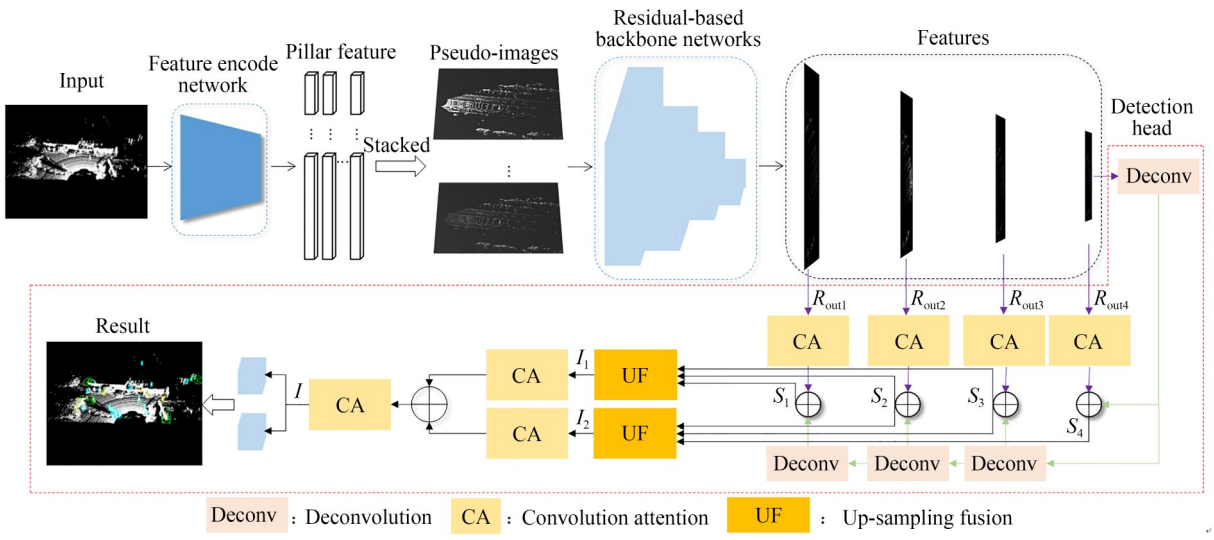


图 2　Pillar-FFNet 结构

Fig. 2　Structure of pillar-FFNet

寸的包含丰富语义信息和空间信息的特征图；设计了一个基于多尺度特征融合策略的检测头，通过有效融合主干网络输出特征图中的信息来提升小目标的检测效果；设计了一个卷积注意力模块，通过有效增强特征图中的有效信息来提升检测效果。Pillar-FFNet的检测流程为：首先，将点云馈入特征编码网络提取点特征，根据点特征生成伪图像；然后，将伪图像馈入残差主干网络，生成4组不同尺度的特征图；最后，将主干网络生成的特征图馈入检测头，完成点云目标检测。

3.1　主干网络

通过直接堆叠卷积来增加网络深度会加剧反向传播过程中梯度消失的现象，导致网络性能退化[14]。解决这一问题的常用方法是利用残差结构来构建深层网络。基于此，本文将残差结构作为基础模块构建了一个基于卷积残差块（Convolution Residual Block，CR）的主干网络，其结构图如图3所示。一个CR块共包含两个分支，分别由1×1卷积和3×3卷积构成，每个卷积后都包含一个批量归一化层BN和一个激活函数ReLU。对于一个CR块，假设输入特征图为$F\in\mathbf{R}^{C\times H\times W}$、输出特征图为$F_{\text{out}}$。CR块的第一个分支由1个3×3卷积和2个1×1卷积构成，F经过1个1×1卷积将数据降维为$F_1'\in\mathbf{R}^{(C/2)\times H\times W}$，$F_1'$经过1个步长为$S$的3×3卷积输出特征$F_1''\in\mathbf{R}^{(C/2)\times(H/S)\times(W/S)}$，$F_1''$经过1个1×1卷积将数据增维为$F_1'''\in\mathbf{R}^{C\times(H/S)\times(W/S)}$；CR块的第二个分支由1个步长为$S$的3×3卷积构成，$F$经过3×3卷积输出特征$F_2'\in\mathbf{R}^{C\times(H/S)\times(W/S)}$；将第一个分支的输出$F_1'''$和第二个分支的输出$F_2'$级联得到$F_3\in\mathbf{R}^{2C\times(H/S)\times(W/S)}$，再将$F_3$经过一个1×1卷积输出$F_{\text{out}}\in\mathbf{R}^{C\times(H/S)\times(W/S)}$。

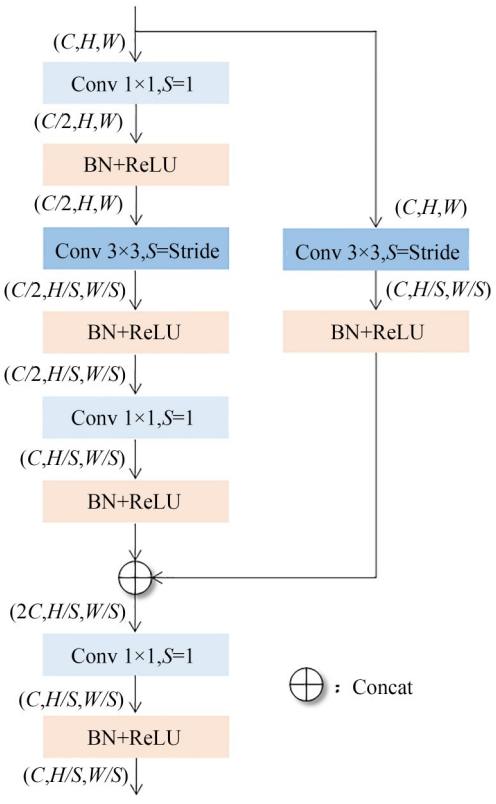


图3　CR块结构

Fig. 3　Structure of CR block

本文以CR块作为基础块设计了一个残差主干网络（Residual BackBone Net，RBNet）。如图4所示，RBNet由14个步长S为1的CR块和3个步长S为2的CR块堆叠而成，输出4个不同尺寸的特征图$\boldsymbol{R}_{\text{out1}}$，$\boldsymbol{R}_{\text{out2}}$，$\boldsymbol{R}_{\text{out3}}$，$\boldsymbol{R}_{\text{out4}}$。RBNet的特征提取过程为：点云首先经特征编码网络生成伪图像$\boldsymbol{M}\in\mathbf{R}^{64\times496\times432}$，$\boldsymbol{M}$为RBNet的输入；$\boldsymbol{M}$经过2个步长为1的CR块，生成特征图$\boldsymbol{R}_{\text{out1}}\in\mathbf{R}^{64\times496\times432}$；$\boldsymbol{R}_{\text{out1}}$经过1个步长为2的CR块和2个步长为1的CR块，生成特征图$\boldsymbol{R}_{\text{out2}}\in\mathbf{R}^{64\times248\times216}$；$\boldsymbol{R}_{\text{out2}}$经过1个步长为2的CR块和5个步长为1的CR块，生成特征图$\boldsymbol{R}_{\text{out3}}\in\mathbf{R}^{128\times124\times108}$；$\boldsymbol{R}_{\text{out3}}$经过1个步长为2的CR块和5个步长为1的CR块，生成特征图$\boldsymbol{R}_{\text{out4}}\in\mathbf{R}^{256\times62\times54}$。

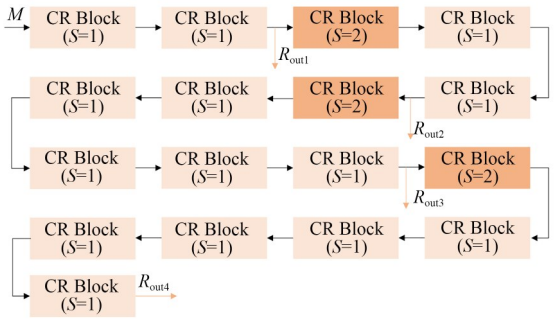


图4　RBNet结构

Fig. 4　Structure of RBNet

3.2　基于多尺度特征融合策略的检测头

卷积神经网络中，低层特征图的分辨率大，空间信息丰富；高层特征图的分辨率低，语义信息丰富[15]。为了将不同层级特征图中的高层语义信息和低层空间信息有效融合，本文提出了一

个基于多尺度特征融合策略的检测头，命名为MFHead。MFHead通过有效融合不同层级的语义信息和空间信息来提升点云中小目标的检测效果。

MFHead结构如图2中红色虚线框内的检测头部分所示(彩图见期刊电子版)，输入为RBNet输出的4组不同尺寸的特征图$\boldsymbol{R}_{\text{out1}}$，$\boldsymbol{R}_{\text{out2}}$，$\boldsymbol{R}_{\text{out3}}$和$\boldsymbol{R}_{\text{out4}}$，输出为检测框位置、目标尺寸、目标偏转角和目标类别。首先，$\boldsymbol{R}_{\text{out4}}$经过4个反卷积生成4组特征图，其尺寸分别为256×62×54，128×124×108，64×248×216和64×496×432，将$\boldsymbol{R}_{\text{out1}}$，$\boldsymbol{R}_{\text{out2}}$，$\boldsymbol{R}_{\text{out3}}$和$\boldsymbol{R}_{\text{out4}}$经过卷积注意力模块后分别与其进行级联融合生成4组特征图$\boldsymbol{S}_1$，$\boldsymbol{S}_2$，$\boldsymbol{S}_3$和$\boldsymbol{S}_4$，将$\boldsymbol{S}_1$，$\boldsymbol{S}_2$，$\boldsymbol{S}_3$进行级联融合生成特征图$\boldsymbol{I}_1$，将$\boldsymbol{S}_2$，$\boldsymbol{S}_3$，$\boldsymbol{S}_4$进行级联融合生成特征图$\boldsymbol{I}_2$。然后将$\boldsymbol{I}_1$，$\boldsymbol{I}_2$分别经过卷积注意力模块后进行级联融合生成最终的特征图$\boldsymbol{I}$，将$\boldsymbol{I}$馈入两个1×1卷积分别进行目标分类和边界框回归。

3.3　卷积注意力模块

卷积神经网络提取的特征图中包含丰富的细节信息，但特征图中信息的重要性不一样，某些信息对目标的检测识别更重要。因此，为了能充分利用网络中与任务相关的有效信息，抑制网络中与任务无关的冗余信息，本文设计了一个卷积注意力模块(CAMA)，其结构如图5所示。首先将特征图$F\in\mathrm{R}^{C\times H\times W}$按通道拆分为两组特征图$F_1\in\mathrm{R}^{(C/2)\times H\times W}$，$F_2\in\mathrm{R}^{(C/2)\times H\times W}$，然后将$F_1$馈入平均池化层得到$F_1'\in\mathrm{R}^{(C/2)\times 1\times 1}$，$F_1'$经过一个1×1卷积和sigmoid函数，输出每个特征图的权重值β，将F_1与β相乘得到F_1''，将F_1''与F_1相加得到F_1^{out}；再将F_2馈入最大池化层得到$F_2'\in\mathrm{R}^{(C/2)\times 1\times 1}$，$F_2'$经过一个1×1卷积和sigmoid函数，输出每个特征图的权重值α，将F_2与α相乘得到F_2''，将F_2''与F_2相加得到F_2^{out}；最后，将F_1^{out}和F_2^{out}相加后馈入1×1卷积得到被赋予权重的特征图F_{out}。本文的卷积注意力模块可以表示为：

$$\begin{cases} F_1^{\text{out}}=S\left(f_1\left(\text{Avgpool}\left(F_1\right)\right)\right)\times F_1+F_1 \\ F_2^{\text{out}}=S\left(f_1\left(\text{Maxpool}\left(F_2\right)\right)\right)\times F_2+F_2 \\ F_{\text{out}}=f_1\left(F_1^{\text{out}}+F_2^{\text{out}}\right) \end{cases},\tag{1}$$

式中：$S(\cdot)$表示sigmoid函数，$f_1(\cdot)$表示1×1卷积，Avgpool(·)表示平均池化，Maxpool(·)表示最大池化。

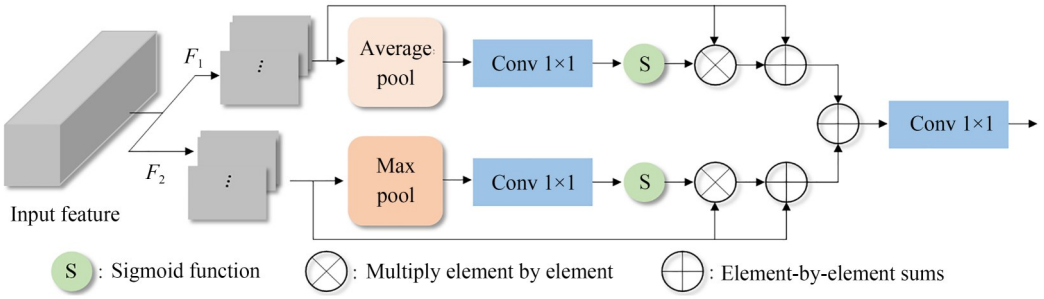


图5　CAMA模块

Fig. 5　CAMA module

3.4　损失函数

本文沿用文献[7]中的损失函数。损失函数分为回归损失和分类损失。回归损失采用Smooth L1函数，分类损失采用Focal Loss。假设目标的三维先验框为$\left(x,y,z,w,l,h,\theta\right)$，真实边界框表示为$\left(x^{\text{gt}},y^{\text{gt}},z^{\text{gt}},w^{\text{gt}},l^{\text{gt}},h^{\text{gt}},\theta^{\text{gt}}\right)$，预测边界框表示为$\left(x',y',z',w',l',h',\theta'\right)$，则边界框回归损失函数$L_{\text{loc}}$表示为：

$$\begin{aligned} &\Delta x=\frac{x^{\text{gt}}-x}{d},\Delta y=\frac{y^{\text{gt}}-y}{d},\Delta z=\frac{z^{\text{gt}}-z}{h} \\ &\Delta w=\log\frac{w^{\text{gt}}}{w},\Delta l=\log\frac{l^{\text{gt}}}{l} \\ &\Delta h=\log\frac{h^{\text{gt}}}{h},\Delta\theta=\theta^{\text{gt}}-\theta \\ &L_\theta=\text{SmoothL}1\left(\sin\left(\Delta\theta-\theta'\right)\right) \\ &L_{\text{loc}}=L_\theta+\sum_{b\in\left(x',y',z',w',l',h'\right)}\text{SmoothL}1\left(\Delta b\right) \end{aligned},\tag{2}$$

式中$d=\sqrt{l^2+w^2}$，x，y，z为中心坐标，w，l，h分

别为宽，长，高，θ 为目标偏转角。

对于偏转角回归，使用 L_{dir} 进一步在离散方向学习边界框回归方向，由于 sin 函数特性无法区分 θ 为 0 °和 180°，因此利用 softmax 函数对角度进行分类。softmax 的输出值大于 0 时，角度为正，反之为负。

分类损失函数 L_{cls} 表示为：

$$L_{\text{cls}}=-\alpha_{\text{a}}\left(1-p^{\text{a}}\right)^{\gamma}\log p^{\text{a}}, \tag{3}$$

其中：p^{a} 是 anchor 的类别概率，设置 $\alpha=0.25$，$\gamma=2$。

总损失函数 L 表示为：

$$L=\frac{1}{N_{\text{pos}}}\left(\beta_{\text{loc}}L_{\text{loc}}+\beta_{\text{cls}}L_{\text{cls}}+\beta_{\text{dir}}L_{\text{dir}}\right), \tag{4}$$

式中：N_{pos} 表示正锚框的数量，$\beta_{\text{loc}}=2$，$\beta_{\text{cls}}=2$，$\beta_{\text{dir}}=2$。

4　实　验

本文的实验环境为：Intel® Core™ i9-9820X CPU @ 3.30GHz×20 CPU NVIDIA GeForce RTX 2080 GPU，64 G 内存，Ubuntu18.04 系统。采用 Python 语言在 Pytorch，OpenPCDet 点云目标检测框架下进行实验验证。

4.1　数据集

本文在公开数据集 KITTI[16] 和 DAIR-V2X-I[17] 上进行实验验证。KITTI 含有带标签的训练样本为 7 481 帧。将 7 481 帧样本划分为训练集和验证集，训练集包含 3 712 帧样本，验证集包含 3 769 帧样本，对汽车（Car）、行人（Pedestrian）和骑行者（Cyclist）三类目标进行检测。DAIR-V2X-I 包含 10 084 帧样本，其中公开的带标签样本共 7 058 帧。将 7 058 帧样本划分为训练集和验证集，训练集包含 5 042 帧样本，验证集包含 2 016 帧样本，对汽车，行人和骑行者三类目标进行检测。

4.2　实验设置

本文按照 OpenPCDet 的设定来设置网络的学习策略和超参数，按照文献[8]中的设定来设置数据集中点云的柱体尺寸、每个柱体中的最大点数、点云数据范围内所含最大柱体数、锚框尺寸、正负样本交并比（Intersection Over Union，IoU）匹配阈值。由于 KITTI 和 DAIR-V2X-I 数据集中的有效范围不一样，因此设置的点云范围不同。KITTI 数据集的点云范围设置为 $x\in[0,69.12]$，$y\in[-39.68,39.68]$，$z\in[-3,1]$；DAIR-V2X-I 数据集的点云范围设置为 $x\in[0,99.84]$，$y\in[-39.68,39.68]$，$z\in[-3,1]$。

4.3　评价指标

本文采用精确率-召回率（Precision-Recall，PR）曲线、平均精度（Average Precision，AP）和每秒帧数（Frame Per Second，FPS）来衡量算法的性能。

PR 曲线是一种评价模型性能的指标，以召回率（R）为横坐标，精确率（P）为纵坐标。其定义如下：

$$R=\frac{TP}{TP+FN}, \tag{5}$$

$$P=\frac{TP}{TP+FP}, \tag{6}$$

式中：TP 表示预测为正且实际为正的样本数量，FN 表示预测为负但实际为正的样本数量，FP 表示预测为正但实际为负的样本数量。AP 是一种评价目标检测模型检测效果的指标，这里采用文献[16]中 AP 的定义。对于汽车，设置当 IoU≥0.7 时检测正确；对于行人和骑行者，设置当 IoU≥0.5 时检测正确。按照 KITTI 设定，根据目标大小、遮挡和截断情况将 3 类目标的检测难度分为简单（easy）、中等（middle）和困难（hard）3 种。本文在这 3 种不同检测难度下评估算法性能。FPS 是一种用来衡量模型推理效率的指标，表示 1 秒内处理样本的数量。在 batch size 为 1 时统计验证集上的指标。一般地，PR 曲线包含的面积越大，模型性能越好；AP 值越大，模型性能越好；FPS 越大，模型推理速度越快。

4.4　实验结果及分析

本文对 KITTI 和 DAIR-V2X-I 数据集中的汽车、行人和骑行者进行三维检测，为公平比较，本文所用算法采用同样的损失函数、相同的超参数在台式工作站上进行训练。

4.4.1　定量分析

在三维目标检测上采用 AP，PR 曲线和 FPS 进行定量分析，将 Pillar-FFNet 与 3 种主流点云目标检测算法对比，实验结果如表 1、表 2 和图 6 所示，加粗字体表示最佳指标。表 1 和表 2 分别为 KITTI 验证集和 DAIR-V2X-I 验证集上 4 种算

表 1　KITTI 验证集上的三维检测结果

Tab. 1　Result for 3D detection on KITTI validation dateset

Algorithm	Car(IoU=0.7)			Pedestrian(IoU=0.5)			Cyclist(IoU=0.5)			mAP/%	FPS/
	easy	middle	hard	easy	middle	hard	easy	middle	hard	(middle)	(frame•s^{-1})
Second	**88.25**	**78.56**	**77.14**	54.72	50.65	45.93	80.72	63.92	60.89	64.37	26.25
PointPillar	87.08	77.55	75.82	54.43	49.31	45.56	81.45	63.16	59.91	63.34	**36.49**
PillarNet	87.25	77.72	76.49	51.43	47.53	44.42	81.95	63.20	59.12	62.81	19.96
Pillar-FFNet	87.92	78.17	76.62	**56.24**	**51.44**	**46.72**	**85.47**	**65.55**	**61.49**	**65.05**	10.00

表 2　DAIR-V2X-I 验证集上的三维检测结果

Tab. 2　Result for 3D detection on the DAIR-V2X-I validation dateset

Algorithm	Car(IoU=0.7)			Pedestrian(IoU=0.5)			Cyclist(IoU=0.5)			mAP/%	FPS/
	easy	middle	hard	easy	middle	hard	easy	middle	hard	(middle)	(frame•s^{-1})
Second	80.13	68.84	68.87	64.35	60.41	60.57	64.64	39.32	39.43	56.19	12.69
PointPillar	80.04	68.86	68.89	63.35	62.06	62.09	61.51	37.51	37.61	56.14	**15.13**
PillarNet	80.21	**69.82**	**69.86**	64.89	60.82	60.84	62.98	38.66	**40.06**	56.43	11.42
Pillar-FFNet	**80.37**	69.03	69.06	**65.44**	**62.23**	**62.26**	**66.22**	**39.35**	39.43	**56.87**	8.18

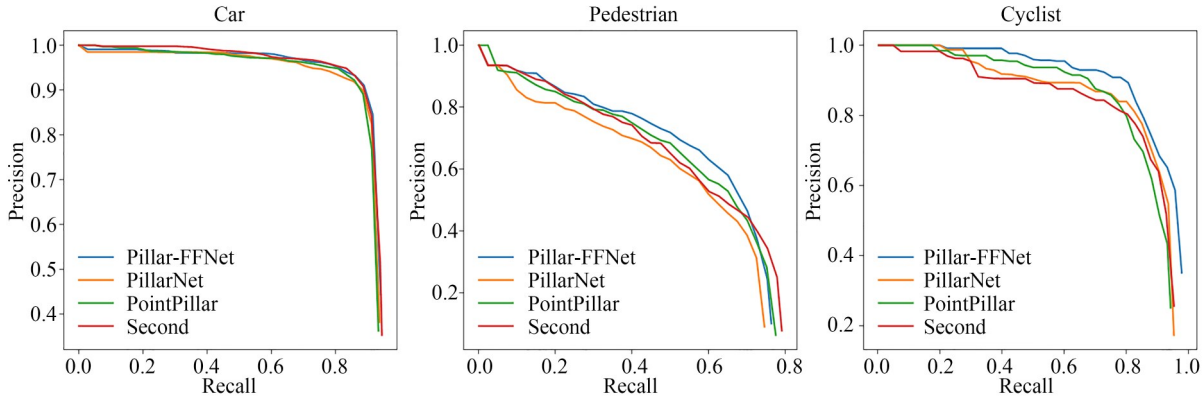


图 6　四种对比算法在 KITTI 验证集上的 PR 曲线

Fig. 6　PR curves for four comparison algorithms on KITTI validation dataset

法的三维检测 AP 值和 FPS；图 6 为 4 种算法在 KITTI 验证集简单检测难度下的 PR 曲线。由表 1 和表 2 可知：Pillar-FFNet 对三类目标的总体检测和小目标的效果均优于其他 3 种算法，但推理速度低于其他算法。在 KITTI 验证集上，与 PointPillar 相比，FPS 降低了 26.49 frame/s，汽车、行人和骑行者的 AP 在 easy 检测难度下分别提高了 0.84%，1.81%，4.02%，在 middle 检测难度下分别提高了 0.62%，2.13%，2.39%，在 hard 检测难度下分别提高了 0.8%，1.16%，1.58%；在 DAIR-V2X-I 验证集上，与 PointPillar 相比，FPS 降低了 6.95 frame/s，汽车，行人和骑行者的 AP 在 easy 检测难度下分别提高了 0.33%，2.09%，4.71%，在 middle 检测难度下分别提高了 0.17%，0.17%，1.84%，在 hard 检测难度下分别提高了 0.17%，0.17%，1.82%。由图 6 可知，Pillar-FFNet 在行人和骑行者类别上的 PR 曲线明显优于其他 3 个算法，在汽车类别上的 PR 曲线与其他算法性能相当。

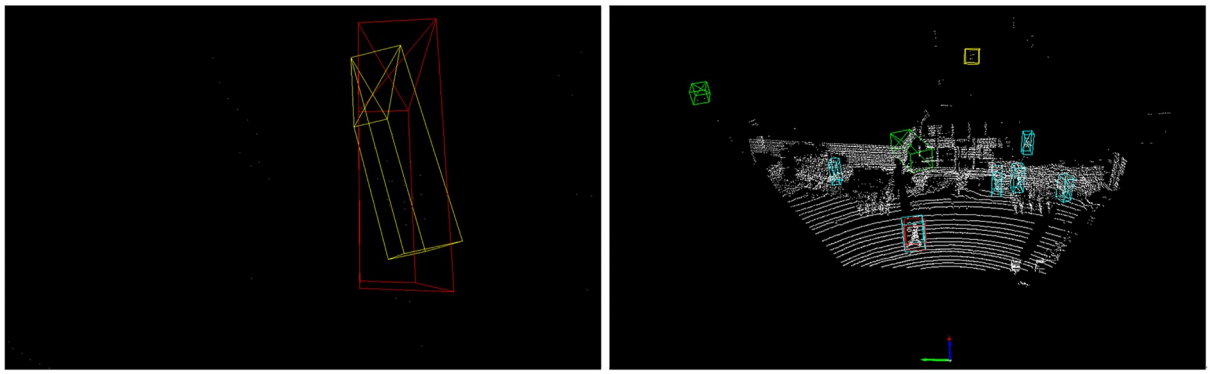
在点云数据中，由于点云自身的稀疏性，用于表征小目标的的点数少，因此对点云小目标的检测更加困难。与 PointPillar 相比，Pillar-FFNet

可以在不影响汽车检测的前提下有效提高行人和骑行者这类小目标检测的性能。其主要原因有：首先，本文设计的 RBNet 生成的 4 个不同尺寸的特征图中包含不同尺寸目标的信息，有利于后续检测头中的特征提取与检测分类；其次，MFHead 检测头将低层特征中的空间信息和高层特征中的语义信息进行有效融合，使得馈入检测头的特征图中含有各类目标的丰富信息；最后，CAMA 模块通过计算特征图中对应像素的权重，有效抑制了主干网络输出特征图和融合后特征图中的冗余信息，增强了特征图中的有效信息。但由于本文设计的检测头采用多尺度特征融合策略，增加了需要处理的特征信息，且包含多个反卷积和卷积注意力模块，计算量增加，降低了网络推理速度。由于 DAIR-V2X-I 数据集中的点云范围更大、更加稠密，因此算法对 DAIR-V2X-I 的处理效率更低。

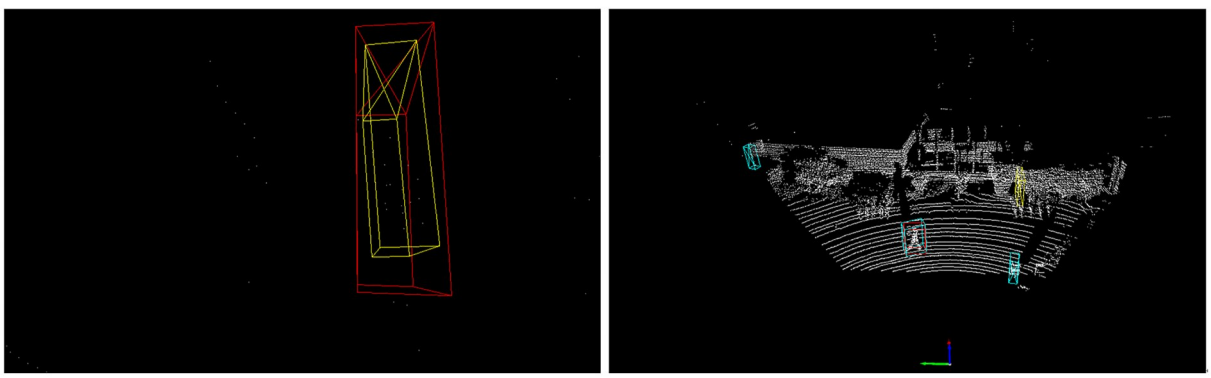
4.4.2　定性分析

这里对 Pillar-FFNet 和 PointPillar 的检测结果进行了可视化分析，算法的部分检测结果如图 7 所示。图 7 中，对应位置表示同一帧点云经过两种不同算法的检测结果，红色框表示目标真实框，绿色框、蓝色框和黄色框分别表示网络预测的汽车、行人和骑行者的位置。

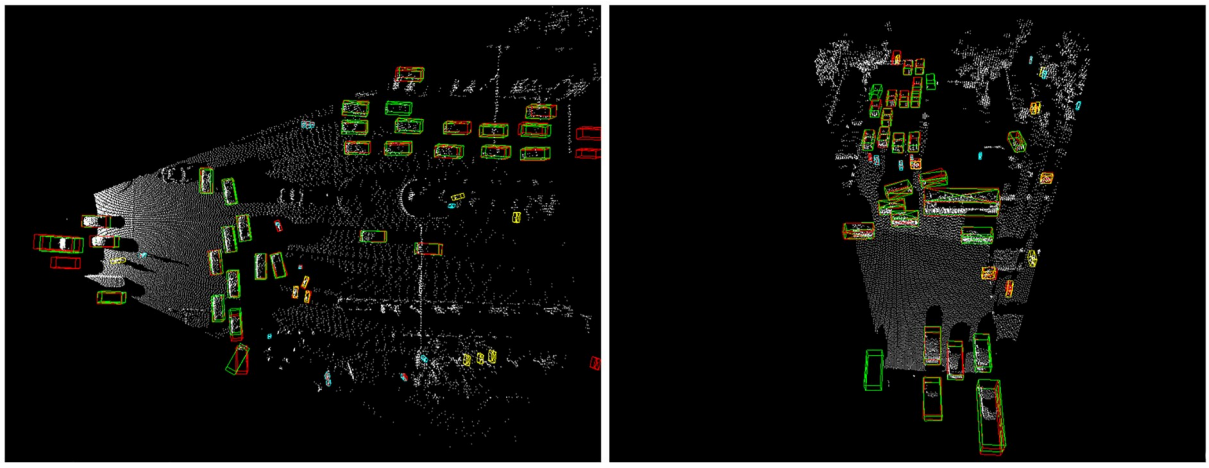
由图 7 可知，Pillar-FFNet 对行人和骑行者的检测效果有显著提升，同时不会降低汽车的检测效果。与 PointPillar 相比，Pillar-FFNet 在小目标检测上整体更加准确，但对于远距离的目标和较密集场景中的目标仍然存在较大的漏检和误检。首先，对于远距离目标的漏检和误检问题，远距离的目标点云数据中包含的有效点太少，导致从点中难以提取到能够有效表征其目标特性的特征，这是产生漏检的主要原因；远距离的一些目标在扫描成点云后在空间结构上与汽车、行人和骑行者的部分高度相似，这是产生误检的主要原因。其次，密集场景中目标之间的遮挡和自身遮挡等问题严重，导致采集的点云数据不全面，也会导致漏检和误检。最后，由于本文算法是从伪图像中提取目标特征，点云中的目标点过于稀疏，导致其在图像上对应的像素点太少，从而影响检测。



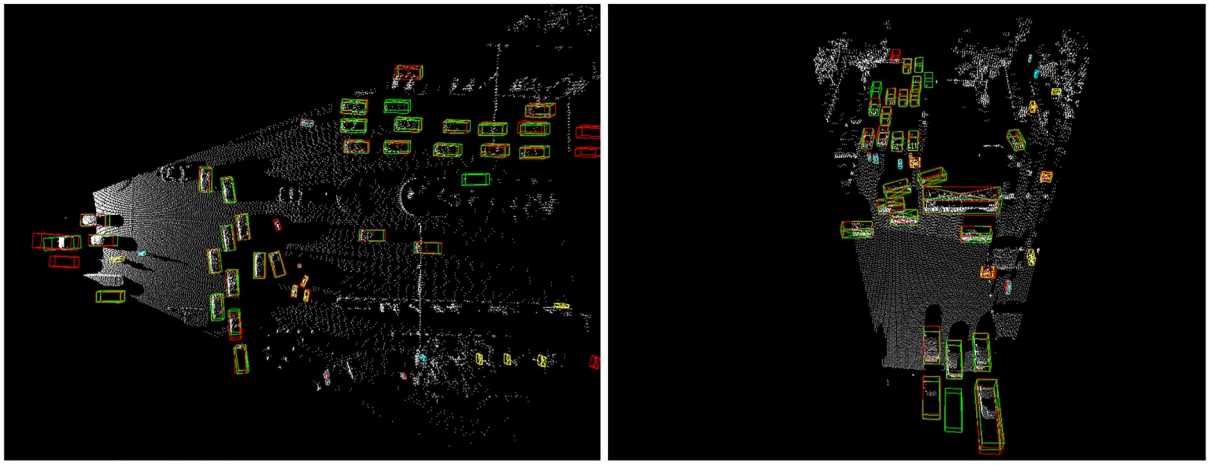
(a) KITTI验证集上PointPillar检测结果
(a) PointPillar detection results on KITTI validation dataset



(b) KITTI验证集上Pillar-FFNet检测结果
(b) Pillar-FFNet detection results on KITTI validation dataset



(c) DAIR-V2X-I验证集上PointPillar检测结果
(c) PointPillar detection results on DAIR-V2X-I validation dataset



(d) DAIR-V2X-I验证集上Pillar-FFNet检测结果
(d) Pillar-FFNet detection results on DAIR-V2X-I validation dataset

图 7　可视化检测结果

Fig. 7　Visualisation of detection results

4.4.3　消融实验

为验证CAMA模块和MFHead检测头对三维检测的影响，在KITTI上进行了消融实验。实验采用4.3节的评价指标来评估算法性能，实验结果如表3～表5所示，其中加粗字体表示最佳指标。

首先，设计4组消融实验来验证检测头中不同的特征融合方式对检测性能的影响：实验1将PointPillar中的检测头替代Pillar-FFNet中的MFHead；实验2将MFHead中的S_1和S_2融合，S_3和S_4融合；实验3将MFHead中的S_1，S_3，S_4融合，S_2，S_3，S_4融合；实验4将MFHead替换为FPN[18]结构，实验结果如表3所示。然后，设计4组实验来验证注意力对检测性能的影响：实验5将CAMA模块用SE模块[19]代替；实验6将CAMA模块用CBAM模块[20]代替；实验7将CAMA模块用ECA模块[21]代替；实验8将CAMA模块

表 3　Pillar-FFNet检测头不同融合方式对检测的影响

Tab. 3　Effect of different fusion methods on detection of Pillar-FFNet detection heads

Algorithm	Car	Pedestrian	Cyclist	mAP/%
Pillar-FFNet	78.17	51.44	**65.55**	**65.05**
experiment 1	77.30	**52.83**	61.69	63.94
experiment 2	**78.21**	52.76	62.98	64.65
experiment 3	74.13	46.86	52.34	57.78
experiment 4	77.31	49.19	62.83	63.24

表 4　Pillar-FFNet不同注意力模块对检测的影响

Tab. 4　Effect of different attention modules of Pillar-FFNet on detection

Algorithm	Car	Pedestrian	Cyclist	mAP/%
Pillar-FFNet	**78.17**	**51.44**	**65.55**	65.05
experiment 5	77.73	51.35	63.63	64.23
experiment 6	77.75	47.04	63.21	62.67
experiment 7	77.53	51.41	63.86	64.27
experiment 8	77.98	49.86	60.59	62.81

表 5　本文设计的模块对检测的影响

Tab. 5　Effect of modules designed in paper on detection

Algorithm	Car	Pedestrian	Cyclist	mAP/%
CAMA＋MFHead	**78.17**	51.44	**65.55**	**65.05**
CAMA	78.07	48.08	61.23	62.46
MFHead	76.85	**52.94**	64.45	64.74

用3×3卷积代替，实验结果如表4所示。最后，设计3组实验验证MFHead检测头和CAMA模块对检测性能的影响：第一组实验同时采用MFHead和CAMA模块；第二组实验仅采用CAMA模块；第三组实验仅采用MFHead检测头，实验结果如表5所示。

根据表3～表5可知，不同的检测头特征融合方式和注意力模块下，检测性能的差异较大。由表3可知，MFHead检测头相对于SSD检测头和FPN结构具有更好的检测效果；由表4可知，CAMA模块相对于SE，CBAM和ECA注意力机制能够进一步提升三维检测的性能。由表5可知，同时采用CAMA和MFHead的模块综合检测性能最好。

5　结　论

针对点云稀疏小目标检测困难的问题，本文结合多尺度特征融合策略和卷积注意力设计了一种点云目标检测网络。在公开数据集上进行了实验验证，结果表明，与基准算法相比，本文算法在KITTI和DAIR-V2X-I数据集上对行人和骑行者这两类小目标的3D检测精度最大分别提高了2.13%和4.71%。但由于点云自身特性的限制，点云中表征小目标的点数量少，算法提取到的小目标特征少，导致小目标准确检测困难。因此，后续的研究重点是对点云中的小目标进行有效补全，以提升小目标检测的准确性。

参考文献：

[1] BELLO S A, YU S S, WANG C, *et al*. Review: deep learning on 3D point clouds[J]. *Remote Sensing*, 2020, 12(11): 1729.

[2] 邓家俊. 面向视频和点云数据的目标检测方法研究[D]. 合肥：中国科学技术大学, 2021.
DENG J J. *Object Detection for Video Sequence and Point Clouds* [D]. Hefei: University of Science and Technology of China, 2021. (in Chinese)

[3] SHI S S, WANG X G, LI H S. PointRCNN: 3D object proposal generation and detection from point cloud[C]. 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). *June* 15-20, 2019, *Long Beach*, *CA*, *USA*. IEEE, 2020: 770-779.

[4] YANG Z T, SUN Y N, LIU S, *et al*. 3DSSD: point-based 3D single stage object detector[C]. 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). *June* 13-19, 2020, *Seattle*, *WA*, *USA*. IEEE, 2020: 11037-11045.

[5] ZHANG Y F, HU Q Y, XU G Q, *et al*. Not all points are equal: learning highly efficient point-based detectors for 3D LiDAR point clouds[C]. 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). *June* 18-24, 2022, *New Orleans*, *LA*, *USA*. IEEE, 2022: 18931-18940.

[6] ZHOU Y, TUZEL O. VoxelNet: end-to-end learning for point cloud based 3D object detection[C]. 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. *June* 18-23, 2018, *Salt Lake City*, *UT*, *USA*. IEEE, 2018: 4490-4499.

[7] YAN Y, MAO Y X, LI B. SECOND: sparsely embedded convolutional detection[J]. *Sensors*, 2018, 18(10): 3337.

[8] LANG A H, VORA S, CAESAR H, *et al*. PointPillars: fast encoders for object detection from point clouds[C]. 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). *June* 15-20, 2019, *Long Beach*, *CA*, *USA*. IEEE,

2020: 12689-12697.

[9] BELTRÁN J, GUINDEL C, MORENO F M, *et al*. BirdNet: a 3D object detection framework from LiDAR information [C]. 2018 21*st International Conference on Intelligent Transportation Systems* (*ITSC*). *November* 4-7, 2018, *Maui*, *HI*, *USA*. IEEE, 2018: 3517-3523.

[10] SHI G S, LI R F, MA C. *PillarNet: Real-time and High-performance Pillar-based 3D Object Detection* [M]. Lecture Notes in Computer Science. Cham: Springer Nature Switzerland, 2022: 35-52.

[11] NOH J, LEE S, HAM B. HVPR: hybrid voxel-point representation for single-stage 3D object detection[C]. 2021 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). *June* 20-25, 2021, *Nashville*, *TN*, *USA*. IEEE, 2021: 14600-14609.

[12] 田枫, 姜文文, 刘芳, 等. 混合体素与原始点云的三维目标检测方法[J]. 重庆理工大学学报(自然科学), 2022(11): 108-117.
TIAN F, JIANG W W, LIU F, *et al*. 3D object detection method based on mixed voxels and original point clouds[J]. *Journal of Chongqing University of Technology* (*Natural Science*), 2022(11): 108-117. (in Chinese)

[13] LIU W, ANGUELOV D, ERHAN D, *et al*. *SSD: Single Shot MultiBox Detector* [M]. Computer Vision - ECCV 2016. Cham: Springer International Publishing, 2016: 21-37.

[14] HE K M, ZHANG X Y, REN S Q, *et al*. Deep residual learning for image recognition [C]. 2016 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*). *June* 27-30, 2016, *Las Vegas*, *NV*, *USA*. IEEE, 2016: 770-778.

[15] 申利华, 李波. 基于特征金字塔网络和密集网络的肺部CT图像超分辨率重建[J]. 计算机应用, 2023, 43(5): 1612-1619.
SHEN L H, LI B. Super-resolution reconstruction of lung CT images based on feature pyramid network and dense network[J]. *Journal of Computer Applications*, 2023, 43(5): 1612-1619. (in Chinese)

[16] GEIGER A. Are we ready for autonomous driving? The KITTI vision benchmark suite[C]. *Proceedings of the* 2012 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*). New York: ACM, 2012: 3354-3361.

[17] YU H B, LUO Y Z, SHU M, *et al*. DAIR-V2X: a large-scale dataset for vehicle-infrastructure cooperative 3D object detection[C]. 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). *June* 18-24, 2022, *New Orleans*, *LA*, *USA*. IEEE, 2022: 21329-21338.

[18] 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*. IEEE, 2017: 936-944.

[19] HU J, SHEN L, SUN G. Squeeze-and-excitation networks [C]. 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. *June* 18-23, 2018, *Salt Lake City*, *UT*, *USA*. IEEE, 2018: 7132-7141.

[20] WOO S, PARK J, LEE J Y, *et al*. *CBAM: Convolutional Block Attention Module* [M]. Computer Vision-ECCV 2018. Cham: Springer International Publishing, 2018: 3-19.

[21] WANG Q L, WU B G, ZHU P F, *et al*. ECA-net: efficient channel attention for deep convolutional neural networks[C]. 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). *June* 13-19, 2020, *Seattle*, *WA*, *USA*. IEEE, 2020: 11531-11539.

作者简介：

张　勇(1995—),男,湖南长沙人,硕士,工程师,2018年于吉首大学获得学士学位,2021年于湖南师范大学获得硕士学位,主要从事自动目标识别及点云数据处理方面的研究。E-mail: zhangyong@hunnu.edu.cn

通讯作者：

石志广(1975—),男,湖南长沙人,副研究员,1996年于石家庄机械工程学院获得学士学位,2002年、2007年于国防科技大学分别获得硕士和博士学位,主要从红外与激光雷达成像信息处理、自动目标识别方面的研究。E-mail: szgstone75@sina.com