文章编号：1007-2780(2024)01-0079-10

基于知识蒸馏和定位引导的 Pointpillars 点云检测网络

赵　晶[1,4]，李少博[1,2]，郭杰龙[2,3*]，俞　辉[2,3]，张剑锋[2,3]，李　杰[2,3]
(1. 厦门理工学院 电气工程与自动化学院，福建 厦门 361024；
2. 中国科学院 福建物质结构研究所，福建 福州 350108；
3. 中国科学院 海西研究院 泉州装备制造研究中心，福建 泉州 362000；
4. 厦门市高端电力装备及智能控制重点实验室，福建 厦门 361024)

摘要：激光雷达数据由于其几何特性，被广泛应用于三维目标检测任务中。由于点云数据的稀疏性和不规则性，难以实现特征提取的质量和推理速度间的平衡。本文提出一种基于体柱特征编码的三维目标检测算法，以 Pointpillars 网络为基础，设计 Teacher-Student 模型框架对回归框尺度进行蒸馏，增加蒸馏损失，优化训练网络模型，提升特征提取的质量。为进一步提高模型检测效果，设计定位引导分类项，增加分类预测和回归预测之间的相关性，提高物体识别准确率。本网络所做改进没有引入额外的网络嵌入。算法在 KITTI 数据集上的实验结果表明，相比于基准网络，在三维模式下的平均精度值从 60.65% 提升到了 64.69%，鸟瞰图模式下的平均精度值从 67.74% 提升到 70.24%。模型推理速度为 45 FPS，在提升检测精度的同时满足了实时性要求。

关　键　词：激光点云；三维目标检测；知识蒸馏；分类置信度
中图分类号：TP391.4　　**文献标识码：**A　　**doi：**10.37188/CJLCD.2023-0058

Pointpillars point cloud detection network based on knowledge distillation and location guidance

ZHAO Jing[1,4], LI Shaobo[1,2], GUO Jielong[2,3*], YU Hui[2,3], ZHANG Jianfeng[2,3], LI Jie[2,3]

(1. *School of Electrical Engineering and Automation, Xiamen University of Technology, Xiamen* 361024, *China*;
2. *Fujian Institute of Research on the Structure of Matter, Chinese Academy of Sciences, Fuzhou* 350108, *China*;
3. *Quanzhou Institute of Equipment Manufacturing, Haixi Institutes, Chinese Academy of Sciences, Quanzhou* 362000, *China*;
4. *Xiamen Key Laboratory of Frontier Electric Power Equipment and Intelligent Controly, Xiamen* 361024, *China*)

Abstract: Lidar data is widely used in 3D target detection tasks due to its geometric characteristics. Due

收稿日期：2023-02-17；**修订日期：**2023-03-21.
基金项目：福建省科技计划(No. 2021T3003)；泉州市科技计划(No. 2021C065L)；福建省科技厅自然科学基金(No. 2020J01285, No. 2022J05285)
Supported by Fujian Provincial Science and Technology Plan (No. 2021T3003); Quanzhou Science and Technology Plan (No. 2021C065L); Natural Science Foundation of Fujian Provincial Department of Science and Technology (No. 2020J01285, No. 2022J05285)
*通信联系人，E-mail：gjl@fjirsm.ac.cn

to the sparsity and irregularity of point cloud data, it is difficult to achieve the balance between the quality of feature extraction and the speed of reasoning. In this paper, a three-dimensional target detection algorithm based on body-column feature coding is proposed. Based on Pointpillars network, the Teacher-Student model framework is designed to distill the regression frame scale, increase distillation loss, optimize the training network model, and improve the quality of feature extraction. In order to further improve the model detection effect, the positioning guidance classification item is designed to increase the correlation between classification prediction and regression prediction, and improve the object recognition accuracy. The improvement of this network does not introduce additional network embedding. The experimental results of the algorithm on the KITTI dataset show that the average accuracy of the reference network in 3D mode is improved from 60.65% to 64.69%, and the average accuracy of the aerial view mode is improved from 67.74% to 70.24%. The model reasoning speed is 45 FPS, which meets the real-time requirements while improving the detection accuracy.

Key words: laser point cloud; 3D object detection; knowledge distillation; classification confidence

1 引　　言

激光点云是一种直观、灵活和存储效率高的三维数据表示方法,在三维视觉中已变得不可或缺。大规模激光雷达数据集的出现和端到端3D表示学习的巨大进步推动了基于点云的分割、生成和检测任务的发展。

不论是单阶段还是两阶段检测方法,点云的特征提取质量影响着算法的检测精度。Qi Charles R等[1]首次提出以端到端的方式通过多层感知来提取点的特征。随后,作者进一步提出PointNet++[2],以分层方式捕获局部结构,采用密度自适应采样和分组的方式提取点云特征。Point和Point++实现了直接对点云数据的处理和特征提取,被广泛应用到其他算法模型中。Zhou Y等人提出了VoxelNet[3],这是一种单级检测网络,可将点云划分为等间距的三维体素,并使用体素特征编码层进行处理,但是其采用了3D子流形稀疏卷积作为特征提取模块,致使网络推理速度相对较慢。Lang A H等人提出了Pointpillars[4]网络模型,提议将点云划分为几个体柱,将其转换为伪图像,可以使用2D卷积层进一步处理。此方法极大提高了网络模型的运算速度,使其能够满足自动驾驶实时性的要求,但其点云编码方式影响了特征提取的质量。Point R-CNN[5]和Pillar RCNN[6]是一种两阶段检测方法,首先基于原始点云生成自底向上的3D提案,然后对其进行细化以获得最终检测结果。随后,Fast point R-CNN[7]和PV-RCN[8]方法出现,利用体素表示和原始点云来发挥各自的优势。图神经网络是点云检测领域新兴的点云结构表示和特征提取方法。如为避免点云中心偏移和比例变化的3D-GCN[9],根据学习的特征生成自适应卷积核的AD-GCN[10]等。尽管点云的结构表示和特征提取方法多种多样,但复杂精细的结构设计可能会降低网络模型的推理速度。

早期的知识蒸馏方法主要是训练学生网络模仿教师网络预测的分类概率分布。近年来,以设计特定的知识提取方法用于提高目标检测的效率和准确性已成为一个新兴的热门话题。Chen等人首先提出将朴素预测和基于特征的知识提取方法应用于目标检测[11]。Wang等人证明前景对象和背景对象之间的不平衡阻碍了知识提取在目标检测中实现更好的性能[12]。为了解决这个问题,丰富的知识提取方法试图基于检测结果[13]、基于查询的注意力[14]和梯度[15]找到待提取区域。此外,最近还提出了提取教师与学生之间像素级和对象级关系的方法[16]。除了用于2D检测的知识蒸馏外,还引入了一些跨模态知识蒸馏,以将知识从基于RGB的教师检测网络转移到基于激光雷达的学生检测网络。然而,这些方法大多侧重于学生和教师在多模态框架中的选择,而基于纯点云数据三维检测的特定知识提取优化方法尚未得到很好的探索。

在Pointpillars的检测网络部分,其分类预测和回归框预测存在低相关性。低相关性主要是

由于在训练阶段分类预测和回归预测使用各自独立的目标函数进行训练，因此正样本的回归框预测和分类置信度之间会存在不对齐的情况[17]，影响置信度分数预测，最终影响网络模型的检测精度。

针对上述问题，本文做了如下工作：

(1)依据单阶段网络设计一组 Teacher-Student 模型框架对回归框尺度进行知识蒸馏。回归框尺度在数据类型上可以从连续表示转到离散表示，将教师网络的输出视为附加的回归框尺度目标，对教师网络和学生网络的回归框尺度输出进行连续值离散化，再做两组概率值拟合，制定蒸馏损失优化学生网络，提升物体的检测精度。

(2)设计定位引导分类项，将鸟瞰图视角下的正样本预测框与真实框的 IoU 值作为引导分数，以软化相应正样本硬类别标签，增加分类预测和回归预测的相关性，提高模型检测精度。定位引导分类项没有额外的网络嵌入，不影响网络模型的推理时间，使其保持高效性。

2　网络模型

2.1　总体框架

图 1 显示了本文的目标检测网络框架：(1)包含一个教师检测网络和一个学生检测网络，其中教师网络和学生网络的特征提取模块使用相同的网络结构。先训练教师网络模型，随后冻结教师网络参数，在训练学生网络模型时教师网络模型进行预加载，对输入学生网络的点云数据做增广，使学生网络探索更大的数据空间，并利用教师网络预测的软目标进行更好的优化。本文所用回归框蒸馏(Regression Box Distillation，RBD)策略作用于检测头的回归分支，而不是深层特征。(2)最终的检测网络是学生网络和其检测模块，为了增加分类预测与回归预测间的相关性而无需额外的网络嵌入，设计了定位引导分类(Positioning Guidance Classification，PGC)项作用于学生网络的分类预测，并改造分类损失函数。

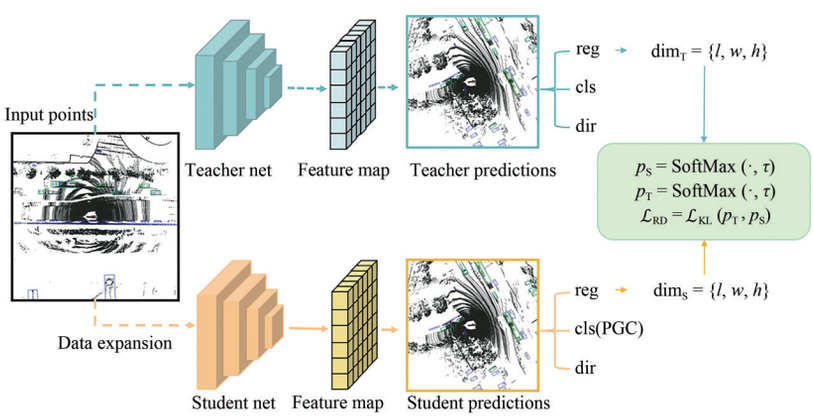

图 1　网络框图

Fig. 1　Network block diagram

2.2　点云编码与特征提取

网络的点云编码和特征提取依照 Pointpillars[4] 进行设置。将点云在 x-y 平面上设置柱体，每个非空柱构成一组子点云 $S_{x\in W, y\in H}=\{P_i, i=1, 2, \cdots, n_{x,y}\}$，其中每个点 P_i 用一个向量 (x, y, z, r) 表示，$n_{x,y}$ 是对应集合中的点的数量。将一帧点云编码成一个维度为 (D, P, N) 的稠密张量。对集合中的每个点用线性层＋BatchNorm＋ReLU 激活函数处理，生成维度为 (C, P, N) 的张量，其中 C 是特征通道。再通过每个点的体柱索引值重新放回到原来对应的体柱的 x, y 位置上生成 (C, H, W) 维度的伪图像。特征提取网络由下采样网络和上采样网络组成。下采样网络块表示为 ConvBlock $(C_{in}, C_{out}, S_d, N_b)$，其中 C 是特征通道数，S_d 是下采样因子，N_b 是每个网络块中卷积层的数量。上采样网络块表示为 DeconvBlock (C_{in}, C_{out}, S_u)，其中 S_u 是 2D 反卷积的上采样因子。

2.3　回归框蒸馏

与只传递语义知识的分类蒸馏不同，回归框蒸馏能够传递目标物体的位置和尺度信息，来自

教师模型的回归框尺度用作学生模型的额外回归目标，以帮助学生模型收敛到更好的优化点。此策略能够让学生网络模型的回归预测更为稳健，并实现更好的泛化能力，提升检测效果。

激光点云的三维目标检测中，网络模型的回归框预测输出为(x,y,z,l,w,h,θ)，共7个维度的数据。本方法中，只对预测输出的回归框尺度(l,w,h)进行蒸馏处理。在二维图像目标检测中，其边界框的表示通常有(x,y,w,h)（中心点坐标，长和宽）、$(x_{\min},y_{\min},x_{\max},y_{\max})$（回归框左上角点和右下角点）和$(t,b,l,r)$（采样点到回归框的上、下、左和右的距离）表示方式。其中(x,y,w,h)和$(x_{\min},y_{\min},x_{\max},y_{\max})$可以直接互相转换，这两种表示方法进一步用其采样点$(x_s,y_s)$和相匹配的真实框$(x_{gt},y_{gt},w_{gt},h_{gt})$计算出采样点到真实框上、下、左和右的距离，也就是$(t,b,l,r)$。不论是Anchor-Base类型的检测网络还是Anchor-Free类型的检测网络，以上回归框的3种表示形式可以依据其相匹配的真实框进行互相转换，从离散值转换到连续值，从连续值转换到离散值。但是在带有旋转角的三维目标检测回归框中，其中心点、回归框尺寸和旋转角互相独立，本文的回归框蒸馏其思想是针对连续域上回归的变量先离散化处理，最后进行概率拟合。

本文所提的回归框蒸馏方法选择对正样本回归框的尺度$\mathrm{Dim}=(l,w,h)$（回归框的长、宽、高）进行处理，(l,w,h)的每个变量的物理意义都是一致的，记每条边为e。设D为网络预测的3个回归框尺寸，分别由教师网络的D_{T}和学生网络的D_{S}表示，使用广义的SoftMax函数$S(\cdot,\tau)=\mathrm{SoftMax}(\cdot,\tau)$将$D_{\mathrm{T}}$和$D_{\mathrm{S}}$转换为概率表示$p_{\mathrm{T}}$和$p_{\mathrm{S}}$。当$\tau=1$时，它等价于原始的SoftMax函数；当$\tau>1$时，输入的参数会携带更多的信息。

$\mathcal{L}_{\mathrm{RD}}$用于衡量两组概率相似度的蒸馏损失，其定义如公式(1)所示：

$$\mathcal{L}_{\mathrm{RD}}=\mathcal{L}_{\mathrm{KL}}(p_{\mathrm{S}}^{\tau},p_{\mathrm{T}}^{\tau})=\mathcal{L}_{\mathrm{KL}}(\mathrm{SoftMax}(D_{\mathrm{S}},\tau),\mathrm{SoftMax}(D_{\mathrm{T}},\tau)),\quad(1)$$

其中：$\mathcal{L}_{\mathrm{KL}}$表示KL发散损失，$\tau$表示温度系数，S和T分别为教师网络和学生网络，$p$为概率值，$D$代表回归框尺度的集合。回归框尺寸$S$的3个维度的蒸馏可以化为公式(2)，其中$e$代表回归框的边：

$$\mathcal{L}_{\mathrm{RD}}(D_{\mathrm{S}},D_{\mathrm{T}})=\sum_{e\in D}\mathcal{L}_{\mathrm{RD}}^{e}.\quad(2)$$

2.4 定位引导分类

为了增加分类预测和回归预测的相关性，设计了定位引导分类项，过程示意图如图2所示。物体在点云的BEV空间中有一个关键优势是位置不重叠，因此在BEV空间中真实物体的定位效果和定位质量较好。将网络的正样本回归预测和真实框在BEV空间下做IoU值计算，将计算得到的IoU值作为引导分数，分配给正样本对应的硬类别标签（One-hot），分配后的硬类别标签变为软标签（Soft Label）。整个过程中，具有高IoU的正样本在分类时被自适应地向上加权，正样本的回归预测质量引导对应的类别标签。定位引导项$\boldsymbol{g}$定义如式(3)所示：

$$\boldsymbol{g}=i_{\mathrm{pos}}=\mathrm{IoU}_{\mathrm{pos}}^{\mathrm{bev}}=(\mathrm{IoU}(\mathrm{bbox}_{\mathrm{pred}},\mathrm{bbox}_{\mathrm{gt}}))_{\mathrm{pos}}^{\mathrm{bev}}=(\mathrm{IoU}(\mathrm{decode}(\mathrm{reg}_{\mathrm{pred}},\mathrm{anchor}),\mathrm{bbox}_{\mathrm{gt}}))_{\mathrm{pos}}^{\mathrm{bev}}.\quad(3)$$

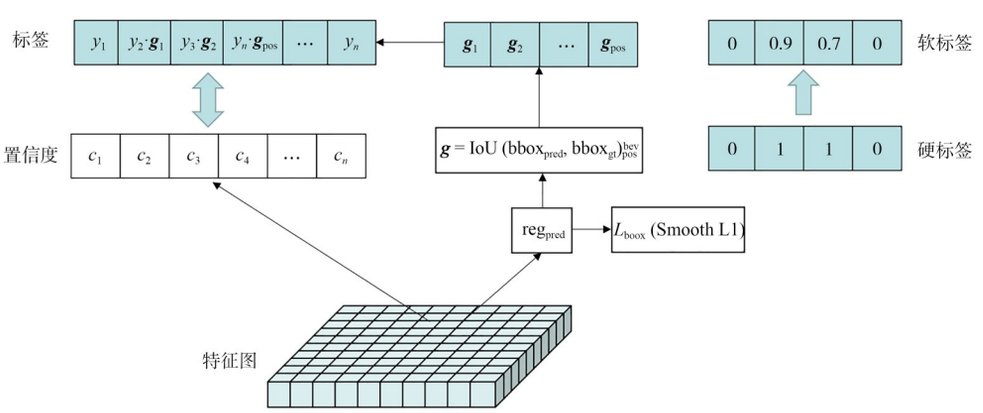

图2 定位引导分类

Fig. 2 Positioning guidance classification

目标监督值为：

$$f_{\text{pos}}=\text{label}_{\text{gt}}^{\text{one-hot}}\cdot g. \tag{4}$$

其中：i 是预测框和真实框的 IoU 值；pos 代表正样本；bev 是在 BEV 空间中边界框的维度表示；reg_{pred} 表示预测框偏差值，通过预测框偏差值与先验框 anchor 解码，得出预测框 $\text{bbox}_{\text{pred}}$，将其与样本所匹配的真实框 bbox_{gt} 做 BEV 视角下的 IoU 值计算，最终得到定位引导分类向量 $\boldsymbol{g}$；$\text{label}_{\text{gt}}^{\text{one-hot}}$ 是用 one-hot 向量表示的真实标签；f 是引入定位引导项的 soft label 表示形式的正样本标签。

目前 Pointpillars 网络的分类损失是焦点损失（Focal Loss，FL）损失函数，其一般形式如式（5）所示：

$$\begin{aligned}&\text{FL}(p)=-(1-p_t)^{\gamma}\log p_t,\\&p_t=\begin{cases}p, y=1\\1-p, y=0\end{cases},\end{aligned} \tag{5}$$

其中：$y\in\{0,1\}$ 是真实值的类别，$p\in[0,1]$ 是当真实标签 $y=1$ 时模型预测的类别概率，γ 是可调节焦点参数。焦点损失（FL）是由标准交叉熵 $-\log p_t$ 和一个调节因子 $(1-p_t)^{\gamma}$ 两部分组成。引入定位引导项 g 后，正样本真实标签从原本的 $y=0$ 代表负样本和 $y=1$ 代表正样本，变为 $f=0$ 代表负样本和 $0<f\leqslant 1$ 代表正样本的 soft label 表示形式。原本的焦点损失不能满足引入定位引导项后的计算要求，需要进行改造。焦点损失采用 sigmoid 算子 $\alpha(\cdot)$ 的多二进制分类实现多分类，把 sigmoid 的输出标记为 α，对焦点损失的两部分进行扩展，将交叉熵部分 $-\log p_t$ 扩展为完整的表示形式 $-((1-y)\log(1-\alpha)+y\log\alpha)$，代入定位引导项 g 后，交叉熵部分变为 $-((1-f)\log(1-\alpha)+f\log\alpha)$。比例因子部分 $(1-p_t)^{\gamma}$ 广义化扩展为估计 α 与其连续标签之间的距离绝对值，即表示为 $|f-\alpha|^{\beta}(\beta\geqslant 0)$，其中 $|\cdot|$ 保证了非负性。最后，将扩展的两部分组合起来，形成完整的分类损失函数，其定义如式（6）所示：

$$\begin{aligned}\mathcal{L}_{\text{cls}}(\alpha)=-|f-\alpha|^{\beta}((1-f)\log(1-\alpha)+\\f\log(1-\alpha)).\end{aligned} \tag{6}$$

2.5　网络总损失函数

本文的损失函数中，回归损失选用与 SECOND[18] 相同的回归损失。每个真实目标或者其先验框的 3D 表示由一个七维向量来表示：(x,y,z,l,w,h,θ)。其中 x、y、z 表示 3D 边界框的中心点坐标，l、w、h 分别表示 3D 边界框的长、宽和高，θ 表示 3D 边界框的朝向角。在边界框定位回归任务中，真实边界框和先验框之间的残差定义为：

$$\begin{cases}\Delta x=\dfrac{x^{\text{gt}}-x^{a}}{d^{a}},\Delta y=\dfrac{y^{\text{gt}}-y^{a}}{d^{a}},\Delta z=\dfrac{z^{\text{gt}}-z^{a}}{h^{a}}\\\Delta w=\log\dfrac{w^{\text{gt}}}{w^{a}},\Delta l=\log\dfrac{l^{\text{gt}}}{l^{a}},\Delta h=\log\dfrac{h^{\text{gt}}}{h^{a}}\\\Delta\theta=\sin(\theta^{\text{gt}}-\theta^{a})\end{cases}, \tag{7}$$

其中：x^{gt} 和 x^{a} 分别表示真实边界框和先验框。$d^{a}=\sqrt{(w^{a})^{2}+(l^{a})^{2}}$。边界框回归损失采用 Smooth L1 函数表示：

$$\mathcal{L}_{\text{box}}=\sum_{b\in(x,y,z,l,w,h,\theta)}\text{Smooth L1}(\Delta b). \tag{8}$$

采用 Softmax 分类损失用于学习目标的朝向，朝向损失记为 $\mathcal{L}_{\text{dir}}$。

对于分类损失，使用改造过的焦点损失函数，即：

$$\begin{aligned}\mathcal{L}_{\text{cls}}=-|f-\alpha|^{\beta}((1-f)\log(1-\alpha)+\\f\log(1-\alpha)).\end{aligned} \tag{9}$$

回归蒸馏损失为：

$$\mathcal{L}_{\text{RD}}(D_{\text{S}},D_{\text{T}})=\sum_{e\in D}\mathcal{L}_{\text{RD}}^{e}. \tag{10}$$

最终网络模型的总损失可表示为：

$$\mathcal{L}=\frac{1}{N_{\text{pos}}}(\lambda_0\mathcal{L}_{\text{box}}+\lambda_1\mathcal{L}_{\text{cls}}+\lambda_2\mathcal{L}_{\text{dir}}+\lambda_3\mathcal{L}_{\text{RD}}), \tag{11}$$

其中：N_{pos} 是正概率锚数；各项损失值的系数 $\lambda_0=2.0$，$\lambda_1=1.0$，$\lambda_2=0.2$，$\lambda_3=0.2$。

3　实验结果分析

使用三维目标检测数据集 KITTI 对本文算法进行验证，在 KITTI 数据集上进行多种算法对比实验、模型推理速度比较和消融实验。

3.1　实验环境和优化器设置

本文实验环境操作系统为 CentOS 7.6，硬件显卡型号为 NVIDIA GeForce RTX 2080 TI，Intel（R）Xeon（R）5220 CPU@2.20 GHz。深度学习框架为 Pytorch 1.7，Python 环境为 3.7，使用 CUDA 10.1 用于 GPU 加速。

网络训练设置 Batch Size 为 6，训练 80 个 epochs。采用 AdamW 优化器，使用 0.01 的衰减权重。使用周期性重启学习率调整策略，初始学

习率设置为0.001,最高学习率和最低学习率分别为10和0.000 1,训练期间循环次数为1次,学习率增加过程在整个循环中的比率为0.4。

3.2 数据集设置

在KITTI数据集上评估本文所提出的3D检测网络模型的性能。KITTI数据集中包含7 481个训练样本和7 518个测试样本。根据通用协议,将KITTI训练集分为3 712个样本的训练集和3 769个样本的验证集。对Car类、Cyclist类和Pedestrian类进行评估,其IoU阈值分别为0.7、0.5、0.5。此外,该基准在评估中有3个难度级别:简单、中等和困难,评估基于目标对象的遮挡和截断水平。按照官方KITTI评估指标,使用40个召回位置计算,以平均精度均值(mean Average Precision, mAP)评价检测结果。

在实验中将范围[0,69.12]、[−39.68,39.68]和[−3,1]米内的所有点分别沿着x、y和z轴体柱化,体柱的分辨率为[0.16,0.16,4],整个体柱网格大小为496×432×1。最大柱数(P)为16 000个,柱内最大点数(N)为100个。每个类的锚点由宽度、长度、高度和z中心来描述,具有0°和90°两个方向。在训练阶段,对输入的点云数据做数据增强处理,在x轴方向以0.5的概率随机翻转点云;将全局点云在z轴方向按照[$-\pi/4,\pi/4$]均匀分布的角度范围进行随机旋转,对全局点云按照[0.95,1.05]的范围进行随机缩放。

3.3 对比评估

为了评估所提模型方法的性能,在KITTI数据集与其他算法进行3D检测和BEV检测对比实验,结果如表1和表2所示。

表1 KITTI数据集不同算法3D检测精度($3D_{R40}$)对比

Tab. 1 Comparison of 3D detection accuracy ($3D_{R40}$) of different algorithms in KITTI dataset %

	算法	数据类型	Car (IoU=0.7)			Pedestrian (IoU=0.5)			Cyclist (IoU=0.5)		
			简单	中等	困难	简单	中等	困难	简单	中等	困难
两阶段	AVOD[19]	L+R	83.07	71.76	65.73	36.10	27.86	25.76	57.19	42.08	38.29
	PointRCNN[5]	L	86.96	75.64	70.70	49.43	41.78	38.63	73.93	59.60	53.59
	UberATG-MMF[20]	L+R	**88.40**	77.43	70.22	N/A	N/A	N/A	N/A	N/A	N/A
	Part-A2[21]	L	87.81	78.49	73.51	53.10	43.35	40.06	79.17	**63.52**	56.93
单阶段	SECOND[18]	L	83.34	72.55	65.82	51.07	42.56	37.29	70.51	53.85	46.90
	TANet[22]	L	84.39	75.94	68.82	**53.72**	44.34	40.49	75.70	59.44	52.53
	Associate-3Det[23]	L	85.99	77.40	70.53	N/A	N/A	N/A	N/A	N/A	N/A
	Point-GNN[24]	L	88.33	78.47	72.29	51.92	43.77	40.14	78.60	63.48	57.08
	Ours	L	88.15	**78.95**	**74.97**	52.77	**46.09**	**41.09**	**81.66**	61.31	**57.21**

注:加粗字体为每项最优值,L代表激光点云,R代表图像

表2 KITTI数据集不同算法BEV检测精度(BEV_{R40})对比

Tab. 2 Comparison of BEV detection accuracy (BEV_{R40}) of different algorithms in KITTI dataset %

	算法	数据类型	Car (IoU=0.7)			Pedestrian (IoU=0.5)			Cyclist (IoU=0.5)		
			简单	中等	困难	简单	中等	困难	简单	中等	困难
两阶段	AVOD[19]	L+R	89.75	84.95	78.32	42.58	33.57	30.14	64.11	48.15	42.37
	PointRCNN[5]	L	92.13	87.39	82.72	N/A	N/A	N/A	N/A	N/A	N/A
	UberATG-MMF[20]	L+R	**93.67**	88.21	81.99	N/A	N/A	N/A	N/A	N/A	N/A
	Part-A2[21]	L	91.70	87.79	84.41	59.04	49.81	45.92	83.43	**68.73**	61.85
单阶段	SECOND[18]	L	89.39	83.77	78.59	55.10	46.27	44.76	73.67	56.04	48.78
	TANet[22]	L	91.58	86.54	81.19	**60.85**	51.38	**47.54**	79.16	63.77	56.21
	Associate-3Det[23]	L	91.40	88.09	82.96	N/A	N/A	N/A	N/A	N/A	N/A
	Point-GNN[24]	L	93.11	**89.17**	83.90	55.36	47.07	44.61	81.17	67.28	59.67
	Ours	L	93.09	88.86	**84.50**	58.46	**51.88**	47.43	**84.10**	64.03	**59.82**

注:加粗字体为每项最优值,L代表激光点云,R代表图像

在 3D 检测对比中，与经典的单阶段检测方法 TANet[22] 和 SECOND[18] 相比，在中等难度级别上，Car类和Cyclist类分别高 3.01%、1.87% 和 6.4%、7.46%；与先进的单阶段检测方法 Point-GNN[24] 相比，Car 类和 Pedsetrian 类在中等难度级别分别高了 0.48% 和 2.32%。与两阶段检测方法 PointRCNN[5] 相比，3 种类别的中等难度分别高出 3.31%、4.31% 和 1.71%；与 Part-A2[21] 相比，Car类中等难度高出 0.46%，本文模型优于多数两阶段模型方法。在 BEV 检测中，本文模型与 TANet[22] 和 SECOND[18] 相比，在 Car 类中等难度分别高出 2.32% 和 5.09%。结果显示，本文的模型在所有 3 个难度级别的 3D 和 BEV 检测中与其他先进方法相比具有竞争力，验证了本文方法的有效性。回归框蒸馏能够传递目标物体的位置和尺度信息，帮助网络收敛到更好的优化点，使回归模型更为稳健；定位引导分类建立了预测框和分类预测间的相关性，提升模型分类效果，最终提升了模型检测精度。

本文方法采用体柱特征编码的方式，点云经过编码后，其分辨率显著低于体素特征编码和基于点的特征形式，所以其小目标如 Pedestrian 类的检测精度会低于部分基于体素特征编码和基于点的模型方法。

本文的回归框尺度蒸馏中引入温度系数 τ，表 3 中显示了 KITTI 数据集中不同温度系数下的蒸馏结果，在温度系数 $\tau=2$ 时模型获得最好的效果。

表 3　蒸馏中温度系数在 Mod_{R40} 模式下对模型探测精度的影响

Tab. 3　Influence of temperature coefficient on model detection accuracy in distillation under Mod_{R40}　　%

τ	Car	Pedestrian	Cyclist
教师网络	74.31	41.92	51.92
2	78.95	56.09	61.31
3	76.33	44.56	59.89
4	75.22	43.67	58.02
5	73.87	41.03	46.63

为了验证本文方法的检测效率，选择主流算法进行模型推理速度对比，结果如图 3 所示。在模型推理速度方面，本文模型方法是两阶段网络 AVOD[19] 和 PointRCNN[5] 的 3～4 倍；与单阶段网络

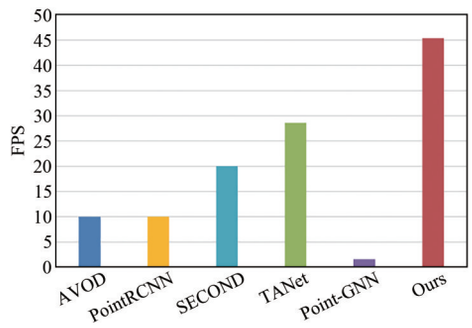

图 3　网络模型推理速度对比

Fig. 3　Comparison of network model reasoning speed

SECOND[18] 和 TANet[22] 相比，推理速度提高了大约 2 倍，达到 45 FPS。虽然检测精度与 Point-GNN[24] 基本持平，但 Point-GNN 由于需要对点云构建“图”结构以及图卷积等操作，需要消耗大量算力，因此模型推理速度慢了许多，不符合实时性要求。与单阶段网络相比，本文网络模型具有检测精度优势；与两阶段网络相比，本文网络模型能够在检测精度上持平，但在推理速度上远高于两阶段网络。

如图 4 所示，将本文针对点云的蒸馏策略与其他蒸馏方法如 Zagoruyko[25]、Zheng[26]、Tian[27]、Heo 等[28]、Zhang[16] 等方法对比，以 Pointpillars 为基准网络，在 KITTI 数据集上进行 Car 类和 3 种难度级别的 3D 检测。可以观察到本文方法在 Car 类平均检测精度方面比所列蒸馏方法都要高。如图 5 所示，在 3D 检测难度级别为中等和困难难度级别中，本文的蒸馏策略比上述蒸馏方法效果提升更明显。

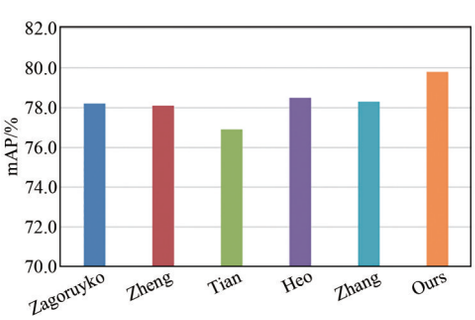

图 4　Car类的平均精度均值对比

Fig. 4　Comparison of average precision of car class

3.4　消融实验

消融实验可以评估本文所提方法各个模块对检测结果的贡献。所有评估测试都在 KITTI

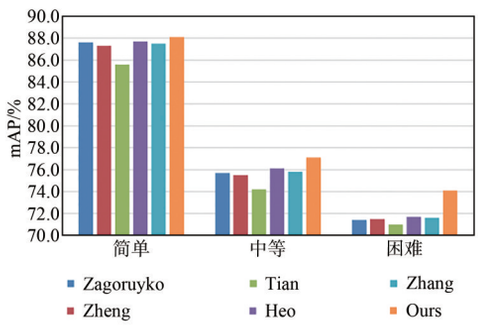

图 5 3种难度级别的平均精度均值对比

Fig. 5 Comparison of average accuracy of three difficulty levels

训练集上进行训练,在验证集上进行评估。基准网络为Pointpillars[4]网络模型,消融实验的设置以单独和总体结合的形式展示本文方法的有效性。其中“回归框蒸馏”记作RBD,“定位引导分类”记作PGC,使用40个召回位置计算平均精度均值(mAP),结果如表4所示。

只增加RBD方法时,网络模型在3D检测中简单、中等、困难3类的平均均值精度提升了4.48%、2.27%和1.49%,表明给预测框的尺度增加额外的回归目标可以更好地优化模型,同时教师网络产生的软目标携带更多的信息,让学生网络在训练过程中学习到更多的信息熵,提升模型特征提取的质量,从而提高物体检测精度。只增加PGC方法时,3D检测中简单、中等、困难3类的平均均值精度提升了2.5%、1.0%和1.16%。定位引导分类项增加了回归预测和分类预测之间的相关性,具有高IoU的正样本在分类时被自适应地向上加权。最终综合评估,本文所提出的两种方法组合使用时,其检测效果提升最大。

表 4 回归框蒸馏和定位引导分类在KITTI数据集上的消融实验

Tab. 4 Ablation experiments of regression frame distillation and location-guided classification in KITTI dataset %

	PGC	RBD	Car (IoU=0.7)				Pedestrian (IoU=0.5)				Cyclist (IoU=0.5)			
			简单	中等	困难	mAP	简单	中等	困难	mAP	简单	中等	困难	mAP
3D	×	×	82.58	74.31	68.99	75.29	51.45	41.92	38.89	44.08	77.10	58.65	51.92	62.55
	√	×	85.04	76.29	72.04	77.79	51.50	44.11	39.65	45.08	77.60	60.30	53.22	63.71
	×	√	88.08	77.10	74.12	79.77	52.28	45.79	40.97	46.35	77.58	59.17	55.37	64.04
	√	√	88.15	78.95	74.97	80.69	52.77	46.09	41.09	46.65	81.66	61.31	57.21	66.73
BEV	×	×	90.07	86.56	82.81	86.48	57.60	48.64	45.78	50.67	79.90	62.73	55.58	66.07
	√	×	91.37	87.52	82.95	88.28	57.64	50.73	46.27	51.45	82.05	63.76	60.62	70.81
	×	√	92.51	88.51	83.93	88.32	58.89	52.67	47.85	53.14	83.33	64.29	60.51	69.38
	√	√	93.09	88.86	84.50	88.82	58.46	51.88	47.43	52.59	84.10	64.03	59.82	69.32

4 结 论

本文受图像目标检测中知识蒸馏思想的启发,针对激光点云数据的3D目标检测任务设计了预测框的尺度作为约束训练的蒸馏方法。此方法可以为检测网络在训练中提供更多的信息熵,使网络模型拥有更好的泛化能力,提升特征提取质量,提高模型检测效果。针对Pointpillars网络中回归预测和分类预测间的低相关性,设计了定位引导分类项,同时改造了分类损失函数,将正样本回归预测质量引导类别标签,以提升检测效果。在KITTI数据集中,本文算法模型比基准网络在Car类提升了5.4%mAP,在一众算法模型中具有竞争力。

参 考 文 献:

[1] QI CHARLES R, HAO S, MO K C, *et al*. PointNet: deep learning on point sets for 3D classification and segmentation [C]//*Proceedings of* 2017 *IEEE Conference on Computer Vision and Pattern Recognition*. Honolulu: IEEE, 2017: 77-85.

[2] 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* 31*st International Conference on Neural Information Processing Systems*. Long Beach: Curran Associates Inc, 2017.

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

[4] LANG A H, VORA S, CAESAR H, *et al*. PointPillars: fast encoders for object detection from point clouds [C]//*Proceedings of* 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Long Beach: IEEE, 2019: 12689-12697.

[5] SHI S S, WANG X G, LI H S. PointRCNN: 3D object proposal generation and detection from point cloud [C]//*Proceedings of* 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Long Beach: IEEE, 2019: 770-779.

[6] 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.
LI R L, WU C, ZHU M. 3D object detection in voxelized point cloud scene [J]. *Chinese Journal of Liquid Crystals and Displays*, 2022, 37(10): 1355-1363. (in Chinese)

[7] CHEN Y L, LIU S, SHEN X Y, *et al*. Fast point R-CNN [C]//*Proceedings of* 2019 *IEEE/CVF International Conference on Computer Vision*. Seoul: IEEE, 2019: 9774-9783.

[8] SHI S S, GUO C X, JIANG L, *et al*. PV-RCNN: Point-voxel feature set abstraction for 3D object detection [C]//*Proceedings of* 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Seattle: IEEE, 2020: 10526-10535.

[9] LIN Z H, HUANG S Y, WANG Y C F. Convolution in the cloud: learning deformable kernels in 3D graph convolution networks for point cloud analysis [C]//*Proceedings of* 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Seattle: IEEE, 2020: 1797-1806.

[10] ZHOU H R, FENG Y D, FANG M S, *et al*. Adaptive graph convolution for point cloud analysis [C]//*Proceedings of* 2021 *IEEE/CVF International Conference on Computer Vision*. Montreal: IEEE, 2021: 4945-4954.

[11] CHEN G B, CHOI W, YU X, *et al*. Learning efficient object detection models with knowledge distillation [C]//*Proceedings of the* 31*st International Conference on Neural Information Processing Systems*. Long Beach: Curran Associates Inc, 2017.

[12] WANG T, YUAN L, ZHANG X P, *et al*. Distilling object detectors with fine-grained feature imitation [C]//*Proceedings of* 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Long Beach: IEEE, 2019: 4928-4937.

[13] DAI X, JIANG Z R, WU Z, *et al*. General instance distillation for object detection [C]//*Proceedings of* 2021 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Nashville: IEEE, 2021: 7838-7847.

[14] KANG Z, ZHANG P, ZHANG X, *et al*. Instance-conditional knowledge distillation for object detection [C]//*Proceedings of the* 35*th Conference on Neural Information Processing Systems*. Online: NIPS 2021: 16468-16480.

[15] GOU J Y, HAN K, WANG Y H, *et al*. Distilling object detectors via decoupled features [C]//*Proceedings of* 2021 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Nashville: IEEE, 2021: 2154-2164.

[16] ZHANG L F, MA K S. Improve object detection with feature-based knowledge distillation: towards accurate and efficient detectors [C]//*Proceedings of the* 9*th International Conference on Learning Representations*. Online: ICCV, 2021.

[17] JIANG B R, LOU R X, MAO J Y, *et al*. Acquisition of localization confidence for accurate object detection [C]//*Proceedings of the* 15*th European Conference on Computer Vision* (*ECCV*). Munich: Springer, 2018: 816-832.

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

[19] KU J, MOZIFIAN M, LEE J, *et al*. Joint 3D proposal generation and object detection from view aggregation [C]//*Proceedings of* 2018 *IEEE/RSJ International Conference on Intelligent Robots and Systems* (*IROS*). Madrid: IEEE, 2018: 1-8.

[20] LIANG M, YANG B, CHEN Y, *et al*. Multi-task multi-sensor fusion for 3D object detection [C]//*Proceedings of* 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Long Beach: IEEE, 2019: 7337-7345.

[21] SHI S S, WANG Z, SHI J P, *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*, 2021, 43(8): 2647-2664.

[22] LIU Z, ZHAO X, HUANG T T, *et al*. TANet: robust 3D object detection from point clouds with triple attention [C]//*Proceedings of the* 34*th AAAI Conference on Artificial Intelligence*. New York: AAAI, 2020: 11677-11684.

[23] DU L, YE X Q, TAN X, *et al*. Associate-3Ddet: perceptual-to-conceptual association for 3D point cloud object detection [C]//*Proceedings of* 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Seattle: IEEE, 2020: 13326-13335.

[24] SHI W J, RAJKUMAR R. Point-GNN: graph neural network for 3D object detection in a point cloud [C]//*Proceedings of* 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Seattle: IEEE, 2020: 1708-1716.

[25] ZAGORUYKO S, KOMODAKIS N. Paying more attention to attention: improving the performance of convolutional neural networks via attention transfer [C]//*Proceedings of the* 5*th International Conference on Learning Representations*. Toulon: OpenReview.net, 2017.

[26] ZHENG W, TANG W L, JIANG L, *et al*. SE-SSD: self-ensembling single-stage object detector from point cloud [C]//*Proceedings of* 2021 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Nashville: IEE, 2021: 14489-14498.

[27] TIAN Y L, KRISHNAN D, ISOLA P. Contrastive representation distillation [C]//*Proceedings of the* 8*th International Conference on Learning Representations*. Addis Ababa: OpenReview.net, 2020.

[28] HEO B, KIM J, YUN S, *et al*. A comprehensive overhaul of feature distillation [C]//*Proceedings of* 2019 *IEEE/CVF International Conference on Computer Vision*. Seoul: IEEE, 2019: 1921-1930.

作者简介：

赵　晶，女，博士，教授，2017年于中国台湾元智大学获得博士学位，主要从事人工智能、电气智能化等方面的研究。E-mail：ztulipwork@139.com

郭杰龙，男，硕士，工程师，2015年于中南民族大学获得硕士学位，主要从事机器学习、三维几何建模方面的研究。E-mail：gjl@fjirsm.ac.cn