DOI: 10.19650/j.cnki.cjsi.J2311940

基于 Transformer 的融合信息增强 3D 目标检测算法*

金宇锋[1],陶重犇[1,2]
(1. 苏州科技大学电子与信息工程学院 苏州 215009; 2. 清华大学汽车研究院 苏州 215134)

摘 要:针对当前 3D 目标检测算法将不同模态数据融合时会产生错位现象,从而破坏数据之间的关联性并造成数据损失的问题,提出了一种基于 Transformer 的融合信息增强 3D 目标检测算法。首先设计了 Transformer 双域融合特征区域建议模块,利用变形注意力机制,将提取到的雷达点云特征和图像特征进行双域特征融合,用于生成 3D 预选框;其次,通过设计的深度补全机制的特征信息增强模块,补全密集的深度和特征语义信息来完成框的细化;最后,设计了多模态特征交叉注意力模块,采用动态交叉注意力机制来获得不同模态间的相关性,从而将特征信息有效对齐融合。在 Kitti、Nuscences 和 Waymo 数据集上的实验结果证明了该算法的有效性和通用性。大量的消融实验证明了该算法各个模块的有效性。在实车平台上的实验结果表明,该算法在复杂的实际环境中具有优秀的鲁棒性。

关键词: 3D 目标检测;Transformer;深度补全;多模态融合;自动驾驶
中图分类号: TH741 TP391.4 **文献标识码**: A **国家标准学科分类代码**: 510.4050

Fusion information enhanced method based on transformer for 3D object detection

Jin Yufeng[1], Tao Chongben[1,2]
(*1. School of Electronic and Information Engineering, Suzhou University of Science and Technology, Suzhou 215009, China*; *2. Suzhou Automotive Research Institute, Tsinghua University, Suzhou 215134, China*)

Abstract: A fusion information enhanced method based on Transformer is proposed to address the issue of misalignment when the current 3D object detection methods fuse different modal data, which mitigates the disruption of correlation between data and data loss. Firstly, a region proposal network of dual fusion feature module based on transformer is designed, which utilizes the deformable attention mechanism to fuse the extracted lidar point cloud features and image features into dual domain features and generate pre-selected boxes. Then, the refinement of box is designed by using a feature information enhancement module, which utilizes a deep completion mechanism to complement the dense depth and feature semantic information. Finally, a multimodal feature cross attention module is designed, which uses a dynamic cross attention mechanism to obtain correlations between different modalities, thereby aligning and fusing feature information effectively. The experimental results based on the Kitti, Nucences, and Waymo datasets demonstrate the effectiveness of method. A large number of ablation experiments have proven the effectiveness and efficiency of each module in the algorithm. The experimental results based on a real vehicle platform show that the algorithm possesses strong robustness in complex practical environments.

Keywords: 3D object detection; Transformer; depth complementation; multimodal fusion; autonomous driving

0 引 言

在当前的自动驾驶研究中,三维目标检测技术已经成为一个不可或缺的重要部分。现有的三维目标检测方法主要是利用激光雷达与像机来感知周围环境。其中基于点云的方法[1-2]和基于图像的方法[3-6]已经取得了令人映像深刻的效果。激光雷达和像机这两种传感器具有显著不同的特性。激光雷达能够在近距离提供精确的 3D 测量数据,但是在远距离上点云会变得稀疏。像机能够

收稿日期:2023-09-19 Received Date: 2023-09-19
*基金项目:国家自然科学基金(62372317)、中国博士后科学基金(2021M691848)、苏州市科技项目基金(SYG202142)资助

提供色彩和纹理丰富的物体特征,但不是良好的深度信息来源。二者特性的互补使得激光雷达-像机传感器的融合成为近年来人们感兴趣的话题。这种组合已经被证明可以运用在包括自动驾驶在内的许多应用场景中,并实现高精度的三维物体检测。因此,融合时减少因处于不同域而导致的错位现象,提升模态数据之间的关联性是关键。

现有的多模态目标检测融合方法在点层、建议层和特征层都采用了不同的融合策略。点融合方法通常将像机特征信息投影到三维空间中的激光雷达点,将图像点特征与点状激光点特征结合,但是局限性在于密集的图像信息与稀疏的点云特征往往无法有效对齐,从而破坏不同模态数据之间的关联性并造成数据损失。建议层融合方法先从像机图像中生成2D建议区域,并将激光雷达特征与相应的建议区域相关联。然后,将二者的特征相融合并细化建议框。这种方法的性能受到生成的建议准确性的限制。特征级融合方法分别从像机图像与激光雷达点云中提取相应的特征,并在体素域中进行对齐融合。PointAugmenting[7]、3D-GAF[8]将图像特征利用校准矩阵变换到体素域,进行元素特征融合。而最近基于Transformer的检测方法[9-10]已经非常流行。DeepFusion[11]、TransFusion[12]等算法使用了Transformer的注意机制,将不同域的特征进行对齐融合,实现了鲁棒的目标检测效果。

面对上述问题,本文提出了一种新颖的、端对端的多模态三维目标检测算法(transformer fusion information enhancement network,TFIENet)。首先,针对多模态数据处于不同域在进行特征融合时,会产生错位现象,破坏数据之间关联性的问题,设计了Transformer双融合特征区域建议网络模块。采用可变形的变换器-解码器,利用变形注意力机制,将提取到的雷达点云特征和图像特征进行双融合。实现聚合LiDAR和像机双域特征信息用以生成初始候选框。在第2阶段,提出了特征信息增强模块。该模块通过深度特征补全机制,利用图像的丰富纹理特征,预测出密集深度特征信息。提取出的密集的深度信息和特征语义信息用于改善特征表示。从而增强目标检测能力,提高远距离小目标的定位精度。最后,为了能够将来自不同模态的特征信息能够有效对齐融合。设计了一种多模态特征交叉注意力模块。采用动态交叉注意力机制来获得不同模态间的相关性,并预测相关权重。再利用权重对这对特征进行加权,以获得融合特征。本文提出的方法在具有挑战性的自动驾驶数据集上进行了测试和评价。总体而言,各模块之间的协同工作提高了系统的有效性和鲁棒性。

本文设计了一种Transformer双域融合特征区域建议模块。采用可变形的变换器-解码器结构,将图像和雷达特征进行双域融合。这种方法有效地聚合了LiDAR和像机的双域特征信息,从而生成预选框。设计了深度特征信息增强模块。该模块通过深度信息补全机制,预测密集深度信息和提取特征语义信息,从而提高了特征表示能力。这种方法有助于后续框精化和提高置信度预测的准确性,并最终提高远距离小目标的定位精度。设计了多模态特征交叉注意力模块,该模块采用动态交叉注意力机制来获得不同模态间的相关性,并预测相关权重。再通过对特征进行加权来获得融合特征,有助于提高多模态数据的利用效率。

1 TFIENET的网络架构

TFIENet算法的整体框架结构,如图1所示,这是一种新的多模态三维目标检测算法。TFIENet算法由主干网络、Transformer双域融合特征区域建议模块、深度特征信息增强模块、多模态特征交叉注意力模块和二维三维检测头5个主要部分组成。

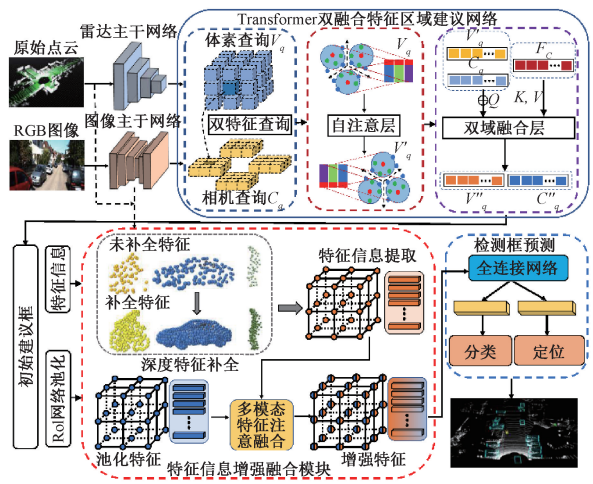


图1 TFIENet算法框架图

Fig. 1 Framework of TFIENet Algorithm

1.1 Transformer双域融合特征区域建议模块

1)双特征查询

双特征查询是指通过像机特征查询和体素特征查询两种方式来得到更加准确的特征。在使用激光雷达特征时,会将其进行体素化处理,从而得到 k 个非空体素,其中数量 k 依据输入激光雷达特征的分布而变化。体素查询 $V_q = \{V_q^1, V_q^2, V_q^3, \cdots, V_q^k\}$ 与相应的非空体素进行匹配,从而改进体素域特征。每一个非空体素的中心点投影到像机域中的图像像素,依次由像机查询 $C_q = \{C_q^1, C_q^2, C_q^3, \cdots, C_q^k\}$ 对应分配,并用于改进图像域特征。

2)3D 局部自注意

体素查询是利用激光雷达数据在非空体素中进行初始化的方法。自注意层通过建立体素查询之间的空间位置关系来逐一对它们进行编码。为了降低全局自注意所带来的计算负担,本文设计了 3D 局部自注意层来减少局部区域内的关注范围。具体地,通过对非空体素的中心点应用最远点采样算法[13],将其聚类到局部区域,并在每个局部区域的质心周围的固定半径内找到 N 个体素。设 $\boldsymbol{T}=\{\boldsymbol{t}_n \in \boldsymbol{N}\}$ 和 $\boldsymbol{P}=\{\boldsymbol{p}_n \in \boldsymbol{N}\}$ 分别是赋给质心 C_m 的一组查询特征和三维位置。然后,3D 局部自注意层进行自我关注,如图 2 所示。

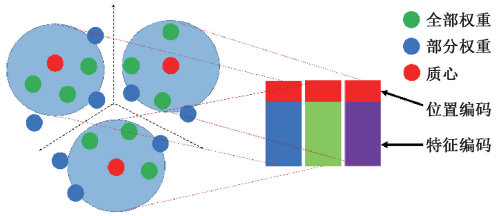


图 2　3D 局部自注意示意图

Fig. 2　3D local self-attention schematic

$$PCF(p_x,p_y)=PFN(p_x-p_y) \tag{1}$$

$$\boldsymbol{q}_x^{(s)}=\boldsymbol{t}_x^{(s)}\boldsymbol{q}_m,\boldsymbol{k}_x^{(s)}=\boldsymbol{t}_x^{(s)}k_m,\boldsymbol{v}_x^{(s)}=\boldsymbol{t}_x^{(s)}v_m \tag{2}$$

$$F_x^{(s)}=\sum_{y\in N}^{c_m}\mathrm{softmax}\left(\frac{q_xN_y}{\sqrt{\theta}}+PCF(P_x,p_y)\right) \tag{3}$$

$$t_x^{(S+1)}=t_x^{(s)}+PFN(F_x^{(s)}) \tag{4}$$

式中:$\boldsymbol{q}_m$、$\boldsymbol{k}_m$ 和 $\boldsymbol{V}_m$ 分别是 query、*key* 和 *value* 的投影矩阵;S 是 S 层转换块的索引;θ 是归一化点积关注的缩放因子;$PCF(\cdot)$ 表示位置前置网络;$PCF(p_x,p_y)$ 是位置编码函数。通过 PFN 对两个三维坐标 P_x、P_y 的插值进行编码。如果半径中非空体素个数大于 N,那么计算 N 个体素。与一组 K 个体素相关联的体素查询由自注意层单独进行编码。在通过多个自注意层后,输入的体素查询 $\boldsymbol{V}_q=\{\boldsymbol{V}_q^1,\boldsymbol{V}_q^2,\boldsymbol{V}_q^3,\cdots,\boldsymbol{V}_q^k\}$ 被更新为 $\boldsymbol{V}'_q=\{\boldsymbol{V}_q^{1'},\boldsymbol{V}_q^{2'},\boldsymbol{V}_q^{3'},\cdots,\boldsymbol{V}_q^{k'}\}$。

3)双域特征注意融合

原本的 DETR[14] 只支持单个模态上的自注意,因此为了能够有效地将多模态特征进行交互融合,本文提出了双域特征注意融合机制,如图 3 所示。设 Q 为非空体素。第 K 个非空体素的中心点处的 3D 参考点 $\boldsymbol{P}_{3D}^k=(\boldsymbol{a},\boldsymbol{b},\boldsymbol{c})$。在像机域上投影 $\boldsymbol{P}_{3D}^k$ 并在网格上量化,从而确定相应的 2D 参考点 $\boldsymbol{P}_{2D}^k=(\boldsymbol{x},\boldsymbol{y})$。像机查询 C_q 由 2D 参考点 $\boldsymbol{P}_{2D}^k$ 指示的像机域特征 $\boldsymbol{F}'_c$,进行初始化。由三维局部自注意得到体素查询 V'_q。深度感知位置编码首先应用于双特征查询。不同于原本的 DETR[14] 中采用基于 $(\boldsymbol{x},\boldsymbol{y})$ 进行编码,这里的位置编码是依据 $\boldsymbol{P}_{3D}^k$ 的深度:

$$PCF_k^{(i)}=\sin\left(\frac{z}{10\ 000^{i/d}}\right) \tag{5}$$

$$PCF_k^{(i+1)}=\cos\left(\frac{z}{10\ 000^{i/d}}\right) \tag{6}$$

式中:i 和 d 分别是查询向量的索引和维度。然后将深度感知位置编码添加到双特征查询之中。

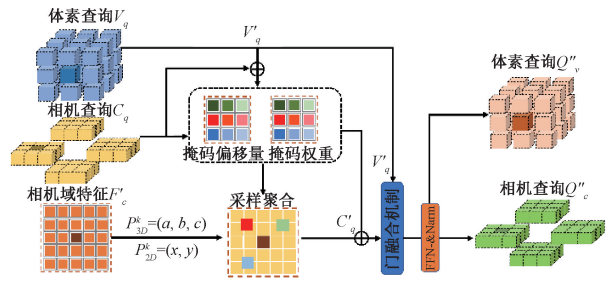


图 3　Transformer 双域融合特征区域建议网络结构

Fig. 3　The region proposal network structure of transformer dual fusion feature based on transformer

双域特征融合注意解码多个注意层上的双查询 V'_q 和 C_q。首先,像机特征 $\boldsymbol{F}'_c$ 为 *key* 和 *value*,对像机查询 C_q 进行变形注意变换。对于给定的 2D 参考点 $\boldsymbol{P}_{2D}^k=(\boldsymbol{x},\boldsymbol{y})$。在像机域特征 $\boldsymbol{F}'_c$ 上应用具有自适应偏移量和权重的变形掩模。掩码偏移量 Δd_m 和掩码权重 Δw_m:

$$\Delta d_m=PFN(C_q),\Delta w_m=PFN(C_q+V'_q) \tag{7}$$

式中:PFN 表示前馈网络。可知,掩码权重 Δd_m 是由体素查询和像机查询共同确定。由此可以通过基于体素区域和像机区域特征的注意力权重来提高特征融合的效果。给定偏移量 Δd_m 和权重 Δw_m,关注值 $\boldsymbol{C}'_q=\{\boldsymbol{C}_q^{1'},\boldsymbol{C}_q^{2'},\boldsymbol{C}_q^{3'},\cdots,\boldsymbol{C}_q^{k'}\}$ 计算如下:

$$\boldsymbol{C}'_q=\sum_{h=1}^{H}M_{lb}\sum_{k=1}^{K}M'_{lb}\times\boldsymbol{V}\boldsymbol{w}_m\times\boldsymbol{F}'_c(\boldsymbol{p}_{2D}^k+\boldsymbol{V}\boldsymbol{d}_m) \tag{8}$$

式中:h 是关注头的索引;k 对采样 *key* 进行索引;K 是采样 *key* 的总数;$\boldsymbol{M}_{lb}\in\boldsymbol{R}^{C\times C/H}$ 和 $\boldsymbol{M}'_{lb}\in\boldsymbol{R}^{C/H\times C}$ 表示可学习投影矩阵。

体素查询 V'_q 和像机查询 C'_q 通过门控融合机制进一步融合转换,即 V'_q 和 C'_q 通过如下不同比例融合:

$$\boldsymbol{C}''_q=\boldsymbol{C}'_q+\boldsymbol{V}'_q\times\boldsymbol{\sigma}(\boldsymbol{Con}\boldsymbol{V}_1(\boldsymbol{C}'_q+\boldsymbol{V}'_q)) \tag{9}$$

$$\boldsymbol{V}''_q=\boldsymbol{V}'_q+\boldsymbol{C}'_q\times\boldsymbol{\sigma}(\boldsymbol{Con}\boldsymbol{V}_2(\boldsymbol{C}'_q+\boldsymbol{V}'_q)) \tag{10}$$

式中:$\boldsymbol{\sigma}(\cdot)$ 是 sigmoid 函数;$\boldsymbol{Con}\boldsymbol{V}_1(\cdot)$ 和 $\boldsymbol{Con}\boldsymbol{V}_1(\cdot)$ 表示不同的卷积层。由于比例是关于 $\boldsymbol{V}'_q$ 和 $\boldsymbol{C}'_q$ 的函数,因此该组合比率会根据输入特征来进行自适应地调整。

1.2　特征信息增强模块

本文的特征信息增强模块的核心思想是预测出密集的深度特征信息来改善特征表示。因此,需要分别完成

两个子任务:即如何预测出密集的深度特征信息以及如何将特征有效提取出来。

1)深度信息补全

候选框中的前景点构成了描述目标线索的特征,但是因为雷达点云的稀疏性,深度特征信息通常是不完整的。因此,通过在成熟的深度特征信息补全框 SANS(Sparse Auxiliary Networks)[15]的基础上,设计了一个有效的特征信息预测网络,如图 4 所示。通过补全密集的深度特征信息来增强特征表示。

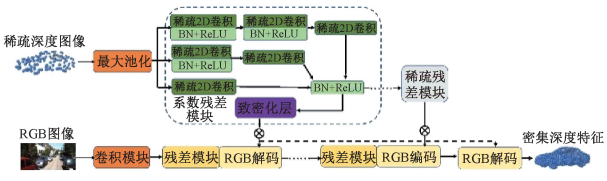


图 4 深度特征补全示意图

Fig. 4 Schematic diagram of depth feature complementation

以稀疏的深度和 RGB 图像作为输入,输出补全的密集深度图。为了利用有效的信息并减少计算,采用稀疏卷积和标准的卷积分别来处理输入的深度图和 RGB 图像。框架中稀疏张量 $\boldsymbol{T}$ 由坐标矩阵 $\boldsymbol{C}$ 和特征矩阵 $\boldsymbol{F}$ 决定:

$$\boldsymbol{C}=\begin{bmatrix}\boldsymbol{x}_1 & \boldsymbol{y}_1 & \boldsymbol{I}_1\\ \vdots & \vdots & \vdots\\ \boldsymbol{x}_N & \boldsymbol{y}_N & \boldsymbol{I}_N\end{bmatrix},\ \boldsymbol{F}=\begin{bmatrix}\boldsymbol{f}_1\\ \vdots\\ \boldsymbol{f}_N\end{bmatrix} \tag{11}$$

式中:$\{\boldsymbol{x}_n,\boldsymbol{y}_n\}$ 是像素点坐标;$\boldsymbol{I}_N$ 是批处理中的样本索引;$\boldsymbol{f}_N\in\boldsymbol{R}^Q$ 是对应的特征向量。可以利用有效像素的坐标和深度值,将其作为特征来稀疏化 $W\times H\times 1$ 深度图 $\widetilde{\boldsymbol{D}}$,

$$\widetilde{\boldsymbol{T}}=\{\{(x,y),\widetilde{\boldsymbol{D}(x,y)}\}\ \forall x,y\in\widetilde{\boldsymbol{D}}\,|\,\widetilde{\boldsymbol{D}}(x,y)>0\} \tag{12}$$

相应地,稀疏张量 $\widetilde{\boldsymbol{T}}=\{\widetilde{\boldsymbol{C}},\widetilde{\boldsymbol{F}}\}$ 可以将其有效像素坐标和深度值投射到密集的 $W\times H\times D$ 矩阵 $\widetilde{\boldsymbol{p}}$ 中来致密化,

$$\widetilde{\boldsymbol{p}}(\boldsymbol{x}_n,\boldsymbol{y}_n)=\begin{cases}\boldsymbol{f}_N, & \{\boldsymbol{x}_n,\boldsymbol{y}_n\}\in\widetilde{\boldsymbol{T}}\\ 0, & \text{其他}\end{cases} \tag{13}$$

当输入为稀疏的深度信息时,便通过一系列稀疏残差模块(spare residual block,SRB)对其进行位置编码。每个 SRB 具体由 3 个并行分支组成,每个分支具有不同数量的稀疏卷积块。输入在经过最大池阶段处理后,输入不同的分支,输出更深层特征作为下一个 SRB 的输入。并且在每个 SRB 后,并行地使用致密化层来生成这些稀疏特征的致密表示。

输入的图像通过 RGB 模块,提取密集的 RGB 特征。通过将 RGB 编码器与 SRB 编码器的分辨率设计得相互匹配,使得在 SRB 中获得的深度特征通过添加特征映射,在致密化后融合到 RGB 特征中。其中,RGB 编码器仅处理图像特征,而 SRB 解码器处理融合过后的增强特征,该特征由深度编码器输出的密集深度特征增强而来。最终输出包括带有丰富语义信息的 RGB 特征外,还有具有密集结构特征的深度信息。

2)增强特征提取

预测出的增强特征信息以伪云的形式存在,其蕴含有丰富的语义和结构信息。传统的卷积算法直接对其进行体素化并进行 3D 稀疏卷积,这样无法有效利用伪云中特征信息。传统算法中的球查询操作会产生大量计算量,且没有将 2D 邻域关系考虑进去。为此,受到网格搜索[16]的启发,本文提出了一种邻域点卷积(neighbor point convolution, NPConv)。该卷积采用感兴趣区域的邻居搜索策略,可以在恒定时间内搜索领域,这样可以减少计算量的同时通过领域关系可以提取 2D 语义特征,如图 5 所示。

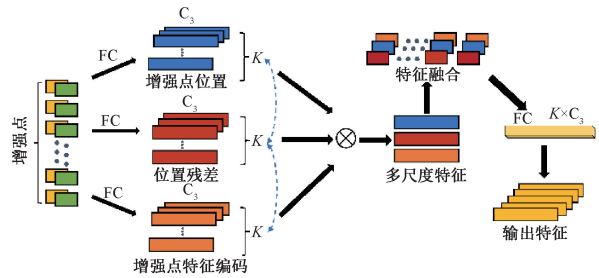


图 5 增强特征提取示意图

Fig. 5 Schematic diagram of enhanced feature extraction

对于增强特征点 N_i,将其表示为 $\boldsymbol{N}_i=(\boldsymbol{w}_i,\boldsymbol{h}_i,\boldsymbol{p}_i,\boldsymbol{r}_i,\boldsymbol{g}_i,\boldsymbol{b}_i,\boldsymbol{x}_i,\boldsymbol{y}_i)$,其由三维几何特征($\boldsymbol{w}_i,\boldsymbol{h}_i,\boldsymbol{p}_i$)、二维语义特征($\boldsymbol{r}_i,\boldsymbol{g}_i,\boldsymbol{b}_i$)及坐标($\boldsymbol{x}_i,\boldsymbol{y}_i$)共同组成。首先,在增强点特征上应用一个全连接层,来减小计算复杂度。其次,在全连接层后,特征通道由 C_2 提升到 C_3。最后利用 N_i 到其邻域的三维和二维的位置残差,使得 N_i 的增强特征感知到三维和二维空间之间的局部对应关系。从而能够提取出增强特征点 N_i 对应的三维几何特征和二维语义特征。对于 N_i 的第 m 个邻点 N_i^m,N_i 和 N_i^m 之间的位置残差可表示为:

$$\boldsymbol{R}_i^m=(\boldsymbol{w}_i-\boldsymbol{w}_i^m,\boldsymbol{h}_i-\boldsymbol{h}_i^m,\boldsymbol{p}_i-\boldsymbol{p}_i^m,\boldsymbol{y}_i-\boldsymbol{y}_i^m,\boldsymbol{w}_i-\boldsymbol{w}_i^m,\|\boldsymbol{N}_i-\boldsymbol{N}_i^m\|) \tag{14}$$

其中,$\|\boldsymbol{N}_i-\boldsymbol{N}_i^m\|=\sqrt{(\boldsymbol{w}_i-\boldsymbol{w}_i^m)^2+(\boldsymbol{h}_i-\boldsymbol{h}_i^m)^2+(\boldsymbol{p}_i-\boldsymbol{p}_i^m)^2}$。

对于增强特征点 N_i 的 M 个邻域点,通过收集它们的位置信息并计算出相应位置残差。接着在位置残差上连接一个完全连接层,并将它们的特征通道与增强特征点特征对齐。对于一组邻域特征 $\boldsymbol{F}_i=\{\boldsymbol{f}_i^m\in\boldsymbol{R}^{c_3},m\epsilon 1,\cdots,\boldsymbol{M}\}$ 和一组邻域位置残差 $\boldsymbol{R}_i^m=\{\boldsymbol{h}_i^m\in\boldsymbol{R}^{c_3},m\epsilon 1,\cdots,\boldsymbol{M}\}$,将相应的邻域位置残差与特征进行加权,加权后的邻域

特征通过串联以获得最大的信息保真。最后,连接一个全连接层将聚集特征通道映射回 C_3。考虑到高级特征可以提供更大的感受野和更丰富的语义信息,而低级特征可以提供更精细的结构信息,因此本文将多个 NPConv 堆叠以获得更深层特征,并将 NPConv 的输出连接起来得到更加全面的增强特征。

1.3　多模态特征交叉注意力模块

由于图像数据与点云数据之间存在维度间隙,通过深度补全的方式将得到的深度增强特征,将图像特征转换成点云特征相似的 3D 伪云形式,便可以用更加精细化地方式将增强特征与原始特征融合起来,用以改善特征表示。本文中的多模态特征交叉注意力模块采用 3D 网格融合机制和注意力融合机制,如图 6 所示。

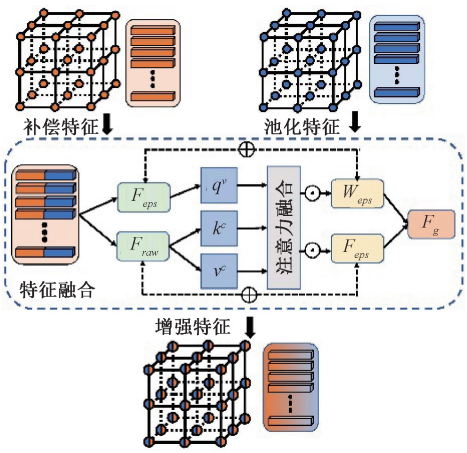


图 6　多模态特征交叉注意力模块示意图

Fig. 6　Schematic diagram of multimodal feature cross-attention module

1)3D 网格融合机制

本文使用一个 3D 网格模块来分别裁剪增强特征数据与原始点云数据。相较于使用 2D 感兴趣区域来提取图像特征的方法,可以减少许多其他对象或背景的干扰。并且以往的多模态数据融合方法中,由于不同模态数据之间存在域差距,密集图像与稀疏点云数据之间也有数量及不同的表示。因此往往无法将相关的特征数据进行有效的加强。由于深度补全后的增强数据特征与原始的感兴趣域特征具有相同的表示,因此可以将对应的网格化特征进行专注融合。从而能够有效利用图像中的语义信息和点云深度信息来改善目标表示,提高检测能力。

2)注意力融合机制

为了更好地将相应特征数据进行对齐,减少融合错位现象,还采用了动态交叉注意力融合机制。该机制够引入特征之间的依赖关系,从而更好地捕捉数据的相关性。该机制可以有效地获得不同模态数据间的相关性,并预测相关权重。通过对特征进行加权,可以更加准确地获得融合特征,进而提高多模态数据的利用效率。

将 $\boldsymbol{F}_{raw} \in \boldsymbol{R}_g^k$ 和 $\boldsymbol{F}_{eps} \in \boldsymbol{R}_g^k$ 用来分别表示原始感兴趣域特征和增强数据特征,其中,K 是 3D 网格块的数量,g 是网格特征通道。$\boldsymbol{F}_{raw}$ 和 $\boldsymbol{F}_{eps}$ 的第 n_{th} 个特征分别表示为 $\boldsymbol{F}_{raw}^{n_{th}}$ 和 $\boldsymbol{F}_{eps}^{n_{th}}$。给定一对网格特征($\boldsymbol{F}_{raw}^{n_{th}}$,$\boldsymbol{F}_{eps}^{n_{th}}$),分别使用 3 个连接层将增强数据特征转换成查询 q^v,将原始感兴趣域特征转换成键 k^c 和值 v^c。对于每个查询,查询 q^v 和键值 k^c 之间进行内积,以获得增强数据特征与其对应的原始特征间的注意力亲和矩阵。在通过 softmax 层归一化后,该矩阵用来衡量和聚合值 v^c,将其联立通过全连接层和 sigmoid 层,产生一对权重($\boldsymbol{W}_{raw}^{n_{th}}$,$\boldsymbol{W}_{eps}^{n_{th}}$),最后,将权重($\boldsymbol{W}_{raw}^{n_{th}}$,$\boldsymbol{W}_{eps}^{n_{th}}$)加权于($\boldsymbol{F}_{raw}^{n_{th}}$,$\boldsymbol{F}_{eps}^{n_{th}}$),得到增强网格融合特征 $\boldsymbol{F}_e$。将聚合后的特征信息通过一个完全连接层进行处理,并与深度特征进行级联输出。增强网格融合特征 $\boldsymbol{F}_e$ 如下:

$$\boldsymbol{F}_e = \boldsymbol{MLP}(\boldsymbol{CONCAT}(\boldsymbol{W}_{raw}^{n_{th}} \rightarrow \boldsymbol{F}_{raw}^{n_{th}}, \boldsymbol{W}_{eps}^{n_{th}} \rightarrow \boldsymbol{F}_{eps}^{n_{th}})) \quad (15)$$

2　实验与分析

本文分别在具有挑战性的 Kitti[17] 和 Nuscences[18] 目标检测数据集和最新的开放数据集 Waymo[19] 上评估测试了提出的 TFIENet 框架。

2.1　Kitti 数据集上的实验结果

不同方法在 Kitti 测试集中汽车类别的检测性能比较,如表 1 所示。结果表明,本文提出的方法整体性能上优于其他目标检测算法,与本文的基线 Voxel-RCNN[22] 相比,本文算法将 AP_{3D} 分别提高了 0.65、1.14 和 0.16。与性能优异 SE-SSD[23] 相比,在容易和中等难度上 AP_{3D} 也分别提高了 0.63 和 0.22。

表 1　Kitti 测试集中汽车类别的检测性能

Table 1　Test results of vehicle categories on the Kitti test set　%

方法	类型	Car AP_{3D}			
		简单	中等	困难	mAP
PV-RCNN[20]	L	90.25	81.62	77.06	82.83
3D-SSD[21]	L	88.36	79.57	74.55	80.83
Voxel-RCNN[22]	L	90.90	81.62	77.06	83.19
SE-SSD[23]	L	91.49	82.54	77.15	83.73
3D-CVF[24]	L+R	89.20	80.05	73.11	80.79
PointPainting[25]	L+R	82.11	71.70	67.08	73.63
TFIENet	L+R	**91.55**	**82.76**	77.22	**83.84**

进一步在行人与自行车类别的 Kitti 测试集上所提出方法的检测性能比较如表 2 所示。虽然这两类目标比汽车小,在检测过程中更难定位目标。本文提出的 TFIENet 框架可以有效提高模型性能,充分利用多模态信息进行互补。总体而言,本文提出的方法在这两类检测中优于其他方法。除此之外,本文还在 Kitti 汽车类别验证集上进行了的检测性能的比较。如表 3 所示,依然取得了优异的性能。Kitti 数据集上的可视化结果如图 7 所示。

表 2 Kitti 测试集中行人和自行车类别的检测性能

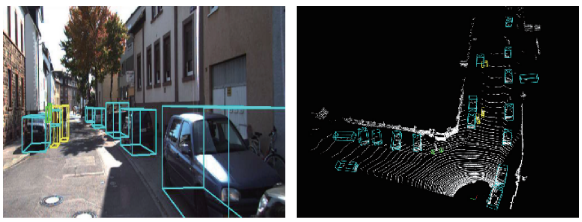
Table 2 Test results of pedestrian and bicycle categories on the Kitti test set %

方法	*Pedestrian* AP_{3D}				*Cyclist* AP_{3D}			
	简单	中等	困难	mAP	简单	中等	困难	mAP
PV-RCNN[20]	52.17	43.29	40.29	45.25	78.60	63.71	57.65	66.65
3D-SSD[21]	54.64	44.27	40.23	46.38	82.48	64.10	56.90	67.82
Voxel-RCNN[22]	53.89	43.67	40.09	45.88	82.31	63.68	57.32	67.77
SE-SSD[23]	54.31	44.25	40.12	46.22	83.36	64.63	56.82	68.27
3D-CVF[24]	54.42	43.87	40.36	46.21	83.56	64.28	56.35	68.06
PointPainting[25]	50.32	40.97	37.84	43.05	77.63	63.78	55.89	65.77
TFIENet	**54.77**	44.23	40.17	**46.39**	**83.94**	64.53	57.46	**68.64**

表 3 Kitti 验证集上汽车类别的检测性能

Table 3 Test results of vehicle categories on the Kitti validation set %

方法	类型	Car AP_{3D}			
		简单	中等	困难	mAP
PV-RCNN[20]	L	89.03	83.24	78.59	83.62
3D-SSD[21]	L	88.55	78.45	77.30	81.43
Voxel-RCNN[22]	L	90.70	84.75	78.25	83.61
SE-SSD[23]	L	90.72	85.67	79.22	85.23
3D-CVF[24]	L+R	89.20	80.05	73.11	80.79
PointPainting[25]	L+R	89.76	83.63	**79.64**	84.34
TFIENet	L+R	**91.35**	**85.89**	78.41	85.21



(a) 2D图像检测结果 (a) 2D image detection results

(b) 3D点云检测结果 (b) 3D point cloud detection results

图 7 Kitti 数据集上的可视化结果

Fig. 7 Visualization results based on the Kitti dataset

在 Kitti 数据集车辆检测结果精确度-召回率曲线对比,如图 8 所示。相较于其它目标检测算法,本文算法在高召回率时仍具有高精度。算法整体具有良好的鲁棒性和稳定性。

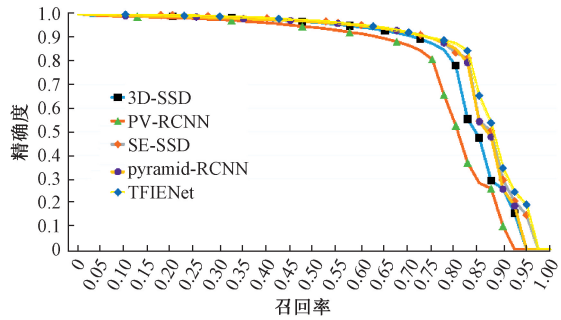
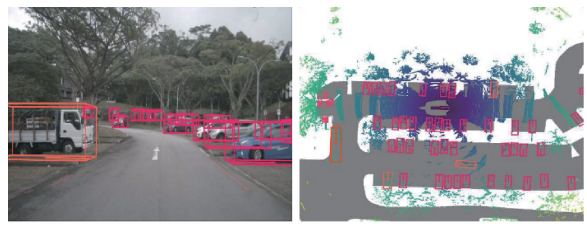


图 8 Kitti 数据集车辆检测结果精确度-召回率曲线对比

Fig. 8 Comparison of precision-recall curves for vehicles in the Kitti dataset

2.2 NuScences 数据集上的实验结果

本文还将提出的方法通过更具挑战性的 NuScences 数据集进行了实验,测试结果如表 4 所示。通过在不同数据集上实验比较进一步证实 TFIENet 的有效性和泛化性。NDS 和 mAP 是 NuScences 数据集最重要的官方评估指标。TFIENet 在 NDS 和 mAP 中较之前的最佳方法分别提高了 0.2 和 1.0。NuScences 数据集上的可视化结果如图 9 所示。



(a) 2D图像检测结果 (a) 2D image detection results

(b) 3D点云检测结果 (b) 3D point cloud detection results

图 9 NuScenses 数据集上的可视化结果

Fig. 9 Visualization results based on the NuScenses dataset

表 4　NuScenes 数据集上的性能比较

Table 4　Comparison results on the NuScenes dataset

方法	类型	NDS	mAP	汽车	货车	公交	拖车	清障车	自行车	行人
CenterPoint[26]	L	67.3	63.5	85.2	53.5	63.6	56.0	71.1	30.7	84.6
Pointpillars[27]	L	55.0	44.3	76.0	31.0	32.1	36.6	56.4	14.0	64.0
TransFusion[12]	L+R	71.7	70.8	87.1	60.0	68.3	60.8	78.1	52.9	88.4
3D-CVF[24]	L+R	62.3	56.6	83.0	45.0	48.4	49.6	65.9	30.4	74.2
PointPainting[25]	L+R	58.1	49.3	77.9	35.8	36.2	37.3	60.2	24.1	73.3
VFF[28]	L+R	72.4	70.0	86.8	58.1	70.2	61.0	73.9	52.9	87.1
TFIENet	L+R	**72.6**	**71.8**	86.4	**61.8**	**70.6**	60.5	79.5	55.6	88.2

2.3　Waymo 数据集上的实验结果

本文还将提出的方法在 Waymo 数据集上进行了实验。如表 5 所示,将本文方法的检测结果与 CenterPoint 进行了比较。本文算法在所有对象类和两个难度级别上表现出色。特别是在行人和骑自行车的人两类目标对象上分别取得了显著的收益,L2 级分别增益+2.23/+4.42。Waymo 数据集上的结果进一步验证了 TFIENet 的有效性和通用性。

表 5　Waymo 数据集上的性能比较

Table 5　Comparison results on the Waymo dataset

方法	汽车(mAP)		行人(mAP)		自行车(mAP)		All(mAP/mAPH)	
	L1	L2	L1	L2	L1	L2	L1	L2
CenterPoint[26]	66.70	62.00	73.55	68.64	72.51	70.00	70.92/68.26	66.88/64.36
TFIENet	67.56	63.56	75.67	70.87	76.65	74.42	73.29/70.61	69.62/67.10
提升	+0.86	+1.56	+2.12	+2.23	+4.41	+4.42	+2.37/2.35	2.74/2.74

2.4　在实车平台上的实验

靠虑到仅仅在数据集上进行仿真实验无法证明 TFIENet 在实际场景中的性能。因此,本文在实车平台上进行了一系列测试。如图 10 所示,实验平台是一个高度集成测试平台,配备有多个高效传感器,包括 64 线 LiDAR、高清立体像机和毫米波雷达等。

图 10　用于真实测试的实车平台

Fig. 10　The autonomous vehicle platform for real testing

为了更加直观地表现出本文提出模块的有效性,不同对象类别之间的数据如图 11 所示。表 6 为实车平台上的车辆类别的性能。结果表明,即使在复杂的真实环境中,通过设计添加不同的模块,依旧可以有效地提高算法模型的性能。

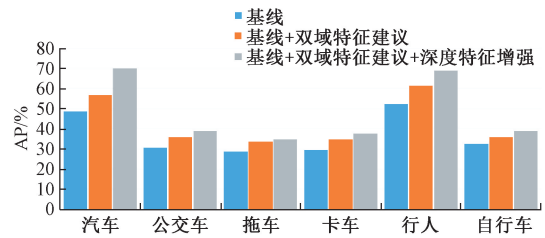


图 11　针对不同对象类别,在基线和 TFIENet 之间比较具有不同对象模块的平均精度

Fig. 11　Comparison of average accuracy between baseline and TFIENet with various object modules for different object classes

表 6　实车平台上车辆类别的表现

Table 6　Performance of vehicle categories on the real vehicle　%

方法	AP_{3D}			
	汽车	公交	货车	拖车
基线[22]	48.6	30.5	28.6	29.3
基线+双特征建议	56.8	35.6	33.5	34.6
基线+双特征建议+深度特征增强	69.8	38.7	34.5	37.6

2.5 消融实验

本文对 TFIENet 进行了全面的消融研究,以验证每个单独的组件的有效性。如表 7 所示,实验 1 是本文在 Voxel-RCNN[22] 上修改的基线。实验 2~4 分别是添加 Transformer 双域融合特征区域建议模块 A、特征信息增强模块 B、多模态特征交叉注意力模块 C,相比于基线,本文通过提出的组件的帮助在不同难度级别均实现了大幅度的提升。

表 7 不同的部分在 Kitti 验证集上的影响

Table 7 The impact of different parts on the Kitti validation set

实验	A	B	C	汽车 AP_{3D}/%		
				简单	中等	困难
1				89.41	84.52	78.93
2	√			90.86	85.76	79.87
3	√	√		91.75	86.96	80.56
4	√	√	√	92.35	87.89	81.41

Transformer 双融合特征区域建议网络模块,采用可变形的变换器-解码器结构,将图像和雷达特征进行双域融合。这种方法有效地聚合了 LiDAR 和像机的双域特征信息。为了说明该模块的有效性。如表 8 所示,本文将该模块拆分为如下 3 个组成部分:双特征查询模块 1、3D 局部自关注模块 2 和双域特征注意融合模块 3。结合所提出的融合方法,模块 1~3 在平均精度上分别提供了 0.31、0.34、0.55 的增益。

表 8 TDFN 模块的有效性

Table 8 Effectiveness of the TDFFN module

方法	模块 1	模块 2	模块 3	Car AP_{3D}/%			
				简单	中等	困难	mAP
基线[22]				89.41	84.52	78.93	84.28
是否添加组件	√			89.82	84.73	79.24	84.59
	√	√		90.31	85.12	79.38	84.93
	√	√	√	90.86	85.76	79.87	85.48

在不同的交并比(IoU)阈值和物体深度下对所提出的方法进行了评价,如图 12 所示。此外,还比较了基线性能。

本文的双特征查询机制与单像机查询进行比较如表 9 所示。单目查询意味着仅有像机查询通过变形关注来更新。可变形掩模偏移和权重仅由像机查询确定。而双特征查询通过将变形关注的输出添加到体素查询中来产生更新的像机查询。请注意,在 Kitti 验证集上,双查

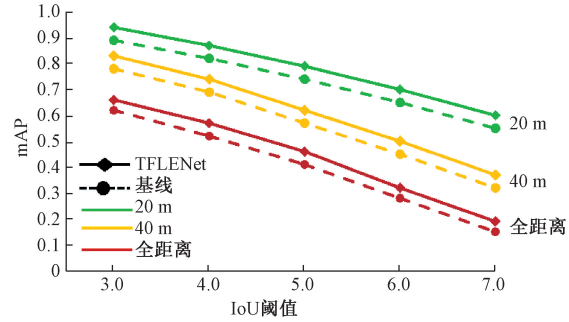


图 12 在 Kitti 汽车类别验证集上,不同物体深度范围和 IOU 阈值的 AP 在基线和 TFIENet 之间的比较

Fig. 12 Comparison between baseline and TFIENet for APs with different target depth ranges and IOU thresholds on the Kitti automotive category validation set

表 9 Kitti 验证集上不同查询类型的比较

Table 9 Comparison results of different query types on the Kitti validation set %

方法	查询类型	Car AP_{3D}			
		简单	中等	困难	mAP
基线[22]		89.41	84.52	78.93	84.28
查询类型	单目查询	89.48	84.55	78.97	84.33
	双特征查询	89.57	84.65	79.11	84.44

询机制在 mAP 指标上比单像机查询和基线分别提高 0.05 和 0.16。

Kitti 验证集关于双查询注意机制解码器层数的性能趋势如图 13 所示。在 Kitti 数据集中,随着层数增加到 4 层,性能逐渐提高。这表明本文的双查询注意逐步细化了用于三维目标检测的特征质量。

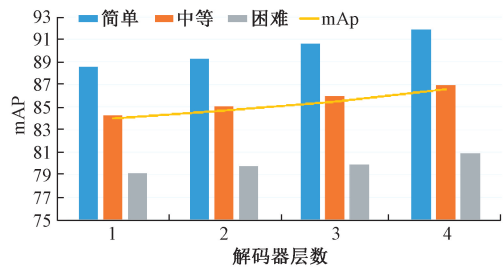


图 13 Kitti 验证集上解码器层数的比较

Fig. 13 Comparison of the number of decoder layers based on the Kitti validation set

验证多模态特征交叉注意力模块各部分的有效性,如表 10 所示。实验 1 是采用了一种原始朴素融合方法,该方法将像机的特征与非空体素相关联,并通过求和将它们与激光雷达的体素特征进行融合。实验 2、3 以此添

3D 网格注意力机制和注意力融合机制,结果表明,多模态特征交叉注意力模块各部分都起到改善作用。

表 10　多模态特征交叉注意力模块的消融研究

Table 10　The ablation study of the multimodal feature cross-attention module

实验	3D 网格融合	注意力融合	Car AP_{3D}/%		
			简单	中等	困难
1			91.75	86.96	80.56
2	√		91.94	87.35	80.87
3	√	√	92.35	87.89	81.41

为了弄清楚在什么情况下本文的方法改善基线最多,在不同的距离和不同的遮挡程度上评估本文的 TFIENet。如表 11 所示,远距离和严重遮挡的对象得到了最大的改善,这验证了本文的假设,即伪点云有助于提高对稀疏的原始点目标的检测能力。

表 11　在不同距离和不同遮挡程度上的性能

Table 11　Performance at different distances and at different levels of occlusion

TFIENet	距离/m			遮挡程度		
	0~20	20~40	40 以上	0	1	2
无	91.51	75.24	14.67	62.56	59.63	56.87
有	92.26	77.59	21.41	62.24	61.78	62.28
提升	+0.75	+2.35	+6.74	+0.68	+2.15	+5.41

3　结　　论

本文提出了一种新颖有效的多模态三维目标检测方法-TFIENet。设计了 Transformer 双域融合特征区域建议模块,通过双域特征查询与注意融合,有效提高不同域间的特征融合。设计了深度特征信息增强模块,利用深度补全机制预测出密集的深度特征信息,并提取相应的密集深度信息和特征语义信息来完成框的细化。该方法不仅可以获得可靠的深度信息,而且可以获得更丰富的细粒度信息。设计了多模态特征交叉注意力模块,采用动态交叉注意力机制来获得不同模态间的相关性,将来自不同模态的特征信息有效地对齐融合。本文在多种数据集上进行的实验证实,所提出的目标检测方法 TFIENet 能够有效提高目标检测精度。

参考文献

[1] HU J S K, KUAI Y, WASLANDER S L. Point density-aware voxels for LiDAR 3D object detection[C]. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022:8459-8468.

[2] ZHANG Y, HU Q, XU G, 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), 2022: 18931-18940.

[3] 童小钟,魏俊宇,苏绍璟,等. 融合注意力和多尺度特征的典型水面小目标检测[J]. 仪器仪表学报,2023,44(1):212-222.
TONG X ZH, WEI J Y, SU SH J, et al. Typical water surface small target detection by integrating attention and multi-scale features[J]. Chinese Journal of Scientific Instrument, 2023, 44(1): 212-222.

[4] 石欣,卢灏,秦鹏杰, 等. 一种远距离行人小目标检测方法[J]. 仪器仪表学报,2022,43(5):136-146.
SHI X, LU H, QIN P J, et al. A long range pedestrian small target detection method[J]. Chinese Journal of Scientific Instrument, 2022,43(5): 136-146.

[5] 闫钧华,张琨,施天俊, 等. 融合多层级特征的遥感图像地面弱小目标检测[J]. 仪器仪表学报, 2022, 43(3):221-229.
YAN J H, ZHANG K, SHI T J, et al. Ground weak and small target detection in remote sensing images fused with multi-level features[J]. Journal of Instrumentation and Meters, 2022,43(3): 221-229.

[6] 曹杰程,陶重犇. 基于 Stereo RCNN 的锚引导 3D 目标检测算法[J]. 仪器仪表学报,2021,42(12):191-201.
CAO J CH, TAO C B. An anchor-guided 3D target detection algorithm based on stereo RCNN.[J]. Chinese Journal of Scientific Instrument,2021,42(12):191-201.

[7] WANG C, MA C, ZHU M, et al. Pointaugmenting: Cross-modal augmentation for 3d object detection[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021:11794-11803.

[8] WU X P, PENG L, YANG H H, 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), 2022:5408-5417.

[9] JIANG W, ZHOU W, HU H. Double-stream position learning transformer network for image captioning[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2022, DOI: 10.1109/TCSVT.2022.3181490.

[10] MA C, SUN H, RAO Y, et al. Video saliency forecasting transformer [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2022, DOI:

10. 1109/TCSVT. 2022. 3172971.

[11] Li Y W, WU A W, MENG T J, et al. DeepFusion: Lidar-camera deep fusion for multi-modal 3D object detection[C]. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022: 7161-17170.

[12] BAI X Y, HU Z Y, HUANG Q Q, 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), 2022: 1080-1089.

[13] QI C R, YI L, SU H, et al. Pointnet++: Deep hierarchical feature learning on point sets in a metric space [J]. Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017.

[14] ZHU X, SU W, LU L, et al. Deformable DETR: Deformable transformers for end-to-end object detection[C]. International Conference on Learning Representations, 2021.

[15] GUIZILINI V, AMBRU 瘙塁 R, BURGARD W, et al. Sparse auxiliary networks for unified monocular depth prediction and completion [C]. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021: 11073-11083.

[16] FAN L, XIONG X, WANG F, et al. RangeDet: In defense of range view for LiDAR-based 3D object detection[C]. 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 2021: 2898-2907.

[17] 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, 2012: 3354-3361.

[18] CAESAR H, BANKITI V, LANG A H, et al. nuScenes: A multimodal dataset for autonomous driving [C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 11621-11631.

[19] SUN P, KRETZSCHMAR H, DOTIWALLA X, et al. Scalability in perception for autonomous driving: Waymo open dataset[C]. CVPR, 2020: 2446-2454.

[20] Shi S S, Guo C X, Jiang L, et al. PV-RCNN: Point-voxel feature set abstraction for 3d object detection[C]. CVPR, 2020: 10529-10538.

[21] YANG Z T, SUN Y N, LIU S, et al. 3DSSD: Point-based 3d single stage object detector[C]. CVPR, 2020: 11040-11048.

[22] DENG J, SHI S, LI P, et al. Voxel R-CNN: Towards high performance voxel-based 3D object detection[C]. National Conference on Artificial Intelligence, 2021.

[23] ZHENG W, TANG W L, JIANG L, et al. Se-SSD: Self-ensembling single-stage object detector from point cloud[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021: 14494-14503.

[24] YOO J H, KIM Y, KIM J S, et al. 3D-CVF: Generating joint camera and LiDAR features using cross-view spatial feature fusion for 3D object detection[C]. ECCV, 2020.

[25] VORA S, LANG A H, HELOU B, et al. Point painting: Sequential fusion for 3d object detection [C]. CVPR, 2020: 4604-4612.

[26] YIN T W, ZHOU X Y, KRAHENBUHL P. Center-based 3d object detection and tracking[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021: 11784-11793.

[27] 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.

[28] LI Y, QI X, CHEN Y, et al. Voxel field fusion for 3d object detection [C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 1120-1129.

作者简介

金宇锋, 2021 年于苏州科技大学获得学士学位,现为苏州科技大学硕士研究生,主要研究方向为目标检测和机器视觉。
E-mail: jinyufeng0315@ outlook. com

Jin Yufeng received his B. Sc. degree from Suzhou University of Science and Technology in 2021. He is currently a M. Sc. candidate at Suzhou University of Science and Technology. His main research interests include object detection and machine vision.

陶重犇(通信作者),2014 年于江南大学获得博士学位,现为苏州科技大学教授,清华大学苏州汽车研究院博士后,主要研究方向为三维语义建图、先进机器人和自动化。
E-mail: tom1tao@ 163. com

Tao Chongben (Corresponding author) received his Ph. D. degree from Jiangnan University in 2014. He is currently a professor at Suzhou University of Science and Technology. And he also is a postdoctoral fellow of Suzhou Automobile Research Institute of Tsinghua University. His main research interests include 3D semntic mapping, advanced robotics and automation.