Computer Science and Application 计算机科学与应用, 2025, 15(11), 143-155
Published Online November 2025 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2025.1511292

Hans汉斯

FM-VXNet：一种基于MVX-Net改进的多模态3D目标检测算法的研究

郑广海*，张　薇，张　倩

大连交通大学轨道智能工程学院，辽宁 大连

收稿日期：2025年10月12日；录用日期：2025年11月12日；发布日期：2025年11月21日

摘　要

多模态3D目标检测通过融合不同模态信息，有效克服了单一模态的局限性，在自动驾驶和机器人导航等领域展现出重要价值。然而，现有方法仍存在图像分支对全局语义的建模能力有限、跨模态融合多依赖简单拼接，未能充分挖掘模态间的互补潜力、点云体素特征在密度分布不均时易受噪声或冗余信息干扰等不足。针对上述问题，本文提出频域多模态体素网络(Frequency-domain Multimodal Voxel Network, FM-VXNet)模型。该模型是基于多模态体素网络(Multimodal Voxel Network, MVX-Net)设计，它包含三个核心模块：(1) 在图像分支中引入频域－空间域融合模块(Frequency and Spatial Fusion Module, FFCM)，借助快速傅里叶变换增强全局语义感知能力；(2) 提出双向跨模态门控注意力模块(Bidirectional Cross-Modal Gated Attention, Bi-CMGA)，实现图像与点云特征间的双向交互融合，并引入通道级门控机制以抑制噪声干扰，提升融合特征的判别力；(3) 在体素特征编码阶段设计双模态密度感知注意力模块(Bimodal Density-aware Attention, BiDA)，通过密度感知与通道重标定机制，有效缓解稀疏体素中的噪声问题和密集体素中的冗余现象。改进后的FM-VXNet算法在KITTI数据集上的实验表明，FM-VXNet在鸟瞰图(BEV)检测任务中，全类平均精度(mean Average Precision, mAP)在简单、中等和困难场景下分别达到96.3%、95.2%和92.9%；在3D检测任务中，mAP分别达到96.2%、88.9%和87.7%，相较BEVFusion、MVX-Net等主流算法平均提升5.7%~8.2%。本研究创新性地引入频域分析、双向门控注意力与密度感知机制，为多模态3D目标检测提供了新的研究思路。

关键词

多模态，3D目标检测，频域建模，跨模态融合，密度感知注意力，自动驾驶

FM-VXNet: A Study on an Improved Multimodal 3D Object Detection Algorithm Based on MVX-Net

Guanghai Zheng*, Wei Zhang, Qian Zhang

School of Railway Intelligent Engineering, Dalian Jiaotong University, Dalian Liaoning

*通讯作者。

文章引用: 郑广海, 张薇, 张倩. FM-VXNet: 一种基于 MVX-Net 改进的多模态 3D 目标检测算法的研究[J]. 计算机科学与应用, 2025, 15(11): 143-155. DOI: 10.12677/csa.2025.1511292

Received: October 12, 2025; accepted: November 12, 2025; published: November 21, 2025

Abstract

Multimodal 3D object detection, by fusing data from different modalities, effectively overcomes the limitations of single-modal approaches and has demonstrated significant value in fields such as autonomous driving and robot navigation. However, current methods still face several shortcomings: the image branch has a limited capacity for global semantic modeling; cross-modal fusion often relies on simple feature concatenation, failing to fully exploit the complementary potential between modalities; and point cloud voxel features are susceptible to noise or redundant information when the density distribution is uneven. To address these issues, this paper proposes the Frequency-domain Multimodal Voxel Network (FM-VXNet). Designed based on the Multimodal Voxel Network (MVX-Net), the model incorporates three core modules: (1) the Frequency and Spatial Fusion Module (FFCM), which leverages the Fast Fourier Transform (FFT) to enhance global semantic perception in the image branch; (2) the Bidirectional Cross-Modal Gated Attention (Bi-CMGA) module, which enables bidirectional interactive fusion between image and point cloud features and introduces a channel-wise gating mechanism to suppress noise and improve the discriminative power of the fused features; (3) the Bimodal Density-aware Attention (BiDA) module, which operates during the voxel feature encoding stage and effectively mitigates noise in sparse voxels and redundancy in dense voxels through density-aware and channel re-calibration mechanisms. Experiments on the KITTI dataset show that the enhanced FM-VXNet algorithm achieves mean Average Precision (mAP) scores of 96.3%, 95.2%, and 92.9% for the Bird's Eye View (BEV) detection task under easy, moderate, and hard settings, respectively. For the 3D detection task, it achieves mAP scores of 96.2%, 88.9%, and 87.7% across the respective difficulty levels, outperforming state-of-the-art methods like BEV Fusion and MVX-Net by an average of 5.7% to 8.2%. This research innovatively introduces frequency-domain analysis, bidirectional gated attention, and density-aware mechanisms, offering a new direction for multimodal 3D object detection research.

Keywords

Multimodal, 3D Object Detection, Frequency Domain Modeling, Cross-Modal Fusion, Density-Aware Attention, Autonomous Driving

1. 引言

随着 3D 传感技术的发展，3D 目标检测在自动驾驶、机器人导航等领域的需求日益迫切。激光雷达与相机作为主流感知设备各具优势：激光雷达能提供精确的三维几何信息，但存在稀疏性和高成本的局限；相机可捕获丰富的纹理与语义特征，但其深度感知能力较弱且易受环境影响[1]-[3]。因此，融合两种模态数据以突破单模态局限，成为提升 3D 目标检测性能的关键，对推动自动驾驶感知系统实用化具有重要意义[4] [5]。

现有研究主要围绕单模态优化与多模态融合两个方向展开。在激光雷达方面，研究者们提出了点云稀疏补全[6]、半监督域自适应[7]、轻量化网络设计[8]及去噪技术[9]等方法以提升性能。在相机方面，工作重点包括引入深度感知 Transformer [10]、几何一致性约束[11]、去噪自编码器[12]、无监督域适应[13]

以及轻量化和恶劣天气下的鲁棒性研究[14] [15]。多模态融合方面，早期研究如 MVX-Net [16]通过体素或点级融合实现多模态交互；近期方法则探索了合作感知[17]、遮挡感知融合[18]、跨模态与跨尺度平衡[19]等多种策略。然而，这些方法在全局语义建模、跨模态交互机制和体素特征鲁棒性方面存在共性瓶颈。现有方法存在三方面不足：一是图像特征提取依赖卷积操作，缺乏全局语义建模能力，难以捕获长程依赖；二是跨模态融合多采用简单拼接或静态加权，无法动态挖掘模态间互补信息；三是点云体素特征受密度不均影响，稀疏体素易引入噪声，密集体素存在冗余，导致特征判别性低。

综上所述，这些不足共同构成了当前多模态 3D 检测性能提升的主要瓶颈。为了解决上述问题，本文提出一种兼顾全局特征建模与双向跨模态融合的多模态 3D 目标检测模型，对推动自动驾驶感知系统的实用化具有重要意义。本文提出 FM-VXNet 模型，引入 FFCM 模块：将频域建模引入多模态 3D 检测的图像分支，通过 FFT 实现全局语义捕获，弥补卷积网络的长程依赖缺陷，尤其提升远距离、遮挡目标的表征能力；设计 Bi-CMGA 模块：构建双向跨模态注意力交互与门控融合机制，动态平衡相机与点云的模态贡献，解决传统拼接方法的信息丢失问题；提出 BiDA 模块：在 VFE 阶段引入密度感知双重注意力，分别处理稀疏体素噪声与密集体素冗余，提升 LiDAR 点云体素特征的判别性。

2. 相关工作与基线模型分析(MVX-Net 模型)

本节首先简要综述多模态融合的典型思路，并重点分析本文的基线模型 MVX-Net。该模型的 PointFusion 策略是本研究的起点，但其局限性也直接引出了本文的创新方向。MVX-Net [16]是早期多模态 3D 检测的经典框架，通过扩展 VoxelNet 架构，实现 LiDAR 点云与相机特征的早期融合。其核心优势在于单阶段检测效率与多模态早期交互。MVX-Net 通过 PointFusion 与 VoxelFusion 两种方式实现多模态融合，本文以 PointFusion 为基础进行改进，其架构如图 1 所示。

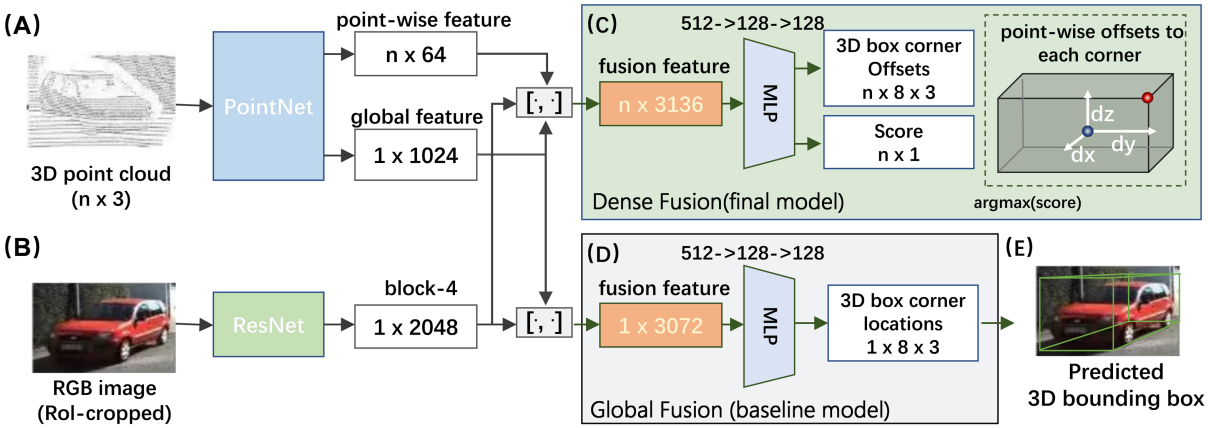

Figure 1. MVX-Net model diagram
图 1. MVX-Net 模型图

在图像特征提取模块，采用预训练的 Faster R-CNN (ResNet-50 + FPN)作为 2D 检测器，从 RGB 图像中提取高层语义特征图(conv5 层，256 维)。通过相机内外参矩阵，将 LiDAR 点云投影至 2D 图像平面，获取每个 3D 点对应的像素坐标，进而采样图像特征向量，实现“点–像素”特征关联。点云特征增强模块：原始 LiDAR 点云以三维坐标(x, y, z)表示，将每个点的坐标特征(3 维)与对应的图像语义向量(256 维)逐点拼接，形成 259 维增强特征。若多个点投影至同一像素，共享该像素的图像特征，减少计算冗余，但可能降低局部区分度。多模态融合检测模块：将增强点云划分为均匀体素网格(0.2 m × 0.2 m × 0.2 m)，每个体素内的点通过 VFE 层提取局部特征(如均值、最大值)；通过 3D 卷积聚合体素级全局特征，生成

BEV 特征图；基于 BEV 特征图生成 3D 候选框，通过分类与回归头输出检测结果。

然而，作为早期工作，MVX-Net 仍存在明显的局限性，这些局限性也正是本文旨在解决的核心问题：在图像特征提取方面，其依赖于基于 CNN 的 ResNet-50 主干网络，CNN 擅长捕获局部特征，但由于感受野有限，难以建模全局长程依赖，从而导致远距离或被遮挡目标的特征表征不足。在跨模态融合方面，MVX-Net 仅采用“坐标特征 + 图像特征”的逐点拼接策略，未能显式考虑 LiDAR 点云的几何结构与相机的语义特性差异，容易丢失关键的互补信息。在体素特征建模方面，其 VFE 层仅对体素内点进行统计聚合，缺乏对体素密度差异的建模能力：稀疏体素易受噪声干扰，而密集体素则存在冗余信息，从而降低了特征判别性。针对上述不足，本文提出 FM-VXNet 模型，并通过 FFCM、Bi-CMGA 与 BiDA 三个创新模块实现全面改进。

综上所述，MVX-Net 作为多模态融合的早期代表，其存在的三方面局限性恰恰对应了第 1 节所提出的当前领域面临的共性挑战：其一，其基于 CNN 的图像特征提取模块缺乏全局语义建模能力，对应引言中所述“图像分支全局建模不足”的问题；其二，其采用简单的点级拼接融合策略，未能实现模态间的深度互补，对应跨模态融合简单的问题；其三，其 VFE 层“对体素密度变化不敏感，对应体素特征鲁棒性差”的问题。这些局限性为本研究的改进提供了明确的方向。因此，本文在第 3 节提出 FM-VXNet 模型，旨在通过 FFCM、Bi-CMGA 和 BiDA 三个模块分别针对上述问题予以系统性解决。

3. FM-VXNet 网络

FM-VXNet 基于 MVX-Net 的 PointFusion 策略进行改进，其整体架构如图 2 所示，主要包含四个部分：图像特征提取分支(ResNet-50 + FFCM + FPN)、LiDAR 点云特征处理分支(体素化 + BiDA + VFE)、跨模态融合模块(Bi-CMGA)与 3D 检测头(SECOND + 3D Region Proposal Network, 3DRPN)。以下将详细阐述三个核心模块的设计动机与具体结构。

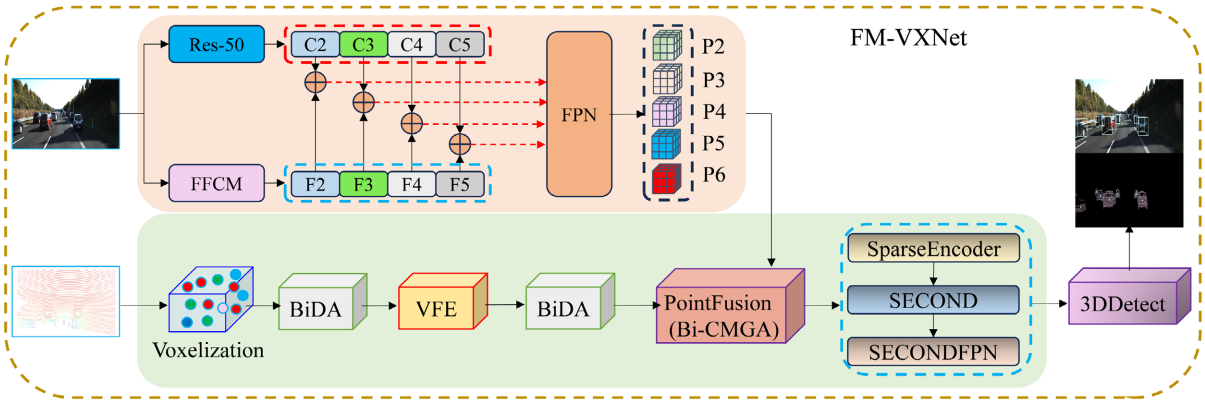

Figure 2. FM-VXNet network architecture
图 2. FM-VXNet 网络架构

本研究的三个模块分别针对第 1 节分析的三个核心问题：首先，针对图像全局建模不足的问题，MVX-Net 的 ResNet-50 分支依赖局部卷积，难以捕获图像的长程语义(如远距离目标的上下文关联、遮挡区域的语义补全)。因此，本文引入 FFCM 模块，通过 FFT 将图像特征映射至频域，实现全局结构建模，同时保留局部空间细节，形成“全局–局部”互补特征。其次，针对跨模态融合简单的问题，MVX-Net 的 PointFusion 采用“点云特征 + 图像特征”逐点拼接，未考虑模态差异：LiDAR 点云擅长几何定位，相机擅长语义分类，简单拼接易导致“模态主导”(如近距场景点云主导，远距场景图像主导)，丢失互补信息。为此，本文提出 Bi-CMGA 模块，通过双向注意力交互与门控融合，动态平衡模态贡献。最后，针对体素

特征鲁棒性问题，MVX-Net 的 VFE 层仅对体素内点进行统计聚合(如 max-pooling)，未处理体素密度差异稀疏体素(如远距离目标)易受噪声点干扰，密集体素(如近距目标)易引入冗余信息，导致特征判别性低。相应地，借鉴 C2BG-Net [20]的 LGVAE 全局聚合思想，本文提出 BiDA 模块，在 VFE 前后分别引入密度门控与通道重标定，提升体素特征鲁棒性。

3.1. FFCM 模块

FFCM 模块旨在解决图像分支全局语义建模能力不足的问题。该模块嵌入 ResNet-50 的 C2~C5 层，与卷积分支形成残差融合，FFCM 模块结构如图 3 所示。在局部空间特征提取：输入特征经逐点卷积(Pointwise Convolution, PConv)压缩通道后，分为两路深度可分离卷积(DConv3 × 3、5 × 5)，分别捕获不同感受野的局部细节；两路特征经 GeLU 激活与 PConv 处理后拼接，得到多尺度局部特征。全局频域特征建模：拼接后的局部特征通过 2DFFT 映射至频域，经 PConv 学习通道间全局依赖，再通过批归一化(BN)与 ReLU 激活增强表达；最后通过逆 FFT (iFFT)将频域特征映射回空间域，以实现全局频域特征建模。最后，全局频域特征与局部空间特征以残差形式相加，经 PConv 进一步融合，输出兼具全局语义与局部细节的图像特征并与 C2~C5 特征进行融合。这样既保证了局部几何信息的保留，又显式增强了全局语义建模能力，从而显著改善了图像分支对远距和遮挡目标的表征能力，并在多模态融合阶段提供更判别、更鲁棒的图像特征。

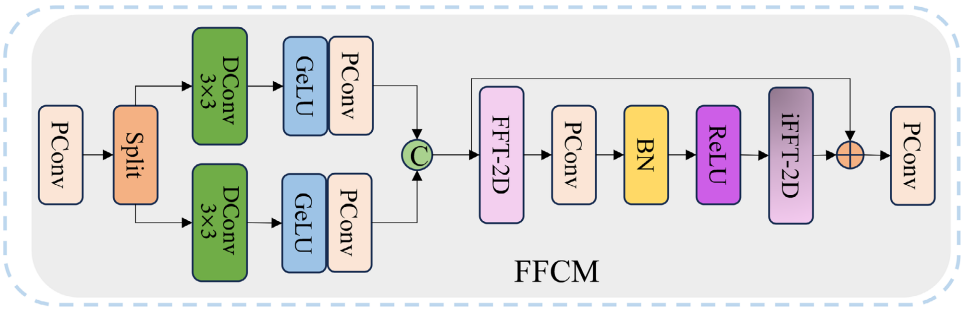

Figure 3. FFCM module structure diagram
图 3. FFCM 模块结构

设输入图像特征为 $X_{in} \in \mathbb{R}^{H\times W\times C}$ (H、W 为特征图尺寸，C 为通道数)，FFCM 的核心过程可表达如下：

1) 局部特征提取：

$$F_{local}^{1} = \mathrm{DConv}_{3\times 3}\left(\mathrm{PConv}\left(X_{in}\right)\right) \tag{1}$$

$$F_{local}^{2} = \mathrm{DConv}_{5\times 5}\left(\mathrm{PConv}\left(X_{in}\right)\right) \tag{2}$$

$$F_{local} = \mathrm{PConv}\left(\mathrm{Concat}\left(F_{local}^{1}, F_{local}^{2}\right)\right) \tag{3}$$

2) 频域全局建模：

$$F_{freq} = \mathcal{F}\left(F_{local}\right) \tag{4}$$

$$F'_{freq} = \mathrm{PConv}\left(\mathrm{ReLU}\left(\mathrm{BN}\left(F_{freq}\right)\right)\right) \tag{5}$$

$$F_{global} = \mathcal{F}^{-1}\left(F'_{freq}\right) \tag{6}$$

3) 残差融合：

$$X_{out} = \mathrm{PConv}\left(F_{local} + F_{global}\right) + X_{in} \tag{7}$$

其中 X_{out} 为 FFCM 输出特征，残差项确保 X_{in} 原始卷积特征不丢失。

与 MVX-Net 只利用 ResNet-50 提取图像特征相比，引入 FFCM 既保证局部几何信息，又增强全局语义建模能力，从而有效解决图像分支长程依赖缺失的问题，显著改善了图像分支对远距和遮挡目标的表征能力。

3.2. BiDA 模块

BiDA 模块旨在提升点云体素特征在密度不均场景下的鲁棒性和判别性。该模块嵌入 LiDAR 点云特征处理分支，分为“逐点增强”(VFE 前)与“体素重标定”(VFE 后)两阶段，如图 4 所示。逐点增强(VFE 前)：对每个 LiDAR 点云，通过两层线性变换投影至“语义子空间”，增强特征表达；计算点所属体素的密度(体素内点数)，经 log 压缩与 Sigmoid 生成密度门控标量，抑制稀疏体素的噪声干扰；密度门控与语义子空间特征逐通道相乘，实现“密度自适应”的逐点增强。体素重标定(VFE 后)：对 VFE 输出的体素特征，沿通道轴施加 1D 局部卷积，捕获通道间局部依赖；卷积结果经 Sigmoid 生成通道门控向量，动态突出关键通道(如几何特征通道)，抑制冗余通道；通道门控与原始体素特征逐通道相乘，实现体素级特征重标定。

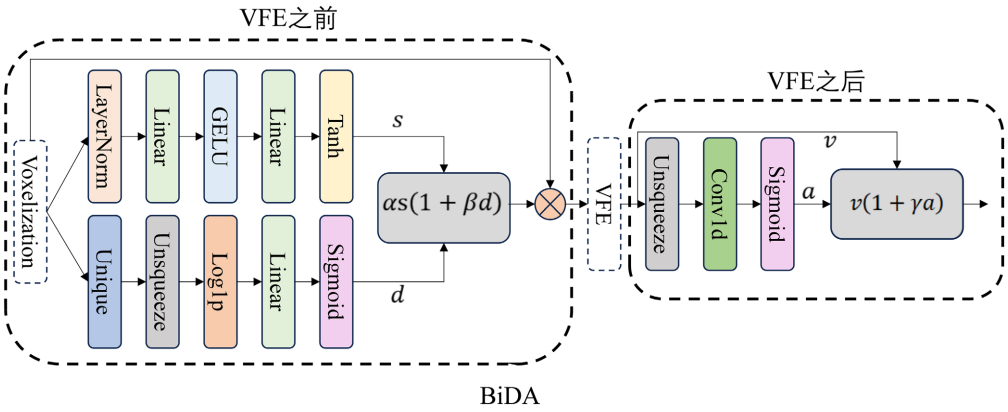

Figure 4. BiDA module framework
图 4. BiDA 模块框架

设第 i 个点的输入特征记为 $f_i \in \mathbb{R}^{C_{in}}$ 可由坐标、反射强度及先验编码拼接得到)，BiDA 的过程可以表达如下：

1) VFE 之前：逐点增强。

(a) 归一化与两层仿射－非线性变换(“语义”子空间投影)：

$$s_i = \tanh\left(\mathrm{Linear}_2\left(\mathrm{GELU}\left(\mathrm{Linear}_1\left(\mathrm{LN}\left(f_i\right)\right)\right)\right)\right) \tag{8}$$

(b) 对体素密度 N_v 做对数压缩后经 Sigmoid 得到标量门控，并在通道维广播至 C 维：：

$$d_v = \sigma\left(\log\left(N_v + 1\right)\right) \tag{9}$$

$$\tilde{d}_i = \mathrm{Broadcast}_C\left(d_v\right) \tag{10}$$

其中，d_v 表示密度门控标量，$\tilde{d}_i$ 表示 d_v 在通道维的广播。该门控用于在稀疏/密集体素间自适应调节增益强度，避免稀疏噪声被放大。

(c) 逐通道残差加权(乘性)将“语义”向量与密度门控耦合得到逐点的通道增益，并以残式乘性方式施加到原始特征：

$$m_i = \alpha\left(s_i \odot \left(1 + \beta\tilde{d}_i\right)\right) \tag{11}$$

$$\tilde{f}_i = f_i \odot (1 + m_i) \tag{12}$$

其中 α 在实验中设置为 0.1，β 为 1.0，两者用于控制语义与密度两路增益的幅度；$\tilde{f}_i$ 为为逐点增强后的输出。

2) VFE 之后

为在通道轴刻画局部相互作用，对 VFE 输出体素特征 $V \in \mathbb{R}^C$ 施加同长一维卷积并经 Sigmoid 生成通道门控：

$$g_{\text{channel}} = \sigma\left(\text{Conv1D}(V, k)\right) \tag{13}$$

其中卷积核长度为 k(本文设置为 3)，边界按零填充处理。最后以残差乘性方式对通道进行重标定，突出几何判别通道、抑制冗余响应：

$$V' = V \odot (1 + \gamma \cdot g_{\text{channel}}) \tag{14}$$

其中，V' 为 VFE 之后的体素级重标定输出；γ 控制通道门控注入强度，在本文设置为 0.5。

BiDA 通过“密度门控 + 通道重标定”双重机制，有效解决了 VFE 层的固有缺陷：a) 稀疏体素鲁棒性：密度门控抑制噪声点影响，提升远距离目标的特征质量；b) 密集体素判别性：通道重标定突出关键几何特征，减少冗余信息，提升近距目标的分类精度。

3.3. Bi-CMGA 模块

Bi-CMGA 模块旨在实现精细化的跨模态融合，以替代简单的特征拼接。该模块将来自图像分支 FPN 模块中的特征与 BiDA 输出的点云特征进行跨模态融合，其结构如图 5 所示。Bi-CMGA 将 FPN 输出的多尺度图像特征(P2~P6)通过双线性插值统一至点云特征分辨率，得到图像语义向量 $F_i \in \mathbb{R}^{N \times C}$ (N 为点云数量，C 为通道数)；点云特征经 BiDA 增强后为 $F_p \in \mathbb{R}^{N \times C}$；双向注意力交互中，点云到图像注意力以点云特征为查询，图像特征为键值，计算注意力权重，为点云补充语义信息；图像到点云注意力以图像特征为查询，点云特征为键值，计算注意力权重，为图像补充几何信息；门控融合引入通道级门控函数，动态加权两路注意力输出，抑制噪声模态(如遮挡场景抑制点云，增强图像)；残差注入融合特征以残差形式注入点云特征，保留原始几何信息，输出最终跨模态特征 $fusion_{L+C}$。

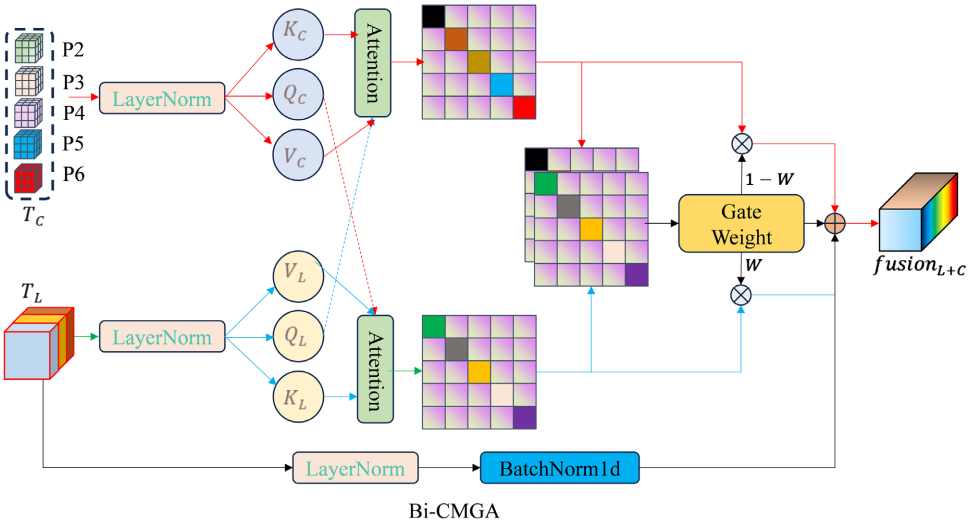

Figure 5. Bi-CMGA module framework
图 5. Bi-CMGA 模块框架

Bi-CMGA 的过程可以表达如下：

设对齐后的图像特征 $F_i \in \mathbb{R}^{N\times C}$，点云特征 $F_p \in \mathbb{R}^{N\times C}$，Bi-CMGA 过程如下：

1) 双向注意力计算：

$$A_{p2i} = \text{Softmax}\left(\frac{Q_p K_i^{\mathrm{T}}}{\sqrt{d}}\right) \tag{15}$$

$$F_{p2i} = A_{p2i} V_i \tag{16}$$

$$A_{i2p} = \text{Softmax}\left(\frac{Q_i K_p^{\mathrm{T}}}{\sqrt{d}}\right) \tag{17}$$

$$F_{i2p} = A_{i2p} V_p \tag{18}$$

其中 $Q_p = F_p W_p$，$K_i = F_i W_k$，$V_i = F_i W_v$，同理可得 Q_i，K_p，V_p。

2) 门控融合：

$$g = \sigma\left(\text{Linear}\left(\text{Concat}\left(F_{p2i}, F_{i2p}\right)\right)\right) \tag{19}$$

$$F_{fuse} = g \odot F_{p2i} + (1-g) \odot F_{i2p} \tag{20}$$

3) 残差注入：

$$fusion_{L+C} = F_p + \gamma \cdot F_{fuse} \tag{21}$$

其中，d 为键向量的维度。γ 为超参数(本文设置为 0.5)，控制融合强度，$fusion_{L+C}$ 为最终跨模态特征。

与传统拼接及静态加权相比，Bi-CMGA 具有显著优势：a) 双向交互：点云与图像的双向注意力，充分挖掘互补信息，解决了简单拼接信息丢失问题；b) 动态门控：通道级门控函数适配复杂场景(如遮挡、远距)，平衡模态贡献；c) 几何保留：残差注入确保点云原始几何信息不丢失，提升定位精度

4. 实验与结构分析

4.1. 实验细节

实验硬件平台配置为计算节点搭载两颗 Intel Xeon Gold 6330 CPU@ 2.0 GHz (总计 56 核心)及 256 GB DDR4 内存；GPU 加速单元采用 2 张 NVIDIA GeForce RTX 4090D 显卡(总计显存 48 GB)，通过 PCIe 4.0 x16 互联；软件环境基于 PyTorch 1.11.0 框架，CUDA 11.3 及 Python 3.8。训练阶段设置批量大小(batch_size)为 16，训练轮数(epoch)为 200，初始学习率为 0.01 并采用余弦退火策略衰减至 1×10^{-6}。输入图像尺寸统一调整为 640 × 640 像素，点云体素化网格大小设为 0.2 m × 0.2 m × 0.2 m。

4.2. 数据集与评估指标

本研究选用 KITTI 数据集开展实验评估，该数据集包含 7481 个训练样本以及 7518 个测试样本，并依据检测难度划分为简单、中等、困难三个级别。为了更科学地评估模型性能，按惯例将训练集划分为 3712 个样本用于训练，3769 个样本用于验证。在实验过程中，采用平均精度(mAP)作为核心指标，BEV 检测与 3D 检测的 IoU 阈值均设为 0.7。

4.3. 对比实验

为验证 FM-VXNet 的性能优势，本文在 KITTI 数据集的汽车类别上与 19 种主流方法进行对比，涵盖单模态(仅相机 C、仅激光雷达 L)和多模态(C + L)两类方案，结果如表 1 所示。单模态方法对比：仅依

赖相机的方法(如 MMono3D [10]、3DOP [21])受限于深度信息缺失，BEV 和 3D 检测 mAP 均显著低于多模态方法，其中 MMono3D [10]在 BEV 简单场景下 mAP 仅 5.22%，验证了单相机模态的固有局限。仅依赖激光雷达的方法(如 PV-RCNN [22]、Voxel-RCNN [23])通过高精度几何信息实现了较高检测性能，PV-RCNN [22]的 BEV 中等场景 mAP 达 91.1%，3D 中等场景 mAP 达 84.8%，但受限于语义信息不足，在复杂遮挡场景下性能仍有提升空间。多模态方法对比：早期多模态方法如 MV3D [24]通过简单特征拼接实现融合，BEV 中等场景 mAP 为 78.1%，3D 中等场景 mAP 为 62.7%，性能受限明显。FM-VXNet 的基准模型 MVX-Net (PF) [16]通过 PointFusion 实现点云与图像特征的早期融合，BEV 中等场景 mAP 提升至 84.9%，3D 中等场景 mAP 达 73.3%，但因缺乏全局语义建模和动态融合机制，与当前最优方法仍有差距。近年来的多模态优化方法中，GraphAlign [20]通过图对齐机制增强跨模态关联，BEV 中等场景 mAP 达 92.8%；SSLFusion [25]引入自监督学习优化特征一致性，3D 简单场景 mAP 达 94.1%；DVF [26]通过动态体积融合提升 BEV 检测性能，简单场景 mAP 达 96.2%然而，在困难场景下，这些方法的性能仍有提升空间。DVF [26]的 BEV 困难场景 mAP 为 89.2%，3D 困难场景 mAP 为 83.1%。

实验结果表明，FM-VXNet 在所有评估场景下均达到了最先进的性能，BEV 检测中简单/中等/困难场景 mAP 分别为 96.3%/95.2%/92.9%，3D 检测中分别为 96.2%/88.9%/87.7%。与主流多模态方法相比，平均提升 5.7%~8.2%：较 MVX-Net (PF) [16]的 BEV 中等场景提升 10.3%，3D 中等场景提升 15.6%；较 BEVFusion [27]的 3D 中等场景相对性提升 4.1%，困难场景提升 5.4%。FM-VXNet 的性能优势主要源于其更精细的融合策略(Bi-CMGA)和更强的噪声抑制能力(BiDA 与 FFCM 协同)，使其在复杂场景下能更有效地利用互补信息。尤其在困难场景(如远距离、遮挡目标)中，FM-VXNet 的 BEV 和 3D 检测 mAP 分别超出次优方法(GraphAlign [20]) 1.5%和 3.0%。这一结果验证了 FFCM 模块的全局语义建模、Bi-CMGA 模块的动态融合以及 BiDA 模块的体素优化的协同作用，表明本模型能更有效地处理复杂场景下的感知挑战。

实验结果表明，FM-VXNet 通过三大模块的创新设计，有效弥补了现有方法在全局语义捕获、跨模态互补信息挖掘以及体素特征鲁棒性方面的不足，在复杂交通场景中展现出更优的检测精度和鲁棒性，为自动驾驶多模态感知提供了高效解决方案。

在 3D 检测任务中，FM-VXNet 同样表现优异，其 mAP 分别为 96.2%、88.9%和 87.7%，较 MVX-Net (PF) [16] (85.5%/73.3%/67.4%)提升显著，尤其在中等和困难场景下优势更为明显。这表明 FFCM 模块的频域全局建模、Bi-CMGA 模块的双向门控融合以及 BiDA 模块的密度感知机制有效协同，增强了模型对复杂场景(如遮挡、远距离、点云稀疏)的适应能力。同时，FM-VXNet 在多模态融合方法中取得了当前最佳性能，验证了其在语义－几何互补性挖掘与噪声抑制方面的优势。可视化结果如图 6 所示，进一步展示了 FM-VXNet 在复杂场景(如遮挡、远距离)下的检测效果，该模型能够实现更精确的目标定位与分类，即使在点云极为稀疏的区域也保持了良好的鲁棒性。

将 FM-VXNet 与主流方法在 KITTI 测试集上对比，结果如表 1 所示。

Table 1. Comparison of the improved models
表 1. 改进后的模型对比

在汽车类别上(0.7-0.5-0.5)							
		AP-BEV			AP-3D		
模型	模态类型	easy	moderate	hard	easy	moderate	hard
MMono3D	C	5.22	5.19	4.13	2.53	2.31	2.31
3DOP	C	12.6	9.49	7.5	6.55	5.07	4.1

续表

VeloFCN	C	40.1	32	30.4	15.2	13.6	15.9
MV3D	C	86.2	77.3	76.3	71.2	56.6	55.3
VoxelNet	C	89.6	84.8	78.6	82	65.5	62.9
MV3D	C + L	86.6	78.1	76.7	71.3	62.7	56.6
F-PointNet	C + L	88.2	84	76.4	83.8	70.9	63.7
MVX-Net (VF)	C + L	88.6	84.6	78.6	82.3	72.2	66.8
MVX-Net (PF)	C + L	89.5	84.9	79	85.5	73.3	67.4
PointRCNN	L				88.9	78.6	77.4
PV-RCNN	L	95.8	91.1	88.9	92.6	84.8	82.7
VoxeI-RCNN	L	95.5	91.3	89	92.4	85.3	82.9
M3DETR	L				92.3	85.4	82.9
Octr	L				89.8	87	79.3
CLOCs	C + L	93.5	92	89.5	92.8	85.9	83.3
CAT-Det	C + L				90.1	81.5	79.3
DVF	C + L	96.2	91.7	89.2	93.1	85.8	83.1
MLF-Det	C + L				89.7	87.3	79.3
GraphAlign	C + L	95.7	92.8	91.4	92.4	87	84.7
SSLFusion	C + L	95.6	91.6	91.4	94.1	85.7	85.4
Ours	C + L	96.3	95.2	92.9	96.2	88.9	87.7

Figure 6. Example of FM-VXNet detection visualization
图 6. FM-VXNet 检测可视化示例

4.4. 消融实验

为验证各模块的独立贡献与协同作用，本研究以 MVX-Net (PointFusion)为基线，在 KITTI 数据集上逐步引入 BiDA、FFCM 与 Bi-CMGA 模块，并评估不同组合下的检测性能(见表 2)。

单模块结果表明，BiDA 在 hard 场景下提升最为显著(AP-3D 由 67.4%提高至 84.6%，+17.2%)，说明其密度感知机制能有效缓解远距稀疏体素的噪声干扰；FFCM 在 moderate/hard 场景分别提升 13.3%和 17.6% (由 73.3%/67.4%提高至 86.6%/85.0%)，验证频域–空间融合显著增强了图像分支的全局语义建模；Bi-CMGA 在各指标上均取得最优的单模块表现(AP-3D moderate 达 87.7%)，体现了双向跨模态注意力在语义与几何互补中的有效性。

两两组合进一步揭示了模块间的互补特性。BiDA + FFCM(a + b)在 3D 检测 moderate/hard 提升至 87.1%/86.0%，相较单模块最高值(86.6%/85.0%)分别提高约 0.5%~1.0%，表明几何净化与全局语义增强的结合能显著提升复杂场景的鲁棒性；BiDA + Bi-CMGA(a + c)获得最高的组合性能(3D moderate/hard: 88.5%/87.3%)，较单独 Bi-CMGA 提高约 0.8%~0.9%，说明“净化后融合”的策略能有效强化跨模态交互；FFCM + Bi-CMGA (b + c)亦取得显著增益(3D moderate/hard: 88.4%/87.1%)，但受稀疏噪声影响略低于 a + c。

当三者联合使用(a + b + c, FM-VXNet)时，模型在所有指标上均达到最优(BEV moderate: 95.2%, 3D moderate: 88.9%)，相较 a + c 进一步提升 0.4%。结果表明，FFCM、BiDA 与 Bi-CMGA 在几何稳定、语义一致性与跨模态融合三方面形成递进式协同，从而实现多模态 3D 检测性能的系统性提升。

Table 2. Ablation experiment
表 2. 消融实验

	AP-BEV			AP-3D		
	easy	moderate	hard	easy	moderate	hard
Baseline	89.5	84.9	79	85.5	73.3	67.4
a (Baseline + BiDA)	95.1	93.7	91.3	94.7	86.2	84.6
b (Baseline + FFCM)	94.9	93.9	91.5	94.5	86.6	85.0
c (Baseline + Bi-CMGA)	95.7	94.6	92.1	95.5	87.7	86.4
a + b (BiDA + FFCM)	95.6	94.4	91.9	95.4	87.1	86.0
a + c (BiDA + Bi-CMGA)	96.1	95.0	92.6	96.0	88.5	87.3
b + c (FFCM + Bi-CMGA)	96.0	94.9	92.5	95.8	88.4	87.1
a + b + c (Ours, FM-VXNet)	96.3	95.2	92.9	96.2	88.9	87.7

5. 结论与展望

本文针对多模态 3D 目标检测中图像全局建模不足、跨模态融合简单、体素特征鲁棒性差这三个核心问题，提出了 FM-VXNet 模型，该模型通过引入 FFCM 模块，并精心设计了 Bi-CMGA 和 BiDA 模块，分别从图像特征提取、跨模态融合以及 LiDAR 点云体素特征优化三个关键层面，对多模态 3D 目标检测任务进行了全面改进。在 KITTI 数据集上的实验结果充分表明，FM-VXNet 在性能上显著超越 MVX-Net、BEVFusion 等现有主流方法，为多模态 3D 目标检测领域提供了一种创新且高效的解决方案。未来工作将

集中于以下方面：一是探索端到端训练策略，进一步提升模型效率与精度；二是将模型扩展至更复杂的多类别目标检测任务中。

参考文献

[1] 顾芳铭, 况博裕, 许亚倩, 等. 面向自动驾驶感知系统的对抗样本攻击研究综述[J]. 信息安全研究, 2024, 10(9): 786-794.

[2] 尹彦鑫, 孟志军, 赵春江, 等. 大田无人农场关键技术研究现状与展望[J]. 智慧农业(中英文), 2022, 4(4): 1-25.

[3] 王若萱, 吴建平, 徐辉. 自动驾驶汽车感知系统仿真的研究及应用综述[J]. 系统仿真学报, 2022, 34(12): 2507-2521.

[4] 魏海跃, 杨奎河. 自动驾驶场景下的多模态 3D 目标检测算法[J]. 长江信息通信, 2024, 37(6): 28-30.

[5] 代振钊. 面向自动驾驶的多模态融合感知技术研究[D]: [硕士学位论文]. 北京: 北方工业大学, 2024.

[6] Liu, C., Gao, C., Liu, F., Liu, J., Meng, D. and Gao, X. (2022) SS3D: Sparsely-Supervised 3D Object Detection from Point Cloud. 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), New Orleans, 18-24 June 2022, 8418-8427. https://doi.org/10.1109/cvpr52688.2022.00824

[7] Wang, Y., Yin, J., Li, W., Frossard, P., Yang, R. and Shen, J. (2023) SSDA3D: Semi-Supervised Domain Adaptation for 3D Object Detection from Point Cloud. *Proceedings of the AAAI Conference on Artificial Intelligence*, **37**, 2707-2715. https://doi.org/10.1609/aaai.v37i3.25370

[8] Bai, Z., Wu, G., Barth, M.J., Liu, Y., Sisbot, E.A. and Oguchi, K. (2023) VINet: Lightweight, Scalable, and Heterogeneous Cooperative Perception for 3D Object Detection. *Mechanical Systems and Signal Processing*, **204**, Article 110723. https://doi.org/10.1016/j.ymssp.2023.110723

[9] Xu, W., Jin, J., Xu, F., Li, Z. and Tao, C. (2023) Denoising and Reducing Inner Disorder in Point Clouds for Improved 3D Object Detection in Autonomous Driving. *Electronics*, **12**, Article 2364. https://doi.org/10.3390/electronics12112364

[10] Huang, K.C., Wu, T.H., Su, H.T., *et al*. (2022) MonoDTR: Monocular 3D Object Detection with Depth-Aware Transformer. 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), New Orleans, 18-24 June 2022, 4002-4011. https://doi.org/10.1109/cvpr52688.2022.00398

[11] Lian, Q., Ye, B., Xu, R., Yao, W. and Zhang, T. (2022) Exploring Geometric Consistency for Monocular 3D Object Detection. 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), New Orleans, 18-24 June 2022, 1675-1684. https://doi.org/10.1109/cvpr52688.2022.00173

[12] Nakatsuka, C. and Komorita, S. (2021) Denoising 3D Human Poses from Low-Resolution Video Using Variational Autoencoder. 2021 *IEEE/RSJ International Conference on Intelligent Robots and Systems* (*IROS*), Prague, 27 September-1 October 2021, 4625-4630. https://doi.org/10.1109/iros51168.2021.9636144

[13] Zhang, C., Chen, W., Wang, W. and Zhang, Z. (2024) MA-ST3D: Motion Associated Self-Training for Unsupervised Domain Adaptation on 3D Object Detection. *IEEE Transactions on Image Processing*, **33**, 6227-6240. https://doi.org/10.1109/tip.2024.3482976

[14] Li, P. and Jin, J. (2022) Time3D: End-to-End Joint Monocular 3D Object Detection and Tracking for Autonomous Driving. 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), New Orleans, 18-24 June 2022, 3875-3884. https://doi.org/10.1109/cvpr52688.2022.00386

[15] Zhang, C., Wang, H., Cai, Y., Chen, L., Li, Y., Sotelo, M.A., *et al*. (2022) Robust-FusionNet: Deep Multimodal Sensor Fusion for 3-D Object Detection under Severe Weather Conditions. *IEEE Transactions on Instrumentation and Measurement*, **71**, 1-13. https://doi.org/10.1109/tim.2022.3191724

[16] Sindagi, V.A., Zhou, Y. and Tuzel, O. (2019) MVX-Net: Multimodal VoxelNet for 3D Object Detection. 2019 *International Conference on Robotics and Automation* (*ICRA*), Montreal, 20-24 May 2019, 7276-7282. https://doi.org/10.1109/icra.2019.8794195

[17] Xia, B., Zhou, J., Kong, F., You, Y., Yang, J. and Lin, L. (2024) Enhancing 3D Object Detection through Multi-Modal Fusion for Cooperative Perception. *Alexandria Engineering Journal*, **104**, 46-55. https://doi.org/10.1016/j.aej.2024.06.025

[18] Chu, H., Liu, H., Zhuo, J., Chen, J. and Ma, H. (2024) Occlusion-Guided Multi-Modal Fusion for Vehicle-Infrastructure Cooperative 3D Object Detection. *Pattern Recognition*, **157**, Article 110939. https://doi.org/10.1016/j.patcog.2024.110939

[19] Ding, B., Xie, J., Nie, J., Wu, Y. and Cao, J. (2024) C2BG-Net: Cross-Modality and Cross-Scale Balance Network with Global Semantics for Multi-Modal 3D Object Detection. *Neural Networks*, **179**, Article 106535. https://doi.org/10.1016/j.neunet.2024.106535

[20] Song, Z., Wei, H., Bai, L., Yang, L. and Jia, C. (2023) GraphAlign: Enhancing Accurate Feature Alignment by Graph Matching for Multi-Modal 3D Object Detection. 2023 *IEEE*/*CVF International Conference on Computer Vision* (*ICCV*), Paris, 1-6 October 2023, 3335-3346. https://doi.org/10.1109/iccv51070.2023.00311

[21] Chen, X., Kundu, K., Zhu, Y., *et al*. (2015) 3D Object Proposals for Accurate Object Class Detection. *Advances in Neural Information Processing Systems*, Montreal, 7-12 December 2015, 424-432.

[22] Shi, S., Guo, C., Jiang, L., Wang, Z., Shi, J., Wang, X., *et al*. (2020) PV-RCNN: Point-Voxel Feature Set Abstraction for 3D Object Detection. 2020 *IEEE*/*CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Seattle, 13-19 June 2020, 10526-10535. https://doi.org/10.1109/cvpr42600.2020.01054

[23] Deng, J., Shi, S., Li, P., Zhou, W., Zhang, Y. and Li, H. (2021) Voxel R-CNN: Towards High Performance Voxel-Based 3D Object Detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, **35**, 1201-1209. https://doi.org/10.1609/aaai.v35i2.16207

[24] Chen, X.Z., Ma, H.M., Wan, J., Li, B., *et al*. (2017) Multi-View 3D Object Detection Network for Autonomous Driving. 2017 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Honolulu, 21-26 July 2017, 6526-6534. https://doi.org/10.1109/cvpr.2017.691

[25] Ding, B., Xie, J., Nie, J. and Cao, J. (2025) SSLFusion: Scale and Space Aligned Latent Fusion Model for Multimodal 3D Object Detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, **39**, 2735-2743. https://doi.org/10.1609/aaai.v39i3.32278

[26] Li, Z., Gu, J., Li, K., *et al*. (2023) DVF: Dynamic Voxel Fusion for 3D Object Detection in Point Clouds. *Proceedings of the IEEE*/*CVF Conference on Computer Vision and Pattern Recognition*, Vancouver, 17-24 June 2023, 17580-17589.

[27] Liu, Z., Tang, H., Amini, A., *et al*. (2022) BEVFusion: A Simple and Robust LiDAR-Camera Fusion Framework. *Advances in Neural Information Processing Systems*, **35**, 10421-10434.