DOI:10.11918/202511097

虚拟点增强与双向跨注意力融合的3D目标检测

马保全[1], 王韵涵[1,2],于寅虎[1],郝博文[1],李佳伟[1],杨 洋[3]

(1.华北计算机系统工程研究所,北京 100000; 2.山东大学 网络空间安全学院(研究院),
山东 青岛 266237; 3.北京航空航天大学 网络空间安全学院,北京 100191)

摘 要: 为解决激光雷达(LiDAR)点云稀疏导致远距离目标几何信息缺失,以及现有模态融合方法中存在的投影噪声干扰和特征交互非标准化问题,提出了一种端到端的多模态三维检测网络——VAF-Net。首先,构建了抗噪的虚拟点增强框架,利用图像生成虚拟点以稠密化原始点云,并采用VirConvNet作为骨干网络,在提取全局几何特征的同时有效过滤由虚拟点引入的噪声。其次,设计了色彩语义特征提取支路,通过语义提升模块精准提取与三维空间对齐的图像局部纹理与上下文语义。最后,提出了基于标准Transformer原则的双向跨注意力融合模块,通过对称的双流架构和门控机制,实现了LiDAR几何特征与图像语义特征的深度、稳定且双向的交互。在KITTI三维目标检测基准上的实验表明,VAF-Net取得了当前最优性能。特别是在困难样本检测中,其3D平均精度达到82.74%,显著优于Voxel-RCNN等主流方法。研究结果表明,VAF-Net通过标准化的双向融合架构,有效实现了多模态信息的互补增强,大幅提升了复杂场景下的检测鲁棒性。

关键词: 三维目标检测;多模态融合;交叉注意力机制;自动驾驶;虚拟点

中图分类号: TP391.41 **文献标志码:** A **文章编号:** 0367-6234(2026)08-0164-09

Virtual point-enhanced 3D object detection with bidirectional cross-attention fusion

Ma Baoquan[1], Wang Yunhan[1,2], Yu Yinhu[1], Hao Bowen[1], Li Jiawei[1], Yang Yang[3]

(1. North China Computer System Engineering Research Institute, Beijing 100000, China; 2. School of Cyber Science and Technology, Shandong University, Qingdao 266237, Shandong, China; 3. School of Cyber Science and Technology, Beihang University, Beijing 100191, China)

Abstract: To address the loss of geometric information of distant objects caused by the sparsity of light detection and ranging (LiDAR) point clouds and the problems of projection noise interference and non-standardized feature interaction in existing modal fusion methods, this paper proposed an end-to-end multi-modal three-dimensional detection network, VAF-Net. First, this paper constructed a noise-resistant virtual point enhancement framework, which utilized images to generate virtual points to densify the point clouds and employed VirConvNet as the backbone network to extract global geometric features while effectively filtering the noise introduced by the virtual points. Second, this paper designed a color semantic feature extraction branch to precisely extract local textures and contextual semantics of images aligned with the three-dimensional space through a semantic lifting module. Finally, this paper proposed a bidirectional cross-attention fusion module based on standard Transformer principles. Through a symmetric dual-stream architecture and a gating mechanism, it achieved deep, stable, and bidirectional interaction between LiDAR geometric features and image semantic features. Experiments on the KITTI three-dimensional object detection benchmark indicate that VAF-Net achieves state-of-the-art performance. Particularly in the detection of hard samples, its 3D average precision reaches 82.74%, significantly outperforming mainstream methods such as Voxel-RCNN. The research results demonstrate that VAF-Net effectively realizes the complementary enhancement of multi-modal information through the standardized bidirectional fusion architecture and significantly improves the detection robustness in complex scenarios.

Keywords: three-dimensional object detection; multi-modal fusion; cross-attention mechanism; autonomous driving; virtual point

三维目标检测是自动驾驶[1]、机器人导航[2],以及增强现实等领域[3]的关键感知技术,可为下游规划与控制模块提供环境中的精确三维信息[4-5]。激光雷达(light detection and ranging,LiDAR)因其能

收稿日期: 2025-11-29;**录用日期:** 2026-01-08;**网络首发日期:** 2026-05-27
网络首发地址: https://link.cnki.net/urlid/23.1235.t.20260527.1608.002
基金项目: 国家自然科学基金(U2241213)
作者简介: 马保全(1982—),男,博士,正高级工程师
通信作者: 马保全,baoquan.ma@buaa.edu.cn;王韵涵,i892991879@163.com

直接获取场景的几何结构且对光照变化不敏感,已成为 3D 检测的主流传感器[6-7]。然而,LiDAR 点云的固有稀疏性,尤其是在远距离、小目标或遮挡场景下,会导致目标几何信息不完整,从而制约了检测性能。

相机可捕捉高分辨率图像,蕴含丰富的语义信息,但极易受环境变化的影响,如雨雪雾天气下性能严重下降。因此,融合 LiDAR 的几何精度与相机的语义密度[8],被广泛认为是突破单模态感知瓶颈,实现鲁棒 3D 感知的有效途径。

尽管多模态融合的理念已得到广泛认可,但实现高效的特征交互仍面临两大核心挑战。

1)在于如何将图像的密集语义信息有效地注入稀疏的 LiDAR 几何结构中,以实现从全局场景理解到局部点级语义的全面增强。一种新兴方法是通过单目深度估计等技术生成虚拟点[9-11]以稠密化场景。然而,这种方式在引入丰富信息的同时,也必然会引入由深度估计不精确性导致的噪声。因此,如何设计一个强大的特征提取网络,使其既能利用虚拟点带来的信息增益,又能鲁棒地过滤其伴随的噪声,是实现有效增强的关键。

2)在于如何设计一个架构上标准化的跨模态注意力模块,以确保 LiDAR 几何特征与图像语义特征之间能够进行稳定、高效且双向的深度交互。尽管基于 Transformer 的注意力机制已成为主流[12-13],但许多现有工作倾向于设计复杂的非标准融合模块[8],这些架构多为非对称或不规范的设计,易破坏标准 Transformer 架构的内在平衡性,反而成为限制模型性能的瓶颈。基于此,本文提出一个值得探究的学术命题:遵循标准化设计原则、结构更为精简的融合模块,是否具备强化跨模态特征对等交互效能的潜在能力。

为应对上述挑战,首先,提出了一个全新的、具有鲁棒注意力的虚拟点增强 3D 目标检测框架。该框架首先采用强大的 VirConvNet 骨干网络在全局层面高效提取鲁棒的三维特征,有效缓解了虚拟点引入的噪声问题;继而通过逐点增强策略在局部层面提升细节语义,应对了第 1 个挑战。其次,设计了一个名为双向跨注意力融合(bidirectional cross-attention fusion, BCAF)的新型融合模块。该模块严格遵循标准 Transformer 设计原则,通过独立的投影、多头注意力,以及残差连接,确保了跨模态特征交互的稳定、高效与双向性,精准地解决了第 2 个挑战。最后,在 KITTI 3D 目标检测基准上,VAF-Net 取得了当前最优的性能,尤其是在处理困难样本时性能提升显著,验证了本文所提方法的有效性。

1　相关工作

1.1　基于 LiDAR 的 3D 目标检测

近年来,基于 LiDAR 的 3D 目标检测技术得到了研究人员的广泛研究。早期的方法通常将点云投影到二维鸟瞰图[14]或距离视图中进行处理[15]。当前研究中,基于体素的稀疏卷积[6,16-18]和基于点的直接特征提取方法[7,19-20]已成为构建高效检测框架的主流范式。然而,纯 LiDAR 方法在处理远距离物体时,因点云的固有稀疏性而性能受限,难以准确感知其几何形状。本文旨在通过引入 RGB 图像信息来缓解这一问题。

1.2　基于虚拟点的融合方法

作为一种有效提升远处物体感知的策略,基于虚拟点的方法[21-23]近期备受关注。这类方法通过深度补全网络,利用图像的密集像素生成大量虚拟点,极大增强了场景的几何完整性。其代表性工作,如 SFD[21],展示了利用虚拟点在高精度 3D 检测中的巨大潜力。然而,增加虚拟点所带来的误差与扰动,极大影响了对于小目标、高遮挡情况的目标识别,本文正是基于上述观察,引入二阶段的逐点语义增强,以提升模型对于高难度检测目标的准确识别。

1.3　多模态 3D 目标检测

多模态融合的核心在于如何有效交互和整合不同来源的信息。一类方法是独立编码两种模态的信息,然后在候选区域[24-25]或 BEV 特征层面[26]进行融合;另一类方法则在点级别进行精细化融合。例如,多阶段巩固融合(multi-phase consolidated fusion, MPCF)提出了一种逐点的综合融合框架,旨在为每个 LiDAR 点匹配并融合更丰富的图像上下文特征。这类方法增强了原始点云的特征表达,但在如何高效处理由虚拟点技术带来的海量且含噪声的数据方面,仍存在探索空间。本文正是基于上述观察,旨在结合虚拟点方法在几何补全上的优势,并借鉴精细化逐点融合的思想,提出一种新的优化框架,以实现对稠密且含噪声的多模态数据更高效、更鲁棒的特征学习。

2　方　法

本文提出了 VAF-Net,一个全新的端到端多模态检测框架,旨在通过增强的输入表示和鲁棒的特

征交互来提升 3D 检测精度。其整体架构如图 1 所示,核心在于本文设计的一个并行的双流主干虚拟点增强网络,以及一个为实现深度特征交互而设计的 BCAF 模块。

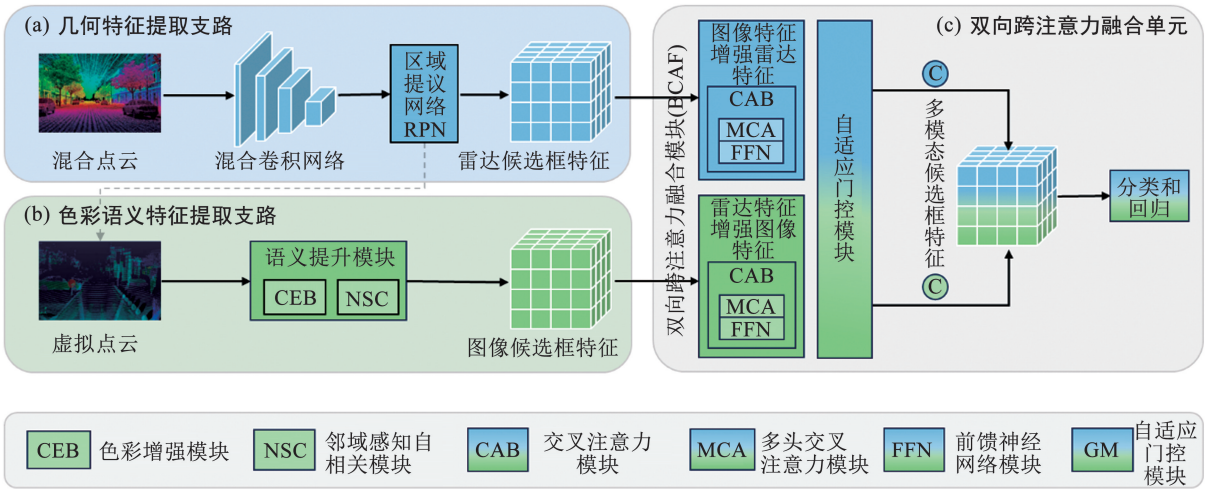


图 1 VAF-Net 流程图

Fig. 1 Flowchart of VAF-Net

2.1 VAF-Net 目标检测框架

为应对 LiDAR 点云的稀疏性问题,本文的架构在输入端融合了原始 LiDAR 点和由图像生成的虚拟点,形成高密度的混合点云[27]表示。混合点云数据通过 LiDAR 支路(图 1(a))提取含有丰富几何信息的感兴趣区域(region of interest, RoI)特征;同时虚拟点云通过图像支路(图 1(b))提取 LiDAR 支路的 RoI 特征所对应的色彩语义信息,生成图像 RoI 特征。经过投影映射,这两个支路实现了整体几何信息与局部语义信息的提取。最终通过 BCAF 模块(图 1(c))进行高效特征融合和目标检测,从而实现多模态数据的有效对齐和高精度目标检测。

VAF-Net 的流程架构具备显著创新性,其核心设计如下。

1)几何特征提取支路。该支路旨在从稠密但含有噪声的混合点云中提取鲁棒的三维空间表征。本文创新性地将强大的 VirConvNet[28] 骨干网络应用于此双流架构中。首先,混合点云经过体素特征编码得到 N 个非空体素的 3D 索引向量 $\boldsymbol{H}\in\mathbf{R}^{N\times 3}$ 及其特征向量 $\boldsymbol{X}\in\mathbf{R}^{N\times C_{\text{in}}}$(其中 C_{in} 为输入特征通道数)。其次,通过混合卷积网络(VirConvNet[28])的双分支结构并行地进行特征编码:3D 分支通过 3D 子流形卷积核 $\mathrm{K}^{3\mathrm{D}}(\cdot)$,在三维邻域内完成信息聚合,高效捕捉自身局部几何结构,生成 3D 几何特征为

$$\hat{\boldsymbol{X}}_i=\mathrm{ReLU}(\mathrm{K}^{3\mathrm{D}}(\boldsymbol{X}_i,\boldsymbol{X}_i^{(f_1)},\cdots,\boldsymbol{X}_i^{(f_j)})) \tag{1}$$

式中:$\boldsymbol{X}_i^{(f_1)},\cdots,\boldsymbol{X}_i^{(f_j)}$ 为由 $\boldsymbol{H}$ 生成的邻域点云特征,ReLU 为非线性激活函数。

2D 分支将 3D 体素索引 $\boldsymbol{H}$ 投影到 2D 平面上,得到 2D 索引 $\hat{\boldsymbol{H}}=P(\tau^{-1}(\zeta(\boldsymbol{H})))$,其中:$\zeta(\cdot)$ 为体素到点坐标的转换,$\tau^{-1}(\cdot)$ 为数据增强的逆变换,$P(\cdot)$ 为最终的相机投影。基于这些 2D 索引,该分支在图像的二维邻域内进行 2D 子流形卷积以识别噪声模式,并计算出噪声感知特征$\widetilde{\boldsymbol{X}_i}$为

$$\widetilde{\boldsymbol{X}_i}=\mathrm{ReLU}(\mathrm{K}^{2\mathrm{D}}(\boldsymbol{X}_i,\widetilde{\boldsymbol{X}_i^{(f_1)}},\cdots,\widetilde{\boldsymbol{X}_i^{(f_k)}})) \tag{2}$$

式中:$\widetilde{\boldsymbol{X}_i^{(f_1)}},\cdots,\widetilde{\boldsymbol{X}_i^{(f_k)}}$ 为由 $\hat{\boldsymbol{H}}$ 生成的 2D 邻域特征,$\mathrm{K}^{2\mathrm{D}}(\cdot)$ 为 2D 子流形卷积核。

然后,模块将 3D 分支提取的几何特征$\hat{\boldsymbol{X}}_i$ 与 2D 分支提取的噪声感知特征$\widetilde{\boldsymbol{X}_i}$进行拼接,即

$$\boldsymbol{Y}=\mathrm{Concat}(\boldsymbol{X}_i,\boldsymbol{X}_i')\in\mathbf{R}^{N\times C_{\text{out}}} \tag{3}$$

式中 C_{out}为输出特征通道数。

最后,经过区域提议网络(regional proposal network,RPN)在特征图的每个空间位置上生成预设尺度与长宽比的锚点(anchor)作为候选区域,通过全卷积层对每个锚点进行二分类和边界框回归,其中分类采用 Sigmoid 函数输出概率,回归则通过 Smooth L1 损失优化。最终输出每个锚点物体的置信度得分及对应的边界框坐标,这些候选区域将作为后续检测模块的输入,实现端到端的目标检测。

本文选择 VirConvNet 是因为其核心的抗噪声子流形卷积和随机体素舍弃的设计,能够有效过滤由不精确深度估计引入的噪声点,并高效编码混合点云的几何结构。经过骨干网络和区域提议网络处理后,该支路为每个候选框输出可靠的 LiDAR RoI 特征。

2)色彩语义特征提取支路。该支路的目标是使用2D 图像中与三维空间精确对齐的语义信息(如颜色、纹理)来局部增强特征。本文选择色彩增强模块(color enhanced block,CEB),旨在从离散的虚拟点中,有效提取具有上下文信息的局部语义特征。首先,先将图像转换为虚拟点云,根据 3D RoI 裁剪特征相关点云 $\hat{P}_i$。其次,输入特征$\hat{P}_i$ 经过 4 级多层感知机(multi-layer perceptron,MLP)($f_{\theta_{11}}$ ~$f_{\theta_{41}}$)逐层提取,生成分层特征$\{F_i^{(1)},F_i^{(2)},F_i^{(3)},F_i^{(4)}\}$。同时,计算当前点 P_i 与邻域点 P_j 的特征差异 $\Delta P'_{ij}$,通过 MLP $f_{\theta_{22}}$转化为调制权重 $\boldsymbol{W}_{ij}$,该权重对邻域中层特征 $F_j^{(2)}$ 进行逐元素加权得到 $M_{ij}=F_j^{(2)}\odot\boldsymbol{W}_{ij}$。所有邻域调制特征经聚合 MLP $f_{\theta_{23}}$处理后生成上下文特征 $F_i^{(\mathrm{agg})}$,最终与原始特征及各层特征拼接输出,以实现几何与色彩特征的多尺度融合,即

$$F_i^{(\mathrm{final})}=\mathrm{Concat}(\hat{P}_i,F_i^{(1)},F_i^{(2)},F_i^{(3)},F_i^{(4)},F_i^{(\mathrm{agg})}) \tag{4}$$

然后,本文引入邻域感知自相关模块(neighborhood-aware self-correlation,NSC)对融合后的多尺度特征进行再提炼,它通过自适应计算并应用注意力权重,在通道与空间维度上对特征图进行校准,从而聚焦于最重要的语义信息。最后,NSC 能够生成上下文丰富的高密度图像 RoI 特征,为后续的跨模态融合提供高质量的语义基础。

3)融合模块(fusion module)。BCAF 模块旨在实现激光雷达几何信息与图像语义信息的深度、双向融合。以图像特征增强雷达特征为例,利用图像特征作为上下文,对 LiDAR 特征进行精炼,输出增强的 LiDAR 特征 F'_{lidar}。其中,F_{lidar} 表示查询(query),F_{image}表示键(key)和值(value)。经过多头交叉注意力(multi-head cross attention,MCA)计算后,自适应门控模块(gating module,GM)会拼接两路增强后的特征,并通过共享 MLP 学习通道权重。此处本文采用 Sigmoid 函数而非 Softmax,是因为激光雷达与图像特征并非互斥关系,Sigmoid 允许模型独立地对每种模态的特征通道进行强调或抑制,从而保留两者的互补信息,即

$$\boldsymbol{F}''_{\mathrm{lidar}}=\boldsymbol{F}'_{\mathrm{lidar}}\odot\mathrm{Sigmoid}(\mathrm{MLP}(\mathrm{Concat}(\boldsymbol{F}'_{\mathrm{lidar}},\boldsymbol{F}'_{\mathrm{image}}))) \tag{5}$$

最终,这些高度精炼的特征 $\boldsymbol{F}_{\mathrm{fusion}}=\mathrm{Concat}(\boldsymbol{F}''_{\mathrm{lidar}},\boldsymbol{F}''_{\mathrm{image}})$被送入检测头,用于生成精确的类别得分与位置回归参数。

2.2　双向跨注意力融合(bidirectional cross-attention fusion, BCAF)模块

针对现有融合方法结构复杂、非标准化的问题,本文设计了遵循标准 Transformer 原则的 BCAF 模块,以实现 LiDAR 与图像特征间稳定、对等的深度交互。如图 1(c)所示,BCAF 模块由两个并行的交叉注意力(cross attention block,CAB)模块和 GM 模块组成, CAB 模块负责两个方向的特征更新,GM 模块负责两路特征的深度交互。

1)交叉注意力(CAB)模块。如图 2 所示,交叉注意力模块的内部结构遵循标准的“Pre-LN”(Pre-layer normalization)Transformer 模块设计,主要包含两个核心子模块:多头交叉注意力(MCA)、前馈神经网络(feed-forward network,FFN)模块,通过残差连接确保稳定性。CAB 模块设计如图 2 所示。

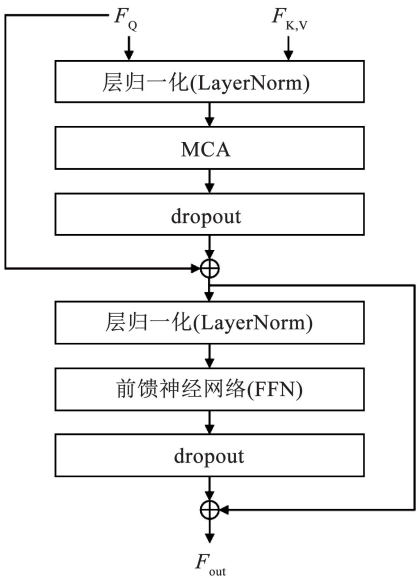


图 2　CAB 算法流程图

Fig. 2　Flowchart of CAB algorithm

①多头交叉注意力(MCA)模块。MCA 模块是完成跨模态信息交互的核心组件,为最大化表达能力,模块内生成查询(Q)、键(K)与值(V)的映射规则定义如下:

$$\boldsymbol{Q}=\boldsymbol{F}_{\mathrm{lidar}}\boldsymbol{W}_{\mathrm{Q}},\boldsymbol{K}=\boldsymbol{F}_{\mathrm{image}}\boldsymbol{W}_{\mathrm{K}},\boldsymbol{V}=\boldsymbol{F}_{\mathrm{image}}\boldsymbol{W}_{\mathrm{V}} \tag{6}$$

式中 $\boldsymbol{W}_{\mathrm{Q}},\boldsymbol{W}_{\mathrm{K}},\boldsymbol{W}_{\mathrm{V}}$ 为独立可学习的权重矩阵。

这种非共享权重设计使得模型能够根据特征在注意力计算中的不同作用执行差异化的线性变换,从而捕捉更丰富的跨模态关系。在获取 $\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$ 后,该模块采用高效的缩放点积注意力机制,该机制通过多个并行子空间进行计算。这使得模型能够同时捕捉来自不同位置和视角的特征相关性。模块负责跨模态信息融合,即

$$\mathrm{head}_i=\mathrm{Attention}(\boldsymbol{Q}_i,\boldsymbol{K}_i,\boldsymbol{V}_i)=\mathrm{softmax}\left(\frac{\boldsymbol{Q}_i\boldsymbol{K}_i^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V}_i \tag{7}$$

将所有注意力头的输出拼接后,通过一个最终的线性投影层 $\boldsymbol{W}_o \in \mathbf{R}^{C\times C}$ 进行融合,生成一个经过深度信息提炼的、更丰富的特征可表示为

$$\text{Output} = \text{Concat}(\text{head}_1, \text{head}_2, \cdots, \text{head}_h)\boldsymbol{W}_o \quad (8)$$

②前馈神经网络 (FFN)模块。每个 MCA 模块之后都连接一个 FFN 模块,FFN 模块对经过注意力聚合后的特征进行进一步的非线性变换,增强模型的表达能力并深度处理特征。它由两个线性层和一个 GeLU (Gaussian error linear unit)非线性激活函数组成。

③残差连接与层归一化。为了确保深度网络训练的稳定性并避免梯度消失,本文在 MCA、FFN 两个子模块周围都使用了残差连接。同时,在将特征输入每个子模块之前,都先进行了层归一化。这种先归一化再计算的模式能够提供更稳定的训练动态和更快的收敛速度,即:

$$\boldsymbol{X}' = \text{LayerNorm}(\boldsymbol{X} + \text{MultiHeadAttention}(\boldsymbol{X})) \quad (9)$$

$$\boldsymbol{X}_{\text{out}} = \text{LayerNorm}(\boldsymbol{X}' + \text{FFN}(\boldsymbol{X}')) \quad (10)$$

2)自适应门控模块(GM)。自适应门控模块首先通过逐点 MLP 处理拼接后的两路输入特征分别为:

$$F_{\text{concat}} = \text{Concat}(F_{\text{lidar}}, F_{\text{image}}) \quad (11)$$

$$F_{\text{out}} = \text{MLP}_3(\text{MLP}_2(\text{MLP}_1(F_{\text{concat}}))) \quad (12)$$

基于上述融合特征被对半分割,通过两个独立的分支网络生成各自的门控信号:$a_1 = \text{Sigmoid}(\text{MLP}(F_{\text{out1}}))$,$a_2 = \text{Sigmoid}(\text{MLP}(F_{\text{out2}}))$。最后将门控信号与各自的原始特征逐元素相乘,实现动态的特征加权与更新,即

$$F'_{\text{lidar}} = F'_{\text{lidar}} \odot a_1, F'_{\text{image}} = F'_{\text{image}} \odot a_2 \quad (13)$$

该门控模块的优势在于动态协同增强,通过自适应校准机制优化特征以抑制噪声并突出关键信息。另外,其模块化的设计支持堆叠,可实现迭代优化和特征质量的逐步提升。

3)优点分析。BCAF 模块的设计优势源于其对称和标准化的架构。对称的双流设计通过相互作用防止了单向融合的信息丢失,而标准化的 Transformer 设计提供了高稳定性、高性能,以及与现有优化策略的兼容性。

3 实 验

3.1 数据集与评估指标

本文所有实验均在权威的 KITTI 3D 目标检测数据集[29]上进行。该数据集包含 7 481 个训练帧和 7 518 个测试帧。遵循标准做法[28,30],本文将官方训练集划分为 3 712 帧的训练子集(train)和3 769 帧的验证子集(val)。

采用平均精度(average precision, AP)作为主要评估指标,该指标在 40 个召回率点下计算得出。对于“汽车”类别,真阳性(true positive)判断的 3D 交并比(intersection over union,IoU)阈值设为 0.7。同时在鸟瞰图(bird's eye view, BEV)视角下对性能进行评估。所有结果均按照官方定义的 3 个难度等级——简单(easy)、中等(moderate),以及困难(hard)进行报告。

3.2 实验设置与基线

3.2.1 网络架构

LiDAR 支路与图像支路的输入分别采用了由 SFD[21]和 PENet[27]相同方法生成的混合点云与虚拟点云。VirConvNet 骨干网络包含 4 个阶段,其特征维度分别为 16、32、64、64。输入端与网络层级的随机体素舍弃(stochastic voxel discard, StVD)率分别设定为 80% 和 15%。体素尺寸为 0.05 m × 0.05 m × 0.05 m,每个体素最多采样 5 个点,训练与测试阶段的最大体素数量均设为 40 000。所有实验均采用与 VirConv-L[28]相同的检测范围参数。

3.2.2 损失函数与数据增强

回归损失函数对真实标签采用(complete intersection over union CIoU)损失,对伪标签采用 Smooth L1 损失。本文采用了标准的数据增强策略[28,30],包括:数据库采样(GT-sampling)、对目标施加局部噪声、全局场景的随机旋转(±45°)、沿 X 轴的随机翻转,以及全局场景的随机缩放(0.95 ~ 1.05)。

3.2.3 训练细节

本文使用 ADAM 优化器在 NVIDIA 3080 GPU 上对 VAF-Net 进行训练。学习率设为 0.001,权重衰减为 0.01,动量为 0.9。模型采用单批次(Single-batch)策略分别训练 20 和 50 个轮次(epoch)。训练阶段,非极大值抑制(non-maximum suppression, NMS)的阈值设为 0.8 以生成 160 个候选框;测试阶段,该阈值设为 0.1 以滤除冗余检测框。

3.2.4 基线方法

本文将 VAF-Net 与 7 种 SOTA 方法进行比较,包括两种纯 LiDAR 方法(Voxel-RCNN, PVT-SSD)和 5 种多模态融合方法(SFD, Graph-VoI, LoGoNet, VirConv-L, MPCF)。

1)Voxel-RCNN[6]。Voxel-RCNN 是一种高效的两阶段纯 LiDAR 3D 目标检测器。该方法提出了 Voxel RoI Pooling 层,避免了对区域内的原始点进行

操作,显著提升了第 2 阶段检测框精调的速度与精度。

2) PVT-SSD[31]。PVT-SSD 是一种结合了点、体素,以及 Transformer 的单阶段纯 LiDAR 检测框架。引入 Transformer 模块捕捉长距离依赖关系和增强上下文特征表达,能够有效融合局部几何信息与全局上下文信息。

3) SFD[21]。SFD 是一种用于多模态 3D 检测的特征解耦方法。该方法将融合后的特征解耦为形状感知和形状无关两个分支,分别用于优化定位和分类任务,避免了图像纹理信息对精确定位的干扰。

4) Graph-VoI[32]。Graph-VoI 是一种基于图神经网络的二阶段优化多模态 3D 检测器。将每个候选区域(voxel-of-interest, VoI)内的非空体素构建成一个图,并利用图卷积网络来聚合和推理这些体素间的几何关系。

5) LoGoNet[33]。LoGoNet 是一个高效融合局部与全局信息的多模态 3D 目标检测框架。设计一个全局上下文网络,通过一组可学习的"信使点"来汇聚和传播场景级别的全局特征,这种方法增强了网络对物体间关系和遮挡情况的理解能力。

6) VirConv-L[28]。VirConv-L 是一种基于虚拟点的多模态 3D 目标检测框架。该方法通过图像深度补全生成稠密的虚拟点云来辅助 LiDAR,以增强对远处物体的几何感知,显著提升了融合检测的性能和鲁棒性。

7) MPCF 是一种逐点的多模态 3D 目标检测框架。该方法针对每一个 LiDAR 原始点,从图像中提取并融合丰富的语义特征,实现两种模态的深度对齐与信息互补,提升了检测的性能,能够在复杂无人机交通场景下显著提升综合检测性能。

3.3　主要结果

1) 整体精度分析。VAF-Net 在 KITTI 验证集上实现了最先进的 3D 检测性能,超越了包括 Voxel-RCNN、PVT-SSD,以及 MPCF 在内的所有基线方法,取得了当前最优(state-of-the-art)的性能,见表 1。从表 1 可以看出,在关键的 Moderate 基准测试中,VAF-Net 达到了 86.21% 的 3D 平均精度,比次优方法 MPCF 提高了 0.71%。这一性能提升得益于本文系统的设计。虚拟点增强与 VirConvNet 骨干网络之间的协同作用,为稀疏输入提供了稳健的几何基础。此外, BCAF 模块确保了多模态信息的有效和深度融合,充分发挥了它们的协同潜力。

表 1　"Car"类别的 KITTI 1 验证集性能比较

Tab. 1　Performance comparison on KITTI 1 validation set for "Car" category

Method	Car AP_{3D}(R40)			Car AP_{BEV}(R40)		
	easy	mod.	hard	easy	mod.	hard
Voxel-RCNN[6]	90.90	81.62	77.06	94.85	88.83	86.13
PVT-SSD[31]	91.84	85.47	80.65	<u>95.80</u>	**92.11**	<u>89.23</u>
SFD[21]	91.73	84.76	77.92	95.64	91.85	86.83
Graph-VoI[32]	91.89	83.27	77.78	95.69	90.10	86.85
LoGoNet[33]	91.80	85.06	<u>80.74</u>	95.48	91.52	87.09
VirConv-L[28]	91.41	85.05	80.22	95.53	91.95	87.07
MPCF	<u>92.46</u>	<u>85.50</u>	80.69			
VAF-Net	**93.66**	**86.21**	**82.74**	**96.49**	<u>91.99</u>	**89.67**

注:参考方法的数据均来自 KITTI 3D 检测排行榜。其中:黑体数值表示各评价指标的最优结果,下划线数值表示各评价指标的次优结果。

2) 困难样本与误差控制。VAF-Net 在困难样本上表现优异,达到了 82.74% 的 3D 精度,比同样使用虚拟点但缺乏先进融合机制的 VirConv-L[28] 高出 2.52%。困难样本通常涉及远距离、小目标,以及严重遮挡,容易导致几何信息缺失(漏检)或定位误差。本文的误差降低主要得益于:1) 虚拟点云补全了远距离物体的几何形状;2) BCAF 模块引入的图像语义信息增强了对物体轮廓的感知。这一结论在 BEV(鸟瞰图)评估中得到印证,hard 难度的 AP_{BEV} 高达 89.67%,说明模型在平面定位上的误差极小。

3) 鸟瞰图(BEV)性能。在 BEV 视角检测中,VAF-Net 同样表现出色,尤其是在困难样本上精度达到了 89.67%。VAF-Net 在 BEV 任务上的优异表现源于 BCAF 模块的有效语义融合,进而提高了对物体轮廓的划分,实现了更精确的 2D 定位。

3.4　消融实验

为了深入探究 VAF-Net 中各个创新组件的具体贡献,本文在 KITTI 测试集上进行了一系列消融实验。以 MPCF 作为本文的基线模型,因为其同样采用了一种交叉注意力融合机制,便于公平比较。实验结果见表 2。

表 2　KITTI 测试集上的消融实验(AP40@0.70, 0.70, 0.70, 20 轮次)

Tab. 2　Ablation experiments on KITTI test set (AP40@0.70, 0.70, 0.70, 20 epochs)

基线	+ VirConvNet	+ BCAF	Car AP_{3D}(R40)/%		
			easy	mod.	hard
MPCF	×	×	92.46	85.50	80.69
	√	×	**93.28**	**86.09**	**82.68**

注:√表示包含该网络或模块,×表示不包含该网络或模块。黑体数值表示各评价指标的最优结果。

3.4.1 几何特征提取支路(VirConvNet)的有效性分析

在 MPCF 框架中,仅将其 LiDAR 主干替换为 VirConvNet(记为 VAF-Net_noBCAF),从表 2 可以看出,模型在 hard 难度下的 3D 精度从 80.69% 显著提升至 82.68%(+1.99%)。该性能提升验证了 VirConvNet 骨干网络处理稠密混合点云的有效性。虚拟点的引入在补全几何信息的同时也带来了噪声,而 VirConvNet 的抗噪声设计能够有效过滤这些干扰。因此,网络能从稠密但含噪声的输入中提取出更鲁棒、更可靠的几何特征。

为了进一步验证 VAF-Net 中 BCAF 模块的有效性,并探究不同训练周期对模型性能的影响,本文进行了进一步消融实验。本文对比了完整模型 VAF-Net 与移除了 BCAF 模块的 VAF-Net_ noBCAF 在训练 20、50 个轮次后的性能表现,见表 3。

表 3 BCAF 和训练轮次的消融实验(AP40@0.70, 0.70, 0.70)

Tab. 3 Ablation experiments of BCAF and training epochs (AP40@0.70, 0.70, 0.70)

模型	轮次	Car AP_{3D}(R40)/%		
		easy	mod.	hard
VAF-Net	20	**93.66**	**86.21**	82.74
	50	93.53	86.17	83.68
VAF-Net_noBCAF	20	93.28	86.09	82.68
	50	92.63	85.54	**85.16**

注:黑体数值表示各评价指标的最优结果。

3.4.2 BCAF 模块的有效性分析

BCAF 模块的有效性体现在其能加速训练并提升模型鲁棒性。在 20 周期时,VAF-Net 已在全难度上领先;到 50 周期时,其在常规目标(easy/mod. 难度)上的优势进一步扩大。需要注意的是,BCAF 避免了基线模型在长时训练中因过拟合几何特征而导致的性能退化问题。本文认为,这得益于 BCAF 的标准化双向融合机制:通过引入图像语义增强了模型的泛化能力,从而在各难度下保持了均衡的高性能。

3.4.3 训练周期的影响分析

对比两个模型在不同训练周期下的表现可以发现,延长训练时间能够显著提升模型在 hard 难度下的检测能力。VAF-Net 在 hard 难度上的 AP 提升 0.94%,而 VAF-Net_noBCAF 的提升幅度更是达到了 2.48%。这说明更长的训练周期使得模型能够学习到更鲁棒的特征,从而更好地处理被遮挡或距离较远的困难目标。

3.5 超参数设置

为探究第 2 阶段图像网络中分类损失与回归损失的最佳权重配比,本文进行了一系列实验。以 rcnn_cls_weight = 0.5 和 rcnn_reg_weight = 0.5 为基准,通过固定其中一个权重并调整另一个权重的方式来观察模型性能的变化实验结果见表 4、5。

表 4 固定分类权重改变回归权重(AP40@0.70, 0.70, 0.70, 20 轮次)

Tab. 4 Changing regression weights under fixed classification weights (AP40@0.70, 0.70, 0.70, 20 epochs)

回归权重 rcnn_reg_weight	Car AP_{3D}(R40)/%		
	easy	mod.	hard
0.2	93.41	86.03	82.57
0.5	**93.66**	**86.21**	82.74
0.8	93.14	86.17	**82.75**

注:黑体数值表示各评价指标的最优结果。

表 5 固定回归权重改变分类权重(AP40@0.70, 0.70, 0.70, 20 轮次)

Tab. 5 Changing classification weights under fixed regression weights (AP40@0.70, 0.70, 0.70, 20 epochs)

分类权重 rcnn_cls_weight	Car AP_{3D}(R40)		
	easy	mod.	hard
0.2	93.19	86.17	**82.85**
0.5	**93.66**	**86.21**	82.74
0.8	93.18	85.99	82.59

注:黑体数值表示各评价指标的最优结果。

1)回归权重(rcnn_reg_weight)的影响。表 4 的结果表明,将回归权重从 0.2 适度提升至 0.5 可实现性能的正向增益,但若进一步增加至0.8反而会导致性能的下降。这说明虽然强化回归任务能提高定位精度,但权重过高会破坏分类与回归之间的平衡,最终以牺牲分类精度为代价损害整体性能。

2)分类权重(rcnn_cls_weight)的影响。从表 5 可以看出,当回归权重固定为 0.5 时,将分类权重从 0.5 调整至 0.2 或 0.8 均会导致性能下降,这再次验证了分类与回归任务间保持平衡的重要性。具体而言,权重过低(0.2)会削弱模型的物体识别能力;而权重过高(0.8)则会使模型过于严苛地筛选分类置信度,从而错误地惩罚了那些定位精准但置信度稍低的有效检测结果。

3)分类权重与回归权重的平衡。在超参数分析中,本文发现将分类权重与回归权重设 1:1 时效果最佳。过分强调回归权重虽然可能微调定位框,但会破坏分类权重置信度的排序,导致整体精度下

降。VAF-Net 通过均衡的损失设计,实现了分类精度与定位误差的最佳权衡。

4　结　论

1)提出了基于虚拟点与稀疏卷积的抗噪增强框架。针对点云稀疏性问题,利用图像生成虚拟点进行稠密化,并结合 VirConvNet 骨干网络的三维子流形卷积与随机体素舍弃策略,有效解决了引入虚拟点伴随的噪声问题。实验结果表明,该框架能够从含噪声数据中提取鲁棒的几何特征。

2)设计了标准化的双向跨注意力融合(BCAF)模块。该模块遵循标准 Transformer 设计原则,通过对称的双流架构和自适应门控机制,实现了 LiDAR 几何特征与图像语义特征的深度双向交互。消融实验证实,BCAF 不仅加速了模型收敛,还有效防止了长时间训练中的过拟合现象,验证了标准化交互机制的稳定性与优越性。

3)实现了复杂场景下检测性能的显著提升。在 KITTI 3D 目标检测基准测试中,VAF-Net 取得了 SOTA 性能,特别是在困难(hard)样本上的 3D 平均精度(AP)达到了 82.74%,鸟瞰图(BEV)AP 达到了 89.67%。这一结果证明了 VAF-Net 在处理远距离、小目标,以及遮挡场景时具有极高的鲁棒性和定位精度。

参考文献

[1] Qian Rui, Lai Xin, Li Xirong. 3D object detection for autonomous driving: a survey[J]. Pattern Recognition, 2022, 130: 108796. DOI: 10.1016/j.patcog.2022.108796

[2] Rogez G, Khademi M, SupančičIII J S, et al. 3D hand pose detection in egocentric RGB-D images[C]// Computer Vision – ECCV 2014 Workshops. Cham: Springer, 2015: 356. DOI: 10.1007/978-3-319-16178-5_25

[3] Yang Y H, Piccinelli L, Segu M, et al. 3D-MOOD: lifting 2D to 3D for monocular open-set object detection[PP/OL]. V2. arXiv (2025-09-05) [2025-10-18]. https://doi.org/10.48550/arXiv.2507.23567

[4] Chen Xiaozhi, Ma Huimin, Wan Ji, et al. Multi-view 3D object detection network for autonomous driving[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu: IEEE, 2017: 6526. DOI: 10.1109/CVPR.2017.691

[5] Xu Qiangeng, Zhong Yiqi, Neumann U. Behind the curtain: learning occluded shapes for 3D object detection[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2022, 36(3): 2893. DOI: 10.1609/aaai.v36i3.20194

[6] Deng Jiajun, Shi Shaoshuai, Li Peiwei, et al. Voxel R-CNN: towards high performance voxel-based 3D object detection[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 35(2): 1201. DOI: 10.1609/aaai.v35i2.16207

[7] Shi Shaoshuai, Guo Chaoxu, Jiang Li, et al. PV-RCNN: point-voxel feature set abstraction for 3D object detection[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle: IEEE, 2020: 10526. DOI: 10.1109/CVPR42600.2020.01054

[8] Bai Xuyang, Hu Zeyu, Zhu Xinge, et al. TransFusion: robust LiDAR-camera fusion for 3D object detection with transformers[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans: IEEE, 2022: 1080. DOI: 10.1109/CVPR52688.2022.00116

[9] Reading C, Harakeh A, Chae J, et al. Categorical depth distribution network for monocular 3D object detection[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville: IEEE, 2021: 8551. DOI: 10.1109/CVPR46437.2021.00845

[10] Wang Xinlong, Yin Wei, Kong Tao, et al. Task-aware monocular depth estimation for 3D object detection[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(7): 12257. DOI: 10.1609/aaai.v34i07.6908

[11] Wang Yan, Chao Weilun, Garg D, et al. Pseudo-LiDAR from visual depth estimation: bridging the gap in 3D object detection for autonomous driving[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach: IEEE, 2020: 8437. DOI: 10.1109/CVPR.2019.00864

[12] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[PP/OL]. V7. arXiv (2023-08-02) [2025-10-18]. https://doi.org/10.48550/arXiv.1706.03762

[13] Liu Zhe, Zhao Xin, Huang Tengteng, et al. TANet: robust 3D object detection from point clouds with triple attention[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(7): 11677. DOI: 10.1609/aaai.v34i07.6837

[14] Beltrán J, Guindel C, Moreno F M, et al. BirdNet: a 3D object detection framework from LiDAR information[C]//2018 21st International Conference on Intelligent Transportation Systems (ITSC). Maui: IEEE, 2018: 3517. DOI: 10.1109/ITSC.2018.8569311

[15] Meyer G P, Laddha A, Kee E, et al. LaserNet: an efficient probabilistic 3D object detector for autonomous driving[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach: IEEE, 2019: 12677. DOI: 10.1109/CVPR.2019.01296

[16] He Chenhang, Zeng Hui, Huang Jianqiang, et al. Structure aware single-stage 3D object detection from point cloud[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle: IEEE, 2020: 11870. DOI: 10.1109/CVPR42600.2020.01189

[17] 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). Long Beach: IEEE, 2020: 12689. DOI: 10.1109/CVPR.2019.01298

[18] 陈刚, 张文标, 王洪雁, 等. 三维激光雷达点云级联匹配多目标跟踪方法[J]. 光电工程, 2025, 52(6): 49
Chen Gang, Zhang Wenbiao, Wang Hongyan, et al. The cascade

matching method for multi-object tracking of 3D LiDAR point clouds [J]. Opto-Electronic Engineering, 2025, 52(6): 49. DOI: 10.12086/oee.2025.250050

[19] Shi Shaoshuai, Wang Xiaogang, Li Hongsheng. PointRCNN: 3D object proposal generation and detection from point cloud [C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach: IEEE, 2020: 770. DOI: 10.1109/CVPR.2019.00086

[20] 陈旭东, 韩丽丽, 肖英豪, 等. 基于体素-关键点特征聚合网络的点云 3 维目标检测[J]. 激光技术, 2026, 50(2): 291
Chen Xudong, Han Lili, Xiao Yinghao, et al. 3-D object detection for point clouds based on voxel-keypoint feature aggregation network [J]. Laser Technology, 2026, 50(2): 291. DOI: 10.7510/jgjs.issn.1001-3806.2026.02.018

[21] Wu Xiaopei, Peng Liang, Yang Honghui, et al. Sparse fuse dense: towards high quality 3d detection with depth completion[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans: IEEE, 2022: 5408. DOI: 10.1109/CVPR52688.2022.00534

[22] Yin Tianwei, Zhou Xingyi, Krähenbühl P. Multimodal virtual point 3D detection[PP/OL]. V1. arXiv (2021-11-12) [2025-10-19]. https://doi.org/10.48550/arXiv.2111.06881

[23] 李文平, 袁强, 陈璐, 等. 基于雷达点云与图像数据的三维目标检测方法[J]. 电光与控制, 2021, 28(10): 110
Li Wenping, Yuan Qiang, Chen Lu, et al. Three-dimensional object detection method based on radar point cloud and image data [J]. Electronics Optics & Control, 2021, 28(10):110. DOI:10.3969/j.issn.1671-637X.2021.10.023

[24] Chen Xuanyao, Zhang Tianyuan, Wang Yue, et al. FUTR3D: a unified sensor fusion framework for 3D detection[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Vancouver: IEEE, 2023: 172. DOI: 10.1109/CVPRW59228.2023.00022

[25] Ku J, Mozifian M, Lee J, et al. Joint 3D proposal generation and object detection from view aggregation [C]//2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Madrid: IEEE, 2019: 1. DOI: 10.1109/IROS.2018.8594049

[26] Liu Zhijian, Tang Haotian, Amini A, et al. BEVFusion: multi-task multi-sensor fusion with unified bird's-eye view representation [C]//2023 IEEE International Conference on Robotics and Automation (ICRA). London: IEEE, 2023: 2774. DOI: 10.1109/ICRA48891.2023.10160968

[27] Hu Mu, Wang Shuling, Li Bin, et al. PENet: towards precise and efficient image guided depth completion [C]//2021 IEEE International Conference on Robotics and Automation (ICRA). Xi'an: IEEE, 2021: 13656. DOI: 10.1109/icra48506.2021.9561035

[28] Wu Hai, Wen Chenglu, Shi Shaoshuai, et al. Virtual sparse convolution for multimodal 3D object detection[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver: IEEE, 2023: 21653. DOI: 10.1109/CVPR52729.2023.02074

[29] Geiger A, Lenz P, Urtasun R. Are we ready for autonomous driving? The KITTI vision benchmark suite [C]//2012 IEEE Conference on Computer Vision and Pattern Recognition. Providence: IEEE, 2012: 3354. DOI: 10.1109/CVPR.2012.6248074

[30] Wu Hai, Deng Jinhao, Wen Chenglu, et al. CasA: a cascade attention network for 3-D object detection from LiDAR point clouds [J]. IEEE Transactions on Geoscience and Remote Sensing, 2022, 60: 5704511. DOI: 10.1109/TGRS.2022.3203163

[31] Yang Honghui, Wang Wenxiao, Chen Minghao, et al. PVT-SSD: single-stage 3D object detector with point-voxel transformer[C]// 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver: IEEE, 2023: 13476. DOI: 10.1109/CVPR52729.2023.01295

[32] Yang Honghui, Liu Zili, Wu Xiaopei, et al. Graph R-CNN: towards accurate 3D object detection with semantic-decorated local graph[C]//Computer Vision-ECCV 2022. Cham: Springer, 2022: 662. DOI: 10.1007/978-3-031-20074-8_38

[33] Li Xin, Ma Tao, Hou Yuenan, et al. LoGoNet: towards accurate 3D object detection with local-to-global cross-modal fusion[C]// 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver: IEEE, 2023: 17524. DOI: 10.1109/CVPR52729.2023.01681

(编辑 张 红)