DOI: 10.19650/ j.cnki.cjsi.J2514141

基于 Transformer 的特征频率解耦融合三维目标检测方法*

李明光[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.清华大学苏州汽车研究院 苏州 215134)

摘 要:针对现有多模态方法在融合过程中普遍依赖多尺度空间特征堆叠,导致特征频率信息耦合,从而限制三维目标检测精度提升的问题,提出了一种基于 Transformer 的特征频率解耦融合三维目标检测方法。首先,对输入图像进行小波频域特征解耦,通过离散小波变换对特征图进行多频率分解,通过独立的高低频特征金字塔获得高低频图像特征。其次,设计非对称频率更新编码块,将高频图像特征作为主要特征,采用自适应动态窗口编码进行高频率更新以增强边缘与纹理特征表达。同时,引入稀疏化的可变形注意力替代传统注意力层进行低频特征更新,实现不同频段特征的高效编码与协同优化。然后,构建高频引导体素融合模块,将小波解耦后的多尺度高频特征经视锥投影映射至三维体素空间,结合自适应半径采样方法,对稀疏点云局部结构进行高效补充,提取关键点体素特征。最后,将体素与图像特征统一映射到二维空间中,设计区域偏移 Transformer,利用注意力机制实现跨模态特征融合。在开放数据集 nuScenes、KITTI 和 Waymo 上评估该方法,在 nuScenes 测试集上实现了 73.2% 的 mAP 和 74.3% 的 NDS,尤其在小目标与远距离检测中表现突出。同时,该方法在实车平台上的测试表明,在复杂多变的实际环境中仍拥有较高的检测精度。

关键词: 三维目标检测;频率解耦;多模态融合;Transformer;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

Feature frequency decoupling and fusion for 3D object detection with Transformer

Li Mingguang [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: Existing multimodal 3D object detection methods often rely on multi-scale spatial feature stacking, which leads to frequency information entanglement and limits detection accuracy. To address this issue, this paper proposes a Transformer-based method for feature frequency decoupling and fusion. Firstly, the input images are processed using discrete wavelet transform for multi-frequency decomposition. Separate high-frequency and low-frequency feature pyramids are constructed to capture detailed local textures and global structural semantics, respectively. Then, an asymmetric frequency update encoder is designed, where high-frequency features are treated as the primary components and updated through adaptive dynamic window encoding to enhance edge and texture representation. Meanwhile, sparse deformable attention is introduced to replace standard attention mechanisms for efficient low-frequency feature updating, enabling coordinated encoding across different frequency bands. A high-frequency guided voxel fusion module is further proposed, where multi-scale high-frequency features are projected into 3D voxel space via frustum-based mapping. Combined with an adaptive radius sampling strategy, this module effectively supplements the local structure of sparse point clouds and extracts critical voxel-level features. Finally, voxel and image features are unified in the bird′s-eye view space. A region-shift Transformer module is introduced to enhance cross-modal feature fusion using attention mechanisms. The proposed method is evaluated on the nuScenes, KITTI, and Waymo datasets. The method achieves 73.2% mAP and 74.3% NDS on the nuScenes test set, demonstrating strong performance in detecting small and distant objects. Moreover, real-vehicle experiments indicate that the method maintains high detection

收稿日期:2025-06-10 Received Date: 2025-06-10
*基金项目:国家自然科学基金项目(62472300)资助

accuracy in complex and dynamic environments.

Keywords: 3D object detection; frequency decoupling; sensor fusion; Transformer; autonomous driving

0 引 言

三维目标检测作为环境感知的关键技术,在自动驾驶、智能机器人等领域中发挥着核心作用,不仅关系到目标的精准识别与定位,还直接影响系统的决策与安全性,具有重要的研究价值和广阔的实际应用前景。作为自动驾驶系统中的核心构件,传感器对环境感知的准确性至关重要。目前,激光雷达(light detection and ranging, LiDAR)与相机是最常用的两类传感器[1],分别具备独特的性能优势,但也存在感知局限性。激光雷达可提供高精度的三维空间信息,具备良好的深度测量能力,但在分辨率与纹理表达方面较弱,且成本较高;相机则能够捕捉丰富的颜色与纹理特征,具有较高的分辨率和成本优势,但对光照变化敏感,缺乏准确的深度感知能力。

单一模态的检测方法[2-4]难以同时满足高精度空间定位与丰富语义理解的需求,促使多模态融合方法[5-8]成为提升三维目标检测性能的重要途径。由于两类传感器具有互补性,这些方法融合了来自 LiDAR 和相机的数据,利用图像提供的丰富纹理信息和 LiDAR 提供的精确深度信息,实现更为准确的三维目标检测。然而,由于 LiDAR 和相机的固有特性,有效整合这两种模态的表示仍然是一个研究难题。文献[5]通过双域特征建议网络与多模态交叉注意力机制,实现图像与点云的有效对齐与融合。董钰婷等[6]则进一步提出自适应加权方法,通过注意力融合多尺度图像语义与动态体素点云特征。而 Liu 等[7]和 Liang 等[8]则提出两种 BEVFusion 框架将相机图像和 LiDAR 点云数据融合到一个统一的鸟瞰视图(bird's-eye view, BEV)[9-10]中,分别通过特殊 BEV 池化与注意力机制实现了更准确的目标检测。

然而,上述多模态融合方法通常侧重于空间域或体素域内多层次特征的融合,依赖于特征金字塔网络(feature pyramid network, FPN)或跨尺度卷积结构,以增强不同尺度的语义表达能力。这些方法主要关注特征在网络深度维度的层次性,而忽略了图像与点云特征在频率域中的本质差异[11]。图像特征富含高频成分,如边缘、纹理与细节变化,同时也包含大量低频的全局背景信息;而点云则更偏向稀疏但准确的低频几何结构表达。在未进行频率分离的情况下直接融合,往往导致高频细节在低频特征主导的融合过程中被削弱,使得模型在处理小目标或边缘复杂区域时感知能力下降。此外,频率信息耦合还加重了融合冗余,降低了特征的判别性。有效的融合策略应考虑在频率域实现高低频特征的解耦与选择性整合,而不仅仅依赖多层特征的堆叠。

基于上述研究,提出了一种基于 Transformer 的特征频率解耦融合三维目标检测方法。使用 Haar 小波构建离散小波变换对图像特征图进行多层次的频率分解,分离高频局部细节与低频全局结构信息,分别构建高频与低频特征金字塔以增强多频特征表达。构建非对称频率更新机制,实现高频特征的动态增强与低频特征的稳定调控。设计高频引导体素融合模块通过高频图像特征与视锥投影机制对点云稀疏区域进行局部信息补全。在公开数据集和实际车辆平台上评估本文方法,均取得优异效果。

1 方法介绍

方法完整结构如图 1 所示,方法主要包括 3 个阶段:首先,分别提取点云与图像特征。采用基于离散小波变换的方法将多尺度图像特征解耦为高频与低频特征,分别构建独立的特征金字塔网络建模不同建频率特征之间的跨尺度信息,构建非对称频率更新机制获得图像特征输出序列。设计图像高频特征引导模块促使体素-像素进行先验融合,随后,采用动态采样体素模块增强融合后体素特征的空间聚合能力,并将体素投影至 BEV 平面。最后,在 BEV 空间中整合点云与图像特征,并送入后续的检测头进行测试。

1.1 小波频域特征解耦

在现有图像特征金字塔网络中,感受野通常通过连续的下采样操作逐层放大。然而,这种机制虽然能够捕获大尺度上下文信息,但也会导致底层细节信息的丢失,这是由于空间分辨率不断降低,浅层网络对局部纹理与边缘信息的感知能力削弱。

本文提出基于离散小波变换(discrete wavelet transform, DWT)[12]的小波频域特征解耦模块,通过多频率分解与特征重构机制,在提升模型对低频全局特征响应能力的同时,有效保留底层空间定位信息,实现高效而精确的频域特征分离。具体表示为,对经过骨干网络处理后得到的多尺度特征 $X \in \mathrm{R}^{H\times W\times C}$ 进行多层离散 Haar 小波分解,获取低频和高频特征,如式(1)所示。

$$\begin{cases} X_{LL}^{(1)}, X_{LH}^{(1)}, X_{HL}^{(1)}, X_{HH}^{(1)} = WT(X) \\ X_{LL}^{(i)}, X_{LH}^{(i)}, X_{HL}^{(i)}, X_{HH}^{(i)} = WT(X_{LL}^{(i-1)}) \end{cases} \tag{1}$$

其中,$X_{LL}^{(i)}$ 为第 i 层的低频特征,包含图像结构信息。$\{X_{LH}^{(i)}, X_{HL}^{(i)}, X_{HH}^{(i)}\}$ 分别为水平方向、垂直方向、对角方向的高频特征。在 Haar 小波分解过程中,每个特征图的像素(x,y)计算如式(2)所示。

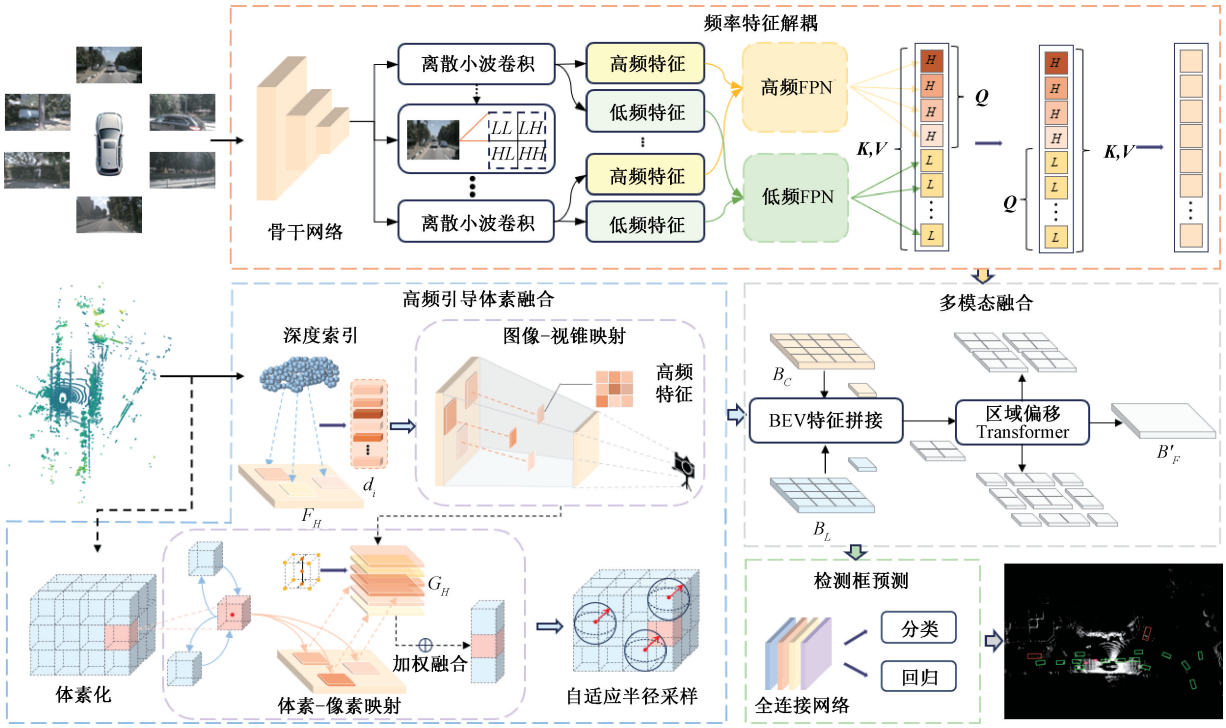


图 1　整体框架

Fig. 1　The framework of our proposed method

$$\begin{cases} X_{LL}[x,y] = \sum_m \sum_n X[m,n] \cdot g[2x-m] \cdot g[2y-n] \\ X_{LH}[x,y] = \sum_m \sum_n X[m,n] \cdot g[2x-m] \cdot h[2y-n] \\ X_{HL}[x,y] = \sum_m \sum_n X[m,n] \cdot h[2x-m] \cdot g[2y-n] \\ X_{HH}[x,y] = \sum_m \sum_n X[m,n] \cdot h[2x-m] \cdot h[2y-n] \end{cases} \tag{2}$$

其中,在 Haar 小波中,使用 $g=[1/\sqrt{2},1/\sqrt{2}]$ 提取特征的低频分量,$h=[1/\sqrt{2},-1/\sqrt{2}]$ 提取特征的高频分量。m 和 n 是用于二维卷积操作的求和索引变量,m 是输入特征图 X 的纵向卷积索引,n 是输入特征图 X 的横向卷积索引。接着,如式(3)所示,对每一层的低频和高频分量进行小核卷积,即:

$$\begin{cases} Y_{LL}^{(i)} = Conv(W_{LL}^{(i)}, X_{LL}^{(i)}) \\ Y_{LH}^{(i)} = Conv(W_{LH}^{(i)}, X_{LH}^{(i)}) \\ Y_{HL}^{(i)} = Conv(W_{HL}^{(i)}, X_{HL}^{(i)}) \\ Y_{HH}^{(i)} = Conv(W_{HH}^{(i)}, X_{HH}^{(i)}) \end{cases} \tag{3}$$

其中,$W_{LL}^{(i)}$ 为第 i 层小波分解后的低频特征 $X_{LL}^{(i)}$ 对应的卷积核权重,$\{W_{LH}^{(i)}, W_{HL}^{(i)}, W_{HH}^{(i)}\}$ 为第 i 层小波分解后的高频特征 $\{X_{LH}^{(i)}, X_{HL}^{(i)}, X_{HH}^{(i)}\}$ 对应的卷积核权重。

为充分挖掘多层次频率域特征,将不同层级的小波分解特征分别送入特征金字塔网络进行频率感知跨尺度融合,并通过横向连接机制实现跨层次信息传递。具体而言,将低频特征 $Y_{LL}^{(i)}$ 作为语义主干分支,输入低频特征金字塔(low-frequency FPN, L-FPN)以保留图像整体结构与语义轮廓;高频特征 $\{Y_{LH}^{(i)}, Y_{HL}^{(i)}, Y_{HH}^{(i)}\}$ 合并为高频特征集合 $Y_H^{(i)}$,输入高频特征金字塔(high-frequency FPN, H-FPN),用于增强边缘、纹理与局部细节表达能力。上述过程表示如式(4)所示。

$$\begin{cases} F_L = L\text{-}FPN(\{Y_{LL}^{(i)}\}_{i=1}^{N}) \\ F_H = H\text{-}FPN(\{Y_{LH}^{(i)}, Y_{HL}^{(i)}, Y_{HH}^{(i)}\}_{i=1}^{N}) \end{cases} \tag{4}$$

其中,F_L 和 F_H 分别为低频与高频特征经 FPN 融合后的多尺度频率感知特征。

1.2　非对称频率更新

受频率域特征表达需求的启发,传统高效编码器在处理多频率图像特征时存在两方面瓶颈:一方面,高频特征中的局部细节信息未能被充分利用,导致模型对细粒度变化响应不足;另一方面,低频特征作为稳定的全局语义支撑,与高频同步更新会增加计算负担。因此,提出了非对称频率更新策略,针对得到的高低频特征 F_L 与 F_H,动态调整其编码更新频次,以实现精度与效率的平衡。

为突出高频特征对局部信息的动态响应能力,F_H 将被作为主要特征进行高频率更新,F_L 更新频率则较低。如图 2 所示,将 F_L 与 F_H 进行拼接得到 F_C,并将更新编码器块堆叠 A 次,其中每个块更新 B 次高频特征,在块尾

更新一次低频特征。整个过程表示为：

$$\begin{cases} F'_H = HUpdate^B(F_H,[F_H,F_L]) \\ F'_L = LUpdate^1(F_L,[F'_H,F_L]) \end{cases} \tag{5}$$

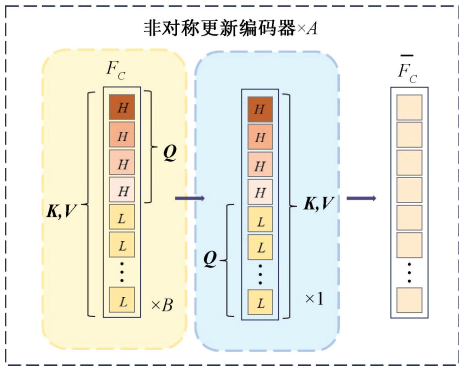


图 2 非对称频率更新示意图

Fig. 2 Schematic diagram of asymmetric frequency update encoder

1)高频更新编码器

高频特征包含丰富边缘、纹理、细节信息，对于实现精细化的目标检测与定位至关重要。为此，提出一种基于自适应动态窗口的高频特征编码机制(high frequency update encoder, HUpdate)，通过精准的局部区域关注与特征聚合，显著提升对小目标与细节区域的特征建模能力。为精确捕获高频特征图中的目标敏感区域，首先定义一组目标候选区域集合 $B = \{B_1,B_2,\cdots,B_m\}$，对于第 i 个目标候选区域 B_i，将其表示为 $B_i = \{x_i,y_i,w_i,h_i\}$。其中，$(x_i,y_i)$ 表示候选区域的中心坐标，(w_i,h_i) 为宽度和高度。

由于目标位置和尺寸往往存在一定的不确定性，进一步提出了一种动态窗口偏移策略，通过学习一个偏移比例 α，对每个候选区域自动扩展，以覆盖更多的局部上下文信息：

$$[x'_i,y'_i,w'_i,h'_i] = [x_i,y_i,w_i,h_i] \times (1+\alpha) \tag{6}$$

这种动态窗口的生成方式，能够自适应地根据目标尺寸大小调整窗口覆盖范围，确保重要的细节特征不被遗漏，提升了局部特征的感知效果。对每个动态窗口区域，提取其对应的局部特征，即：

$$F_{W_i} = F_H[x'_i:x'_i + w'_i, y'_i:y'_i + h'_i] \tag{7}$$

为了强化窗口区域内部特征的联系，充分挖掘高频特征所蕴含的局部结构信息，在每个局部窗口 F_{W_i} 内部实施自注意力，对窗口内的特征进行线性变换生成对应的查询($\boldsymbol{Q}$)、键($\boldsymbol{K}$)、值($\boldsymbol{V}$)，然后在局部窗口内计算自注意力得分矩阵，并对特征进行重新加权，即：

$$Attn(F_{W_i}) = \text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d}}\right)\boldsymbol{V} \tag{8}$$

由于这些窗口覆盖的是高频特征中不同的位置与尺度，为了实现跨区域的信息整合与特征补充，将所有窗口处理后的特征进行通道维度的拼接，即：

$$F'_H = \text{Concat}(\text{Attn}(F_{W_1}),\text{Attn}(F_{W_2}),\cdots,\text{Attn}(F_{W_m})) \tag{9}$$

2)低频更新编码器

低频特征主要承载图像的全局语义和背景信息，通常较为稳定，多次更新的变化不显著，同步高频更新会增加计算负担，为此，设计低频更新编码器(low frequency update encoder, LUpdate)，只在块尾进行一次更新。为了高效增强低频特征更新，提出一种稀疏化的可变形注意力(sparse deformable attention, SDA)替代传统注意力层，在进行可变形注意力操作时，每个查询($\boldsymbol{Q}$)从特征图的相应位置同时提取键($\boldsymbol{K}$)和值($\boldsymbol{V}$)。接着，通过将查询与这些位置的键进行对比，能够获得更加稳定和准确的注意力分数。相比标准的可变形注意力，这种方法不仅拓展了采样策略的灵活性，也可以视为稀疏化的密集注意力形式。

具体而言，在标准的可变形注意力层中，查询 $\boldsymbol{Q}$ 会被划分为 M 个注意力头，每个头从 L 个不同尺度的特征图中采样 K 个点，构成总的采样数 $N_v = M \times L \times K$。采样位置通过查询预测的偏移 Δp 生成，并从多尺度特征 F 中使用双线性插值得到采样特征，即：

$$\begin{cases} \Delta p = \boldsymbol{Q} \times \boldsymbol{W}_{\Delta} \\ \boldsymbol{V} = \text{Interp}(F,p+\Delta p) \cdot \boldsymbol{W}_v \\ \text{DefAttn}(\boldsymbol{Q},\boldsymbol{V}) = \text{Softmax}(\boldsymbol{Q} \times \boldsymbol{W}_a) \cdot \boldsymbol{V} \end{cases} \tag{10}$$

其中，$\boldsymbol{W}_{\Delta}$ 与 $\boldsymbol{W}_a \in \mathrm{R}^{C\times N_v}$、$\boldsymbol{W}_v \in \mathrm{R}^{C\times C}$ 都为投影的参数矩阵。

然而，标准机制忽视了查询与键之间的显式关系。如图 3 所示，为补偿标准机制中忽略的键信息，并确保跨尺度特征更新中的位置一致性，SDA 加入了键的采样过程，如式(11)所示。

$$\begin{cases} \boldsymbol{K} = \text{Interp}(F,p+\Delta p) \cdot \boldsymbol{W}_k \\ \boldsymbol{V} = \text{Interp}(F,p+\Delta p) \cdot \boldsymbol{W}_v \\ F'_L = SDA(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \text{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d}}\right)\boldsymbol{V} \end{cases} \tag{11}$$

其中，d 为键的特征维度。该机制通过结合键的信息，使得查询不仅基于自身特征进行注意力分配，更能精确对齐采样位置的局部特征。

最后，将更新后的高频特征 F'_H 与更新后的低频特征 F'_L 进行融合，形成完整图像特征 $\overline{F_C}$ 用于后续融合，即：

$$\overline{F_C} = \text{Concat}(F'_H,F'_L) \tag{12}$$

1.3 高频引导体素融合

图像特征蕴含丰富的频率成分，其中低频信息侧重于描述全局背景与整体结构，而高频信息则集中体现于

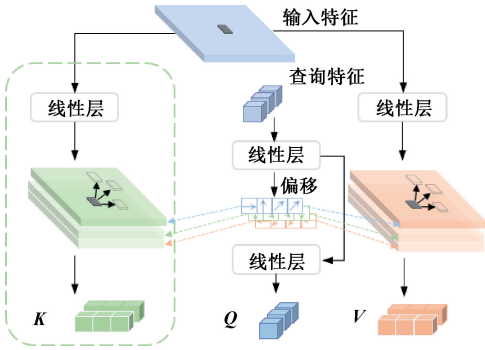


图 3　稀疏化可变形注意力

Fig. 3　Sparse deformable attention(SDA)

物体的边缘轮廓、纹理变化等局部精细特征。相比之下,点云数据由于其固有的稀疏性,尤其在远距离或遮挡区域,局部细节呈现能力有限,难以捕捉完整的几何边界信息。传统方法在融合图像与点云时,直接使用完整图像特征参与体素对齐,往往引入大量冗余的低频背景内容,不仅增加了融合计算的负担,也未能充分激发点云局部特征的表达潜力。为此,提出一种基于高频图像特征引导的体素先验融合方法,利用小波分解精确提取图像的高频细节,通过视锥投影机制将其映射至三维体素空间,实现对稀疏点云局部结构的高效补充,从而显著增强融合特征的细粒度感知能力与模型的局部解析能力。

首先,将输入的点云数据以预定义空间分辨率大小($L\times H\times W$)均匀划分为多个体素,并使用非空体素中所有点的坐标的平均值及其反射强度作为该体素的特征。基于体素空间坐标与图像高频像素坐标之间的几何映射关系,设计了一种基于离散深度估计的视锥投影机制,通过该机制在图像像素与体素网格之间建立精确的关联,并借助相机内参矩阵完成空间转换,从而实现高频图像特征与点云体素的深度对齐。

1)视锥投影机制

该机制旨在将图像域的二维特征准确映射到三维空间中,与体素化后的点云特征进行空间对齐与联合建模,从而实现跨模态特征的高效交互。通过像素深度分布引导图像特征填充到视锥结构中,再与体素特征按空间一致性对齐。本文在此基础上,引入了基于高频图像特征的视锥变换策略,增强了局部特征的对齐精度与空间一致性。对于小波分解得到的图像高频特征 $F_H(x,y)\in \mathrm{R}^{W\times H\times C}$,基于点云数据生成深度图 $D_H(x,y)\in \mathrm{R}^{W\times H\times d}$,从而为视锥映射提供深度信息支持。具体而言,首先通过相机标定矩阵 $\boldsymbol{K}$ 将同步获取的点云投影到图像平面得到每个像素 (x,y) 的深度值。该深度值被离散化为多个深度索引 d_i,并作为图像高频特征在深度维度的填充依据,由此,高频特征直接映射成视锥特征张量 $\boldsymbol{G}_H\in \mathrm{R}^{W\times H\times d\times C}$,形式化表示为:

$$\boldsymbol{G}_H(x,y,d_i)=\begin{cases}F_H(x,y), & D_H(x,y)\in d_i\\ 0, & \text{其他}\end{cases} \tag{13}$$

接着,对于每个体素中心 $s_v^i=(x_i,y_i,z_i)$,根据其空间位置通过相机内参投影至图像平面(x_i,y_i),并查找其在视锥特征 $\boldsymbol{G}_H$ 中对应的高频特征分量。为了保证空间对齐的连续性,采用三线性插值从 $\boldsymbol{G}_H$ 中采样(x_i,y_i,d_i) 位置的特征值,并将该高频特征与体素特征 $V(x_i,y_i,z_i)$ 进行加权融合,即:

$$V'(x_i,y_i,z_i)=V(x_i,y_i,z_i)+\psi(G_H(x_i,y_i,d_i)) \tag{14}$$

其中,$\psi(\cdot)$ 表示特征通道匹配与归一化操作。通过该机制,实现了图像高频细节信息对稀疏点云局部特征的有效补偿,显著增强跨模态融合的细粒度感知能力。

2)自适应半径采样

由于点云的空间分布具有天然的不均匀性,固定邻域范围的特征提取难以兼顾不同区域的特征需求,对融合体素 V' 进行自适应半径采样,通过动态调整邻域采样范围,有效聚焦于局部几何特征显著区域。

具体而言,首先使用稀疏卷积网络对相应体素特征进行处理,然后对输出体素特征进行最远点采样从中选取 N_p 个关键点$\{p_i\}_{i=1}^{N_p}$,根据每个关键点的融合体素特征动态调整邻域采样半径。对于每个关键点 p_i,从融合体素中选取满足 $\|v_j-p_i\|^2<r_i^2$ 的邻域体素 v_j,并将其对应特征 f_j 构成邻域特征集合 S_i,即:

$$S_i=\{\boldsymbol{f}_j:\ \|v_j-p_i\|^2<r_i^2, v_j\in V', \boldsymbol{f}_j\in F'_V\} \tag{15}$$

其中,V' 为融合体素的空间位置集合,$\boldsymbol{f}_j$ 为第 j 个关键点对应的融合体素特征向量。F'_V 为融合体素特征集合,v_j 第 j 个体素在三维空间中的几何中心坐标。r_i 为关键点 p_i 的自适应采样半径,计算如式(16)所示。

$$r_i=\sum_{k=1}^{K}\alpha_k(\boldsymbol{f}_i)\cdot r_k \tag{16}$$

其中,$\{r_k\}_{k=1}^{K}$ 为预设的半径集合,$\alpha_k(\boldsymbol{f}_i)$ 为 softmax 输出的权重,满足 $\sum_{k=1}^{K}\alpha_k(\boldsymbol{f}_i)=1$。

确定邻域体素集合 S_i 后,进行特征聚合,提取局部上下文信息的全局表示,即:

$$f_i^{out}=MLP(\mathrm{MaxPool}(\{\boldsymbol{f}_j|\boldsymbol{f}_j\in S_i\})) \tag{17}$$

最终,得到增强后的关键点特征 f_i^{out} 不仅包含原始体素的三维几何信息,还融合了图像高频特征带来的局部细节信息。

1.4　多模态特征融合

BEV 空间为点云与图像的异构数据融合建立了统一的空间,对关键点体素特征 f_i^{out} 进行 8 次降采样,并沿 Z 轴压缩,得到体素特征的 BEV 特征 B_L。对于图像特征 $\overline{F_C}$,采用 LSS(lift, splat, shoot)方法生成尺寸为 $NHWD$ 的相机特征点云,其中 N 是相机数量(H,W) 是相机特征

图的尺寸,相机点云沿 x、y 轴以步长 r 进行量化,然后使用 BEV 池化操作在每个 $r\times r$ 的 BEV 网格内聚合所有特征,并沿 z 轴扁平化处理得到图像 BEV 特征 B_C。如算法 1 所示,将 B_L 与 B_C 结合,计算出多模态 BEV 特征 B_F。

算法 1:多模态融合算法

输入: $B_C \in \mathrm{R}^{W\times H\times C}$, $B_L \in \mathrm{R}^{W\times H\times C}$

输出: $B'_F \in \mathrm{R}^{W\times H\times C}$

对于每个样本 (B_C, B_L) **做**

$B_F \leftarrow f_{\mathrm{conv}}([B_L, B_C]) \in \mathrm{R}^{W\times H\times C}$ # BEV 特征投影

/* **区域内特征增强** */

$\{g^1_{grid}, g^2_{grid}, \cdots\} \leftarrow B_F$ # BEV 特征网格化

区域构建

$\{g^1_{\mathrm{grid}}, g^2_{\mathrm{grid}}, \cdots, g^{r^2}_{\mathrm{grid}}\} \leftarrow \mathrm{region}(\{g^1_{grid}, g^2_{grid}, \cdots\}, r\times r)$

区域内自注意力机制

$\{\hat{g}^1_{\mathrm{grid}}, \hat{g}^2_{\mathrm{grid}}, \cdots, \hat{g}^{r^2}_{\mathrm{grid}}\} \leftarrow f_{\mathrm{MSA}}(\{g^1_{\mathrm{grid}}, g^2_{\mathrm{grid}}, \cdots, g^{r^2}_{\mathrm{grid}}\})$

/* **区域间特征增强** */

#按偏移量($r/2$,$r/2$)构建新区域

$\{\hat{g}^{\triangle 1}_{\mathrm{grid}}, \hat{g}^{\triangle 2}_{\mathrm{grid}}, \cdots, \hat{g}^{\triangle r^2}_{\mathrm{grid}}\} \leftarrow shift\{\hat{g}^1_{\mathrm{grid}}, \hat{g}^2_{\mathrm{grid}}, \cdots, \hat{g}^{r^2}_{\mathrm{grid}}\}$

区域间自注意力机制

$\{g^{\triangle 1}_{\mathrm{grid}}, g^{\triangle 2}_{\mathrm{grid}}, \cdots, g^{\triangle r^2}_{\mathrm{grid}}\} \leftarrow f_{\mathrm{MSA}}(\{\hat{g}^{\triangle 1}_{\mathrm{grid}}, \hat{g}^{\triangle 2}_{\mathrm{grid}}, \cdots, \hat{g}^{\triangle r^2}_{\mathrm{grid}}\})$

$B'_F \leftarrow reshape(\{g^{\triangle 1}_{\mathrm{grid}}, g^{\triangle 2}_{\mathrm{grid}}, \cdots, g^{\triangle r^2}_{\mathrm{grid}}\})$

返回 B'_F

B_F 的计算公式如式(18)所示。

$$B_F = f_{\mathrm{conv}}([B_L, B_C]) \in \mathrm{R}^{W\times H\times C} \tag{18}$$

为进一步获得 BEV 特征全局的上下文信息,设计区域偏移 Transformer,将 B_F 网格化为$\{g^1_{\mathrm{grid}}, g^2_{\mathrm{grid}}, \cdots\}$,并送入自注意力模块中进行特征增强。采用区域划分策略降低计算开销,将网格特征分为若干 $r\times r$ 的子区域。把每个区域视为整体并在内部使用多头自注意力机制,实现内部信息交互,即:

$$\{\hat{g}^1_{\mathrm{grid}}, \hat{g}^2_{\mathrm{grid}}, \cdots, \hat{g}^{r^2}_{\mathrm{grid}}\} = f_{\mathrm{MSA}}(\{g^1_{\mathrm{grid}}, g^2_{\mathrm{grid}}, \cdots, g^{r^2}_{\mathrm{grid}}\}) \tag{19}$$

随后,将原 BEV 网格按偏移量($r/2$,$r/2$)滑动以重新划分子区域,并在区域中再次使用多头自注意力机制以建模区域间的特征交互。最后,将增强后的网格特征重新排列为完整的 BEV 特征图 B'_F,送入后续的检测头进行测试。

2 实验与分析

2.1 数据集

使用开放数据集 nuScenes[13]、KITTI[14] 和 Waymo[15] 评估所提方法的性能。相关信息介绍如表 1 所示,与数据集相关的性能评价指标介绍见后文。

表 1 nuScenes、KITTI 与 Waymo 数据集的信息介绍

Table 1 Introduction to the nuScenes, KITTI and Waymo datasets

数据集	图像尺寸	单帧点云数量	数据集划分	感知范围	检测类别数量	评价指标
nuScenes[13]	1 600×900	约 30 k 点	700 场景训练/150 场景验证/150 场景测试	360°环视	10 类	mAP/NDS
KITTI[14]	1 242×375	约 100 k 点	7 481 帧训练/7 518 帧测试	仅前视	3 类	mAP
Waymo[15]	1 920×1 280	约 200 k 点	800 场景训练/200 场景验证/100 场景测试	360°环视	5 类	mAP/mAPH

平均精度均值(mean average precision, mAP)是评估目标检测算法整体性能的核心指标。其计算过程包括对预测结果按置信度排序,绘制每类目标的精确率-召回率曲线,计算其下面积并对所有类别取平均。具体为:

$$mAP = \frac{1}{N}\sum_{i=1}^{N}\left(\int_0^1 p_i(r)\,\mathrm{d}r\right) \tag{20}$$

其中,$p_i(r)$ 表示第 i 类在召回率为 r 时的精确率函数,对于离散预测结果,常使用插值或分段平均方式近似计算面积。不同数据集在 mAP 的定义和计算方式上存在差异,主要体现在交并比(intersection over union, IoU)阈值设定、评估粒度、评价维度及是否引入额外惩罚因子等方面。

在 KITTI 中,mAP 指标通常基于 3D 边界框与真实框之间的 IoU 进行计算,即:

$$IoU = \frac{\text{预测框与真实框的交集面积}}{\text{预测框与真实框的并集面积}} \tag{21}$$

对于车辆类别,采用 $IoU\geqslant 0.7$,而对于行人和自行车类别,则采用 $IoU\geqslant 0.5$。

nuScenes 则是以检测框中心点距离为基础的 mAP 计算方法,而非 IoU,常用的距离阈值包括 0.5、1.0 和 2.0 m。此外,nuScenes 检测得分(nuScenes detection score, NDS)是其另一项综合评估指标,计算为:

$$NDS = \frac{1}{10}\left[5\cdot mAP + \sum_{mTP\in TP}(1-\min(1, mTP))\right] \tag{22}$$

其中,真阳性(truth positive, TP)包括位置误差(average translation error, ATE)、尺寸误差(average scale error, ASE)、朝向误差(average orientation error, AOE)、

速度误差(average velocity error, AVE)和属性误差(average attribute error, AAE)。mTP 为平均 TP 度量。

而在 Waymo 中,进一步引入了包含航向角的平均精度(mean average precision weighted by heading, mAPH),用于强化对目标朝向估计准确性的评估。该指标在标准平均精度的基础上,引入了方向正确性的加权项。具体计算如式(23)所示。

$$mAPH = \frac{1}{N}\sum_{i=1}^{N} APH_i \tag{23}$$

其中,N 为检测类别总数,APH_i 表示第 i 类的加权平均精度。对每一类目标,其 APH 由式(24)定义,即:

$$APH_i = \frac{1}{|D_i|}\sum_{j\in D_i}\delta_j \cdot AP_{ij} \tag{24}$$

其中,D_i 为第 i 类目标的检测样本集合,AP_{ij} 为样本 j 在该类下的平均精度值。δ_j 为方向准确性权重,定义为:

$$\delta_j = \begin{cases} 1, & |\theta_j^{\text{pred}} - \theta_j^{\text{gt}}| \leqslant \tau \\ 0, & \text{其他} \end{cases} \tag{25}$$

其中,θ_j^{pred} 和 θ_j^{gt} 分别为预测框与真实框的航向角,为设定的容差阈值。

2.2　实验设置

本文配置基于开源框架 MMDetection3D[16]。对于相机输入,采用 Swin Transformer[17] 作为主干网络。对于 nuScenes,体素尺寸为[0.075 m,0.075 m,0.2 m];对于 KITTI 和 Waymo,体素尺寸为[0.05 m,0.05 m,0.1 m]。对 LiDAR 和相机输入应用了一系列数据增强,以提高泛化能力。这些增强包括随机翻转([-5.4°,5.4°])、缩放([-0.38,0.55])、旋转 $\left(\left[-\frac{\pi}{4},\frac{\pi}{4}\right]\right)$ 和平移(标准差为 0.5)等。优化器使用 AdamW[18],最大学习率设置为 0.001,权重衰减设置为 0.01。为了减轻过拟合,对 3D 物体和点云进行了随机翻转。在实验过程中,不使用测试时间增强或推理过程中的多模型集成。

在对比实验中,涉及的其他基线方法均采用其原论文中公布的结果,未对其进行重新训练。其中,BEVFusion[7-8] 等方法使用与本文相同的数据增强配置,以保证实验条件的一致性与公平性。在实验流程中严格保持输入设置、预处理参数与评估标准一致,确保各方法在可比基础上进行性能对照。

2.3　nuScenes 数据集的实验结果

在表 2 中,将模型性能与 nuScenes 测试集中先进的方法进行比较,相比较于单激光雷达方法 TransFusion-L[19],该模型在 mAP 和 NDS 方面分别提升了 7.7% 和 4.1%。对于多模态融合方法,经过小波解耦后的高低频特征处理使得模型相较于 BEVFusion[7] 在 mAP 上提升了 3.0%,在 NDS 上提升了 1.4%。在 10 个类别的单独检测效果上,获得了 8 个类别的领先,特别是在小目标检测上,行人与路障的检测精度比最优效果分别提升了 0.5% 与 0.9%。

表 2　nuScenes 测试集上的三维目标检测性能

Table 2　3D Object Detection Performance on the nuScenes test set　(%)

方法	数据	mAP	NDS	汽车	卡车	建筑车辆	公交	拖车	障碍	摩托	自行车	行人	交通锥
Pointpillars[20]	点云	30.5	45.3	68.4	23.0	4.1	28.2	23.4	38.9	27.4	1.1	59.7	30.8
CenterPoint[21]	点云	60.3	67.3	85.2	53.5	20.0	63.6	56.0	71.1	59.5	30.7	84.6	78.4
TransFusion-L[19]	点云	65.5	70.2	86.2	56.7	28.2	66.3	58.8	78.2	68.3	44.2	86.1	82.0
MVP[22]	点云+图像	66.4	70.5	86.8	58.5	26.1	67.4	57.3	74.8	70.0	49.3	89.1	85.0
PointAugmenting[23]	点云+图像	66.8	71.0	87.5	57.3	28.0	65.2	60.7	72.6	74.3	50.9	87.9	83.6
TransFusion[19]	点云+图像	68.9	71.7	87.1	60.0	33.1	68.3	60.8	78.1	73.6	52.9	88.4	86.7
BEVFusion[8]	点云+图像	69.8	71.9	88.1	60.9	34.4	68.5	62.1	78.2	71.8	52.2	89.2	85.5
BEVFusion[7]	点云+图像	70.2	72.9	88.6	60.1	39.3	69.8	63.8	80.0	74.1	51.0	89.2	86.5
ObjectFusion[24]	点云+图像	71.0	73.3	89.4	59.0	**40.5**	71.8	63.1	76.6	78.1	53.2	90.7	**87.7**
本文	点云+图像	**73.2**	**74.3**	**89.5**	**62.3**	35.2	**73.9**	**68.4**	**80.9**	**83.7**	**62.1**	**91.2**	86.8

在 nuScenes 验证集上也有优秀的表现,如表 3 所示,实现了 72.6% 的 mAP 与 73.4% 的 NDS,同时保持了较高的计算效率。图 4 为模型在 nuScenes 验证集上的定性检测与可视化结果。

整体来看,本文所提方法在 nuScenes 数据集上在 mAP 与 NDS 两项指标上均取得领先,体现出其在大规模、多类别交通场景中的出色稳定性。这一性能优势主要得益于所引入的频率解耦机制,有效增强了模型对多尺度目标,尤其是小尺寸目标的检测能力。本文通过强化对边缘与细节信息的表达,使得模型在复杂背景下依然具备较强的目标识别能力,显著提升了整体检测精度。

表 3 nuScenes 验证集上的三维目标检测性能

Table 3 3D Object detection performance on the nuScenes validation set

方法	数据	*mAP*/%	*NDS*/%	延时/ms
CenterPoint[21]	点云	57.4	65.2	44.8
FUTR3d[25]	点云+图像	64.5	68.3	176.6
TransFusion[19]	点云+图像	67.5	71.3	86.5
BEVFusion[7]	点云+图像	68.3	71.1	66.2
本文	点云+图像	**72.6**	**73.4**	93.4

同时,使用精度召回曲线评估模型性能,采用 RTX 4090 GPU,以汽车类别为例,在 nuScenes 数据集上与部分方法进行比较,结果如图 5 所示,曲线表明所提出的方法在高召回率的情况下拥有较高的精度。

2.4 KITTI 数据集的实验结果

为充分验证模型有效性,本研究在 KITTI 数据集上也进行了测试。如表 4 所示,在 KITTI 测试集上,对 3D 检测的汽车、行人与自行车 3 个类别在“简单、中等、困难”3 个难度上进行检测,评估 40 个采样召回点的平均精度。

图 4 nuScenes 数据集上的可视化结果

Fig. 4 Visualization results on the nuScenes dataset

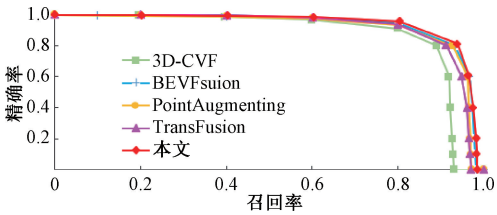


图 5 汽车类别精度召回曲线对比

Fig. 5 Comparison of precision recall curve for car

表 4 结果显示,相比较于 LoGoNet[36],特别是在行人和自行车类别上,得益于频率解耦后的特征处理,有效分离并增强了图像的高频细节信息,使得模型在遮挡严重、目标边界模糊或纹理复杂的区域中仍能保持较高的分辨能力。最终让 mAP 分别提升 1.01% 和 2.42%,可视化效果如图 6 所示。此外,非对称频率更新策略合理分配了计算资源,提升了特征编码效率,并避免了冗余信息对小目标检测的不利干扰。

表 4 KITTI 测试集上的性能比较

Table 4 Performance on the KITTI test set (%)

方法	数据	汽车(3D)				行人(3D)				自行车(3D)			
		简单	中等	困难	mAP	简单	中等	困难	mAP	简单	中等	困难	mAP
SECOND[26]	点云	83.34	72.55	65.82	73.90	–	–	–	–	71.33	52.08	45.83	56.41
PointPillars[20]	点云	82.58	74.31	68.99	75.29	51.45	41.92	38.89	44.09	77.10	58.65	51.92	62.56
PVRCNN[27]	点云	90.25	81.43	76.82	82.83	52.17	43.29	40.29	45.25	78.60	63.71	57.65	66.65
CASA[28]	点云	91.58	83.06	80.08	84.91	54.04	47.09	44.56	48.56	**87.91**	73.47	66.17	75.85
MV3D[29]	点云+图像	74.97	63.63	54.00	64.20	–	–	–	–	–	–	–	–
3D-CVF[30]	点云+图像	89.20	80.05	73.11	80.79	–	–	–	–	–	–	–	–
Fast-CLOCs[31]	点云+图像	89.11	80.34	76.98	82.14	52.10	42.72	39.08	44.63	82.83	65.31	57.43	68.52
EPNet[32]	点云+图像	89.81	79.28	74.59	81.23	–	–	–	–	–	–	–	–
SFD[33]	点云+图像	91.73	84.76	77.92	84.80	**55.31**	43.53	45.13	47.99	86.56	72.32	66.37	75.08
VoxelNextFusion[34]	点云+图像	90.40	82.03	79.86	84.11	52.56	45.72	41.85	46.71	79.28	64.47	58.25	67.33
LVP[35]	点云+图像	91.37	84.92	80.07	85.45	–	–	–	–	–	–	–	–
LoGoNet[36]	点云+图像	**91.80**	85.06	80.74	85.86	53.07	47.43	45.22	48.57	74.47	71.70	64.67	73.61
本文	点云+图像	91.65	**85.23**	**80.81**	**85.90**	54.81	**48.17**	**45.76**	**49.58**	87.49	**74.18**	**66.41**	**76.03**

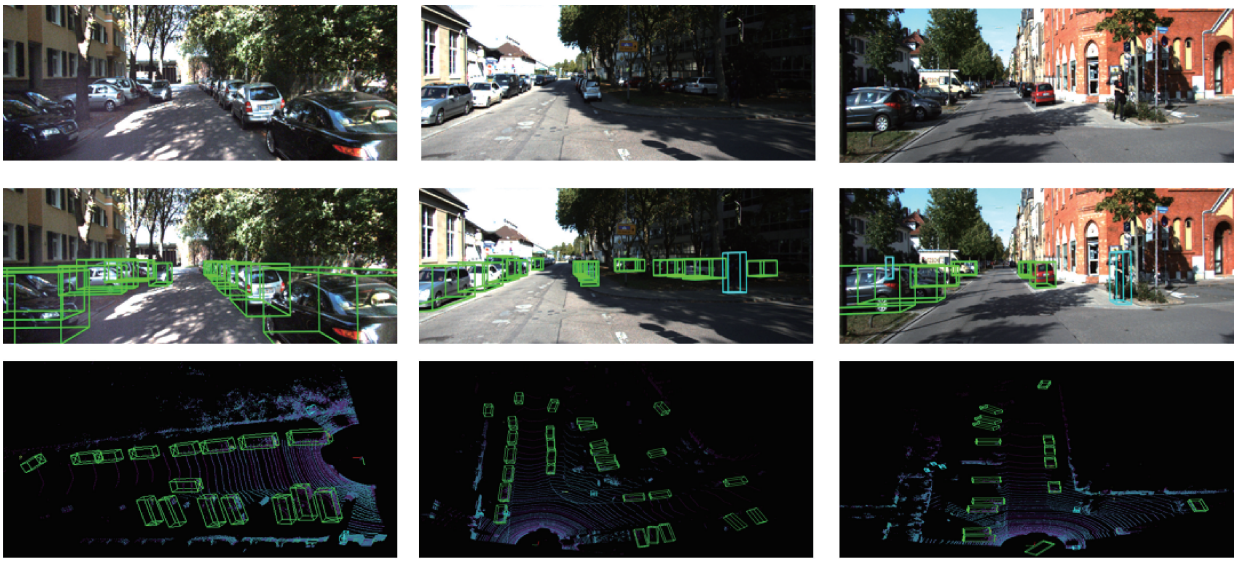


图 6　KITTI 数据集上的可视化结果

Fig. 6　Visualization results on the KITTI dataset

KITTI 数据集的 BEV 检测是对模型定位检测精度的有效验证,将 3D 边界框投影到 BEV 空间中,以汽车类别为例,在 KITTI 验证集上与其他方法进行比较,结果如图 7 所示,在“简单、中等、困难”3 个难度上,都超越了现有方法。

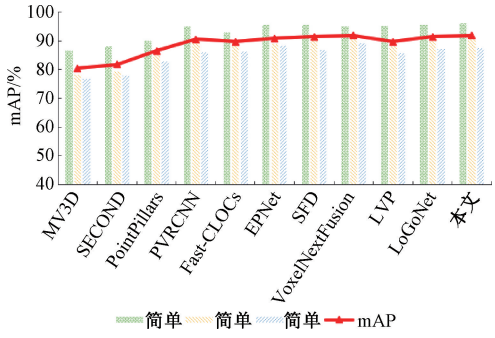


图 7　KITTI 数据集上的 BEV 检测结果

Fig. 7　The BEV detection results on the KITTI dataset

2.5　Waymo 数据集的实验结果

为验证模型远距离目标的识别效果,在 Waymo 数据集上进行测试。如表 5 所示,该方法在 30 m 外的远距离检测精度超过了现有方法,特别是在 30~50 m 范围内,mAP 与 mAPH 分别超过 LoGoNet[36] 0.52%与 0.65%。形成这一优势的关键在于,高频引导体素融合模块通过视锥投影将图像中丰富的边缘纹理细节映射到稀疏点云的局部区域,从而有效弥补了远距离目标几何结构信息不足的问题。相比传统融合方法仅基于点云局部密度进行补全,本文方法能够提供更强的上下文支持和几何细节复原能力,显著提升了长距离目标识别的准确率与稳定性。

表 5　Waymo 数据集上性能比较(类别 2)

Table 5　Performance on the Waymo validation set(Level 2)　(%)

方法	数据	mAP/mAPH		
		0~30 m	30~50 m	50 m 外
Pointpillars[20]	点云	87.93/87.30	63.80/62.36	38.20/36.87
PV-RCNN[27]	点云	84.20/82.87	69.13/68.03	56.48/55.24
CenterPoint[21]	点云	86.21/84.74	71.99/70.55	55.22/43.82
BEVFusion[7]	点云+图像	**89.07/87.76**	76.65/75.33	60.14/58.91
LoGoNet[36]	点云+图像	88.61/87.39	77.00/75.73	63.37/62.10
本文	点云+图像	88.94/87.69	**77.52/76.38**	**63.85/62.89**

2.6　实车实验

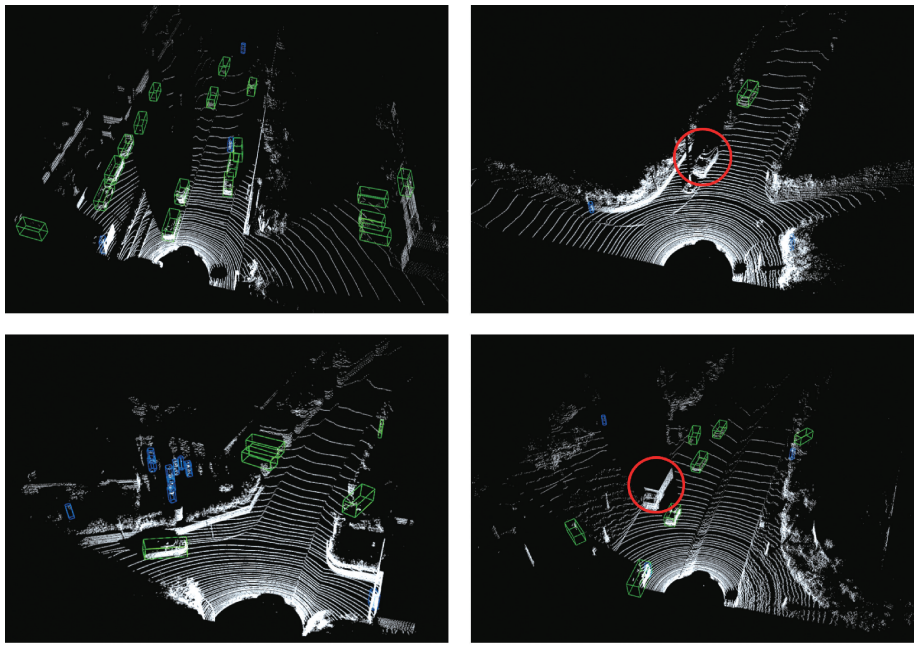
仅在数据集上进行仿真实验无法证明本文方法在实际场景中的性能。因此,本文进行了实车测试。如图 8 所示,实验平台包括 80 线激光雷达、GNSS 天线、双目相

图 8　用于测试的车辆平台

Fig. 8　The autonomous vehicle platform for testing

机、微机电系统惯性测量单元和 GNSS/INS 导航系统等。实验平台由清华大学苏州汽车研究院提供,数据采集工作于苏州市城市道路及快速路段进行,涵盖白天与夜间等光照条件,检测对象包括汽车、卡车和行人等 7 类目标。

如图 9(a)所示,在光照良好的白天场景中,系统能够准确地检测周围的车辆与行人,表现出较高的感知能力。然而,在夜间低照度条件下,如图 9(b) 所示,由于图像质量的下降,导致检测性能降低。例如,圆圈标示的区域中,尽管 LiDAR 已捕获到车辆的空间点云信息,但检测结果中未能正确识别并框选该目标,反映出模型在当前架构下对相机图像的依赖性较强。



(a) 白天
(a) Daytime

(b) 夜间
(b) Nighttime

图 9 实况检测结果

Fig. 9 Real time detection results

同时,如图 10 所示,随着设计模块的引入,模型在车辆、行人和公交车等不同类别目标上的检测精度均有不同程度提升,验证了所提方法在复杂交通环境下的有效性与适应性。

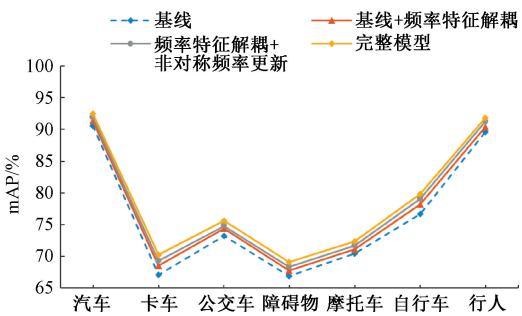


图 10 不同类别检测结果的精度比较

Fig. 10 Comparison of accuracy of detection results for different categories

2.7 消融实验

本节中,将在 KITTI 数据集上研究该模型中每个组成模块的贡献。在表 6 中,仅小波频域特征解耦(wavelet frequency-domain feature decoupling, WFFD)组件测试时,在小波分解后,采用传统 FPN 方法,进行自顶向下与和横向连接来融合多层不同频特征,这一频率解耦融合方法相比较传统多层特征堆叠获得了 1.71% 的 mAP 提升。随后,非对称频率更新(asymmetric frequency update, AFU)与高频引导体素融合(high-frequency guided voxel fusion, HGVF)模块的有效组合使得模型平均检测精度分别提升了 1.71% 和 1.35%。

表 6 不同组件在 KITTI 验证集上的消融研究

Table 6 Ablation study on different components on the KITTI validation set

组件				汽车 AP_{3D}/%			
基线	WFFD	AFU	HGVF	简单	中等	困难	*mAP*
√				88.47	79.24	77.19	81.63
√	√			89.87	81.52	78.63	83.34
√	√	√		90.76	83.33	79.57	84.55
√	√	√	√	91.65	85.23	80.81	85.90

测试不同图像编码器与输入分辨率对模型性能的影响,如表 7 所示。其中,Swin-T[17] 拥有最优的结果,超过了 ResNet-50[37] 与 CSPNet[38]。

表 7　图像编码器的消融研究

Table 7　Ablation study on image encoder

编码器	分辨率	汽车 AP_{3D}/%			
		简单	中等	困难	*mAP*
ResNet-50[37]	320×800	90.57	83.95	79.73	84.75
CSPNet[38]	384×1 056	91.02	84.38	79.89	85.10
Swin-T[17]	256×704	91.44	84.97	80.64	85.68
Swin-T[17]	384×1 056	91.65	85.23	80.81	85.90

为评估 AFU 模块数量与不同高低频特征分支配置对检测性能的影响,本研究设计了多组消融实验。如表 8 所示,随着 AFU 层数的增加,模型精度存在性能饱和点。以(2+1)×3 为例,该结构在引入 3 个 AFU 模块的条件下取得最高汽车类精度 85.90%,且计算开销控制在 39×10^9 浮点运算数(floating-point operations per second, FLOPs)。进一步提升 AFU 数量或增加高频分支数量虽可增强局部细节提取能力,但同时也引入冗余计算与语义干扰,未能持续提升整体性能。由此验证了 AFU 模块的有效性与合理配置的重要性。

表 8　在 KITTI 数据集上对 AFU 中的模块配置进行消融研究

Table 8　Ablation study on module configuration in AFU on the KITTI dataset.

实验	模型(高+低)×AFU	汽车 mAP/%	编码器(10^9/FLOPs)
1	(2+1)×1	84.17	9
2	(2+1)×2	85.24	24
3	(2+1)×3	**85.90**	39
4	(2+1)×4	85.69	55
5	(3+1)×2	85.36	27
6	(3+1)×3	85.77	43
7	(4+1)×3	85.19	47

3　结　　论

针对现有点云-图像融合三维目标检测方法在频率域信息利用不足,导致检测精度的提升受到限制的问题,提出了一种基于特征频率解耦融合的 Transformer 框架。通过小波频域特征解耦,分别提取和处理高频与低频特征,克服了传统方法中高低频特征耦合带来的信息冗余与细节损失。设计的非对称频率更新模块,采用动态自适应窗口编码与稀疏化可变形注意力机制,显著提升了模型对局部高频细节与全局低频结构的有效表达。构建的高频引导体素融合模块,通过视锥投影机制有效利用图像高频特征,补充点云稀疏区域的几何细节,提升了跨模态特征融合的效果。在 nuScenes 数据集中取得了 73.2% 的 mAP 与 74.3% 的 NDS 的优异检测效果。同时,在 KITTI、Waymo 数据集与实际车辆平台上的实验结果也表明,该模型具有良好的适应性,拥有优于现有方法的三维目标检测性能。

参考文献

[1] 吴军, 袁少博, 祝玉恒, 等. 采用自适应背景聚类的激光雷达与相机外参标定优化方法[J]. 仪器仪表学报, 2023, 44(2):230-237.
WU J, YUAN SH B, ZHU Y H, et al. Optimization method for external parameters calibration of LiDAR and camera using adaptive background clustering [J]. Chinese Journal of Scientific Instrument, 2023, 44(2): 230-237.

[2] 汤新华, 代道文, 陈熙源, 等. 基于 PointPillars 的改进三维目标检测算法[J]. 仪器仪表学报, 2024, 45(9):260-269.
TANG X H, DAI D W, CHEN X Y, et al. Improved three-dimensional object detection algorithm based on PointPillars[J]. Chinese Journal of Scientific Instrument, 2024, 45(9):260-269.

[3] 张李辉, 刘紫燕. 融合多尺度特征和自适应 NMS 的 3D 目标检测[J]. 电子测量技术, 2025, 48(4):191-198.
ZHANG L H, LIU Z Y. 3D object detection on fusing multi-scale features and adaptive NMS[J]. Electronic Measurement Technology, 2025, 48(4):191-198.

[4] 宋佳声, 李浩天. 一种在尺度空间下基于边缘的角点目标检测方法[J]. 电子测量与仪器学报, 2024, 38(2):58-66.
SONG J SH, LI H T. Edge-based target corner point detection method in scale space[J]. Journal of Electronic Measurement and Instrumentation, 2024, 38(2):58-66.

[5] 金宇锋, 陶重犇. 基于 Transformer 的融合信息增强 3D 目标检测算法[J]. 仪器仪表学报, 2023, 44(12): 297-306.
JIN Y F, TAO CH B. Fusion information enhanced method based on transformer for 3D object detection[J]. Chinese Journal of Scientific Instrument, 2023, 44(12): 297-306.

[6] 董钰婷, 官磊. 基于自适应加权融合激光雷达和相机

的三维目标检测方法[J]. 计算机应用, 2024, 44(S1):250-255.
DONG Y T, GUAN L. A 3D object detection method based on adaptive weighted fusion of LiDAR and camera[J]. Computer Applications, 2024, 44(S1): 250-255.

[7] LIU ZH J, TANG H T, 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, 2023: 2774-2781.

[8] LIANG T T, XIE H W, YU K CH, et al. Bevfusion: A simple and robust LiDAR-camera fusion framework[J]. Advances in Neural Information Processing Systems, 2022, 35: 10421-10434.

[9] 黄德启, 黄海峰, 黄德意, 等. BEV 感知学习在自动驾驶中的应用综述[J]. 计算机工程与应用, 2025, 61(6):1-21.
HUANG D Q, HUANG H F, HUANG D Y, et al. A review of BEV perception learning in autonomous driving[J]. Computer Engineering and Applications, 2025, 61(6):1-21.

[10] MA Y X, WANG T, BAI X Y, et al. Vision-centric bev perception: A survey[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 46(12): 10978-10997.

[11] GU H X, XIONG R P, TIAN Y. SF-CLIP3D: Spatial-frequency-enhanced vision-language models for multi-modal 3D object detection[J]. The Journal of Supercomputing, 2025, 81(8): 903.

[12] FINDER S E, AMOYAL R, TREISTER E, et al. Wavelet convolutions for large receptive fields[C]. European Conference on Computer Vision, 2024: 363-380.

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

[14] 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.

[15] SUN P, KRETZSCHMAR H, DOTIWALLA X, et al. Scalability in perception for autonomous driving: Waymo open dataset[C]. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 2446-2454.

[16] CONTRIBUTORS M. MMDetection3D: OpenMMLab next-generation platform for general 3D object detection[EB/OL]. [2025-01-13]. https://github.com/open-mmlab/mmdetection3d.

[17] LIU Z, LIN Y T, CAO Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows[C]. 2021 IEEE/CVF International Conference on Computer Vision, 2021: 10012-10022.

[18] KINGMA D P, BA J. Adam: A method for stochastic optimization[J]. ArXiv preprint arXiv: 1412.6980, 2014.

[19] BAI X Y, HU Z Y, ZHU X G, et al. Transfusion: Robust LiDAR-camera fusion for 3D object detection with transformers[C]. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 1090-1099.

[20] 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, 2019: 12697-12705.

[21] YIN T W, ZHOU X Y, KRAHENBUHL P. Center-based 3D object detection and tracking[C]. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021: 11784-11793.

[22] YIN T W, ZHOU X Y, KRÄHENBÜHL P. Multimodal virtual point 3D detection[J]. Advances in Neural Information Processing Systems, 2021, 34: 16494-16507.

[23] WANG CH W, MA CH, ZHU M, et al. Pointaugmenting: Cross-modal augmentation for 3D object detection[C]. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021: 11794-11803.

[24] CAI Q, PAN Y W, YAO T, et al. ObjectFusion: Multi-modal 3D object detection with object-centric fusion[C]. 2023 IEEE/CVF International Conference on Computer Vision, 2023: 18067-18076.

[25] CHEN X Y, ZHANG T Y, WANG Y, et al. FUTR3D: A unified sensor fusion framework for 3D detection[C]. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023: 172-181.

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

[27] SHI SH SH, GUO CH X, JIANG L, et al. PV-RCNN: Point-voxel feature set abstraction for 3d object detection[C]. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 10529-10538.

[28] WU H, DENG J H, WEN CH L, 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: 1-11.

[29] LIANG M, YANG B, CHEN Y, et al. Multi-task multi-sensor fusion for 3D object detection[C]. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019: 7345-7353.

[30] YOO J H, KIM Y, KIM J, et al. 3D-CVF: Generating joint camera and lidar features using cross-view spatial feature fusion for 3D object detection[C]. Computer Vision-ECCV 2020, 2020: 720-736.

[31] PANG S, MORRIS D, RADHA H. Fast-CLOCs: Fast camera-LiDAR object candidates fusion for 3D object detection[C]. 2022 IEEE/CVF Winter Conference on Applications of Computer Vision. 2022: 187-196.

[32] HUANG T T, LIU ZH, CHEN X W, et al. EPNET: Enhancing point features with image semantics for 3D object detection[C]. Computer Vision-ECCV 2020, 2020: 35-52.

[33] 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, 2022: 5418-5427.

[34] SONG Z Y, ZHANG G X, XIE J, et al. VoxelNextFusion: A simple, unified, and effective voxel fusion framework for multimodal 3-D object detection[J]. IEEE Transactions on Geoscience and Remote Sensing, 2023, 61: 1-12.

[35] CHEN Y D, CAI G R, SONG Z Y, et al. LVP: Leverage virtual points in multi-modal early fusion for 3D object detection[J]. IEEE Transactions on Geoscience and Remote Sensing, 2024, 63: 1-15.

[36] LI X, MA T, HOU Y N, 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, 2023: 17524-17534.

[37] HE K M, ZHANG X Y, REN SH Q, et al. Deep residual learning for image recognition[C]. 2016 IEEE Conference on Computer Vision and Pattern Recognition, 2016: 770-778.

[38] WANG C Y, LIAO H Y M, WU Y H, et al. CSPNet: A new backbone that can enhance learning capability of CNN[C]. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020: 1571-1580.

作者简介

李明光,2023 年于江苏海洋大学获得学士学位,现为苏州科技大学硕士研究生,主要研究方向为多模态融合和三维目标检测。
E-mail:Minggliusts@163.com

Li Mingguang received his B. Sc. degree from Jiangsu Ocean University in 2023. He is currently a M. Sc. candidate at Suzhou University of Science and Technology. His main research interests include Multi modal fusion and 3D object detection.

陶重犇(通信作者),2014 年于江南大学获得博士学位,现为苏州科技大学副教授,清华大学苏州汽车研究院博士后,主要研究方向为自动驾驶多模态融合感知。
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. He is also a postdoctoral fellow of Suzhou Automobile Research Institute of Tsinghua University. His main research interest includes multimodal fusion perception for autonomous driving.