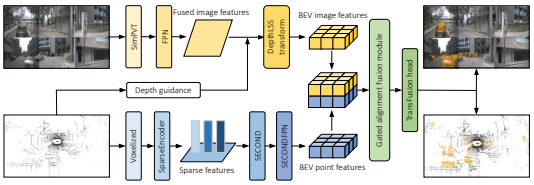
OEE | 光电工程 Opto-Electronic Engineering

Article

2026 年 4 月, 第 53 卷, 第 4 期

多层感知增强与门控对齐融合的 BEV 目标检测方法



刘鸿恩[1,2], 李　旭[1*], 黄　晟[1], 康　婷[3]

[1]磁浮轨道交通装备江西省重点实验室，江西 赣州 341000；

[2]江西理工大学永磁磁浮技术与轨道交通研究院，江西 赣州 341000；

[3]江西理工大学电气工程与自动化学院，江西 赣州 341000

摘要：针对多模态 3D 目标检测中图像语义受限、点云稀疏致几何退化与跨模态对齐困难等问题，提出一种多层感知增强与门控对齐融合的 BEV 目标检测方法 (MEPFusion)。首先图像端利用 SimPVT 强化全局语义与关键区域建模；其次点云端设计空间密度调制卷积，以位置依赖权重注入密度先验，稳固稀疏场景的局部几何；最后融合端构建双模态门控对齐模块，通过通道与空间门控抑制错配并实现动态校准，促成 BEV 特征的动态对齐与互补交互。实验结果表明，本文方法在 nuScenes 上的 mAP 与 NDS 达到 67.1% 与 70.8%，较 BEVFusion 分别提升 2.8% 和 1.7%，显著提高检测精度与鲁棒性。消融结果验证 MEPFusion 在语义特征提取、几何建模与跨模态对齐三方面具备独立增益并产生协同效应，整体上在融合质量、检测性能与计算效率之间实现更优平衡，为自动驾驶三维目标检测提供高效可靠的方案。

关键词：自动驾驶；多模态融合；BEV 表征；多层感知增强；门控对齐融合

The English title and abstract appear at the end of the paper.

DOI: 10.12086/oee.2026.250340 | CSTR: 32245.14.oee.2026.250340 | 中图分类号：TP391.4 | 文献标识码：A

刘鸿恩，李旭，黄晟，等. 多层感知增强与门控对齐融合的 BEV 目标检测方法 [J]. 光电工程，2026，**53**(4): 250340

Liu H, Li X, Huang S, et al. BEV object detection method with multi-layer perceptual enhancement and gated alignment fusion[J]. *Opto-Electron Eng*, 2026, **53**(4): 250340

1　引　言

自动驾驶的规模化应用使环境感知成为衡量整车智能水平的关键指标，其对复杂场景的语义理解与快速响应能力决定行车安全与稳定性[1]。作为底层支撑模块，环境感知的重要性持续提升，并对精度、效率与实时性的协同优化提出更高要求。三维目标检测是感知系统的关键环节，除承担类别识别外，还需准确估计目标的尺寸、位置、姿态与速度等关键状态量，为轨迹预测、路径规划与紧急避障提供可信输入。在真实道路环境中，远距小目标、强遮挡、复杂气象与多源噪声交织出现，单一传感器难以长期保持稳定性能[2]。近年来，纯视觉方案在目标检测任务中取得了显著进展，在多个自动驾驶公开数据集上实现了接近最优的性能。然而，纯视觉方案尽管能够提供丰富的语义信息，但在光照变化、空间结构感知和深度估计方面仍然存在较大挑战。相比之下，激光雷达 (LiDAR) 能够提供高精度的几何数据与深度信息，但在远距离与雨雾等工况下回波稀疏[3]。然而，单一传感器的局限性在遮挡和低可见度的情况下愈发明显。由此引发的语义表达受限与几何建模缺陷在远距离

收稿日期：2025-11-06；修回日期：2026-01-11；录用日期：2026-01-12；网络出版日期：2026-04-25

基金项目：国家自然科学基金资助项目 (62463011)；国家重点研发计划项目子课题 (2023YFB4302104-2)；江西省自然科学基金面上项目 (20224BAB202025)

*通信作者：李旭，885239570@qq.com

版权所有©2026 中国科学院光电技术研究所

与遮挡等条件下进一步放大，成为制约整体检测表现的关键瓶颈[4]。为突破上述限制，结合了激光雷达和相机的多模态融合系统被广泛认为是提升自动驾驶感知质量的有效路径，能够更有效地融合几何信息，从而在处理远距离、小目标、遮挡及恶劣天气等场景时，提供不可替代的安全性优势。其核心是在统一的时空与表征基准上构建可学习的互补推理机制[5]。该范式在尺度锚定、稠密语义补全与冗余互证环节产生协同增益，从而显著降低远距与遮挡场景的不确定性，并增强在复杂工况下的鲁棒性。为将上述协同转化为工程可部署方案，亟须一种统一且计算友好的交互载体。鸟瞰图表示 (Bird's-eye view, BEV) 作为平面化的三维空间表征，通过视图变换将多视角图像与点云对齐至同一地面坐标系，在二维计算预算内保留关键三维几何关系，弱化跨视角配准误差，并为后续的特征对齐与融合提供统一空间基准，从而有助于提升远距小目标与复杂遮挡场景的召回与定位精度[6]。

围绕 BEV 空间的点云与图像的相关算法，现有研究依输入模态大体可分为三类[1]。基于 BEV 的视觉三维检测，典型流程通常遵循“视图变换→BEV 空间聚合→BEV 头检测”。其中，LSS[7] 通过为每个像素预测离散深度分布，将二维特征沿光线方向反投影到三维，并在地面坐标系中进行体素池化或直接 splat 至 BEV 网格。代表性方法如 BEVDet[8] 以高效视图变换将多视角特征映射到 BEV，兼顾速度与精度；BEVDepth[9] 在训练中加入显式深度监督，并配合体素池化等设计，缓解由尺度与距离不确定引起的误差传播。此外，BEVFormer[10] 将 BEV 查询与图像平面对齐并融合时序 Transformer，以提升动态场景建模能力。然而，由于相机缺乏真实深度且易受照明、遮挡与纹理重复影响，纯视觉 BEV 在远距小目标与强遮挡工况下仍容易出现深度分配不稳、框体尺度偏差与虚警增多的问题。

与此相对，基于 LiDAR 的 BEV 检测以几何精度为基础，通过体素或柱体化将不规则点集转化为规则网格，再在 BEV 或体素上进行特征提取与目标回归[11]。PointPillars[12] 以柱状体素和二维卷积实现实时检测；SECOND[13] 通过稀疏卷积在计算代价与几何完整性之间取得平衡；CenterPoint[14] 以中心点回归替代锚框设计，兼顾精度与后处理效率。尽管 LiDAR 在几何上更可靠，但在远距、雨雾与稀疏回波情况下点云密度骤降，卷积核“等权聚合”的倾向会弱化细窄结构与物体边界，导致小目标漏检与姿态估计不稳；体素尺寸的缩小虽有利于细节保留，却显著增加内存与算力，带来部署约束。

在单模态路径逼近瓶颈后，多模态 BEV 融合成为进一步突破方向。其基本思想是在统一的 BEV 空间内将视觉的语义与 LiDAR 的几何进行互补耦合，以在维持推理效率的同时提升性能上限。按融合时机与粒度划分，早期策略在投影前于图像平面或点级别交互，信息充分但对标定与时序同步更敏感；后期策略在各自 BEV 表征形成后再交互，工程稳定但对细粒度错配的纠正能力有限；中期策略则在多尺度间反复交互，力求兼顾语义与几何一致性[15]。代表工作方面，TransFusion[16] 在 LiDAR-BEV 上引入基于注意力的软关联机制，以自适应汇聚图像分支信息，从而减轻模态分布差异导致的“生硬拼接”；BEVFusion[17] 将图像经视图变换与 LiDAR 特征同时统一到 BEV，再端到端融合，在精度与效率之间取得良好平衡；MVP[18]、AutoAlignV2[19]、UVTR[20] 等方法分别从显式几何约束、可学习对齐与 Transformer 深度交互角度强化跨模态协同。尽管体系日趋成熟，多模态 BEV 感知仍普遍面临：视图变换量化误差与视觉深度不确定共同引发的语义——几何错位；点云稀疏性与视角覆盖差异导致远距与遮挡处边界线索被弱化，诱发小目标漏检与尺度偏差；跨模态噪声与冗余在融合阶段被放大；且缺乏逐位置、逐通道的选择性机制时，全局交互难以兼顾局部几何约束与计算预算，限制分辨率、时延与鲁棒性的协同优化空间。

基于上述分析，本文提出面向自动驾驶的多层感知增强与门控对齐融合的 BEV 目标检测算法 (multilayer perception enhancement and gated alignment fusion, MEPFusion)，围绕语义表征增强、稀疏几何建模与跨模态对齐三条主线协同推进。在图像分支，构建 SimPVT 强化全局与多尺度语义一致性并放大关键区域响应，为后续视图变换提供更稳定的语义先验；在点云分支，提出空间密度调制卷积，在不增加参数量的前提下对卷积核实施位置相关的密度调制，强化远距与稀疏场景中的边界与细节表达；在融合阶段，设计门控对齐融合模块，于 BEV 空间学习逐位置与逐通道的门控与软对齐，动态突出优势模态、抑制冗余互扰，从而在复杂交通场景下实现高精度、低时延与强鲁棒的三维感知。在与基线等参数与等时延配置下的系统对比与消融验证中，所提方案在 mAP/NDS 指标上获得稳定增益，并在视觉语义、几何建模与跨模态对齐三个层面呈现相互支撑的协同效应，整体体现出在多场景条件下的精度提升与鲁棒性增强。

本文的主要贡献概括如下：

1) 语义全局增强。在图像分支构建 SimPVT，强化全局多尺度语义一致性并放大关键区域响应，稳固视图变换前的语义先验。

2) 几何建模强化。在点云分支提出空间密度调制卷积，以位置相关的空间密度调制替代等权聚合，在不增加参数量的前提下提升稀疏、远距场景的几何边界与细节建模能力。

3) 融合对齐机制。设计双模态门控对齐融合模块，于 BEV 空间学习逐位置与逐通道的门控与软对齐，动态突出优势模态、抑制冗余互扰，实现高效稳健的跨模态互补。

4) 性能与实证。在等参数与等时延条件下取得稳定性能增益 (mAP/NDS 均提升)，并通过系统消融验证各模块的独立贡献与协同效应，兼顾精度、效率与可部署性。

2 方　法

2.1 总体框架

在多模态三维感知任务中，图像与点云分别在语义表达与几何建模方面具有互补优势，但两者在特征分布、空间组织及时间采样频率上的差异，常导致语义特征提取不足、几何结构刻画不精确以及跨模态对齐偏差等问题。为此，本文提出一种基于 BEVFusion 结构优化的多层感知增强与门控对齐融合的 BEV 目标检测算法 MEPFusion，从语义建模、几何表达与多模态融合三方面对原框架进行协同增强。本文 MEPFusion 方法的总体架构如图 1 所示。

本文 MEPFusion 方法采用相机&LiDAR 双分支结构，并在 BEV 空间实现可学习对齐与融合。图像分支以 PVTv2[21] 为金字塔主干，结合 SimAM[22] 的显著性重标定获取多尺度且判别性更强的语义特征；随后结合显式深度监督与 LSS 视图变换，将二维特征反投影并聚合为统一 BEV 表示，实现由图像平面到鸟瞰坐标的几何对齐。点云分支在体素编码阶段利用空间密度调制卷积，依据局部点密度对卷积响应进行位置相关调制，以强化稀疏与远距场景下的结构建模，并通过 BEVPooling 沿高度维聚合将体素特征压缩至 BEV 网格。两路 BEV 特征输入门控对齐融合模块，以逐位置门控与软对齐实现动态加权与互补交互，得到统一的多模态 BEV 表征；最终经 FPN 与 TransFusion 检测头进行多尺度解码，完成高精度三维目标预测[23]。

2.2 图像分支

在算力受限下，图像分支旨在生成跨视角一致且对目标边界更敏感的语义表征，从而降低视图变换后的深度不确定性，提升 BEV 投影的稳定性。为此，本文构建兼顾全局建模与局部显著性的特征提取网络 SimPVT (图 2)，采用 PVTv2 全局建模和 SimAM 局部显著性重标定的两段式结构。其中，PVTv2 作为金字塔式主干，在多尺度空间内高效刻画长程依赖实现全局表征学习。同时，在各尺度输出端嵌入轻量无参注意力 SimAM，以元素级显著性重标定放大关键区域响应并抑制背景噪声，从而形成兼具全局一致性与局部判别力的输入特征。该分支在不显著增加参数与计算开销下，能够放大关键区域响应、抑制背景噪声，并提升跨视角语义分布稳定性。

1) PVTv2 全局建模。首先输入经重叠式 Patch

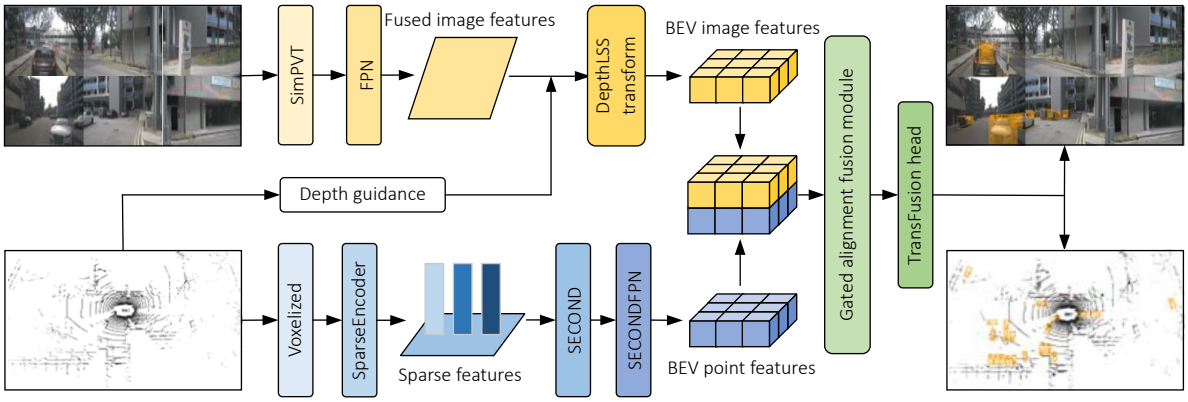


图 1 MEPFusion 方法总体架构图

Fig. 1 Overall framework of the proposed MEPFusion method

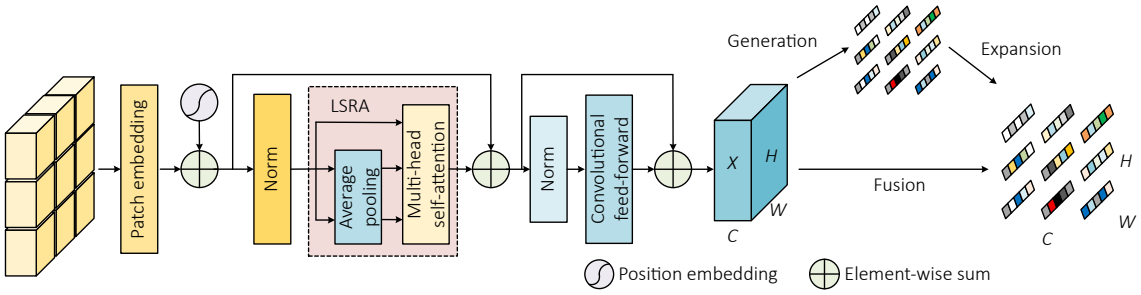


图 2 特征提取 SimPVT 结构图

Fig. 2 Architecture of the SimPVT feature extraction module

Embedding 得到初始特征，随后进入四阶段 Transformer 编码逐级构建由高分辨率到低分辨率的金字塔特征$\{\boldsymbol{F}_1, \boldsymbol{F}_2, \boldsymbol{F}_3, \boldsymbol{F}_4\}$。鉴于后续结构仅需三尺度输入，本工作舍弃最高分辨率的 $\boldsymbol{F}_1$，仅保留 $\{\boldsymbol{F}_2, \boldsymbol{F}_3, \boldsymbol{F}_4\}$ 作为多尺度特征集，以降低高分辨率分支带来的额外计算开销，并抑制低层纹理细节对高层语义表征的干扰。其中，各 Transformer 编码阶段通过 LSRA (linear spatial-reduction attention)，先对 Key、Value 的空间维度进行线性压缩以缩短序列，再执行缩放点积注意力完成全局依赖聚合，最后在理论复杂度上由L^2缩减为L^2/r^2，其中$L=H\times W$表示输入特征展平后的序列长度，r表示空间压缩比，从而在长程依赖建模与计算开销之间取得平衡。LSRA 的正式定义为:

$$LSRA(x)=Concat(\boldsymbol{h}_0,\cdots,\boldsymbol{h}_{N-1})\boldsymbol{W}^{o}, \tag{1}$$

$$\boldsymbol{h}_j=Att(x\boldsymbol{W}_j^{Q},S(x)\boldsymbol{W}_j^{K},S(x)\boldsymbol{W}_j^{V}), j\in\{0,\cdots,N-1\}, \tag{2}$$

$$S(x)=LN(Avg(x,p)\cdot\boldsymbol{W}^{S}), \tag{3}$$

式中：$x\in\mathbb{R}^{H\times W\times C_i}$表示输入特征，$H$、$W$和$C$分别表示特征图的高、宽和通道数；$\boldsymbol{h}_j$是第 j 个注意力头的输出；N为头的数量；$Concat(\cdot)$表示通道级联；$\boldsymbol{W}^{o}\in\mathbb{R}^{C_o\times C_o}$是将级联特征映射到输出维度$C_o$的输出投影矩阵；矩阵$\boldsymbol{W}_j^{Q}$、$\boldsymbol{W}_j^{K}$、$\boldsymbol{W}_j^{V}\in\mathbb{R}^{C_i\times d}$是第 j 个头的查询、键和值投影，$d=C_o/N$作为每个头通道维度；$Att(\cdot)$表示缩放的点积注意力；$S(x)$表示输入 x 的空间简化表示；$Avg(x,p)$表示x上具有$p\times p$窗口的平均池化；$\boldsymbol{W}^{S}\in\mathbb{R}^{C_i\times C_i}$为信道投影矩阵；$LN(\cdot)$为层归一化。

2) SimAM 局部重标定。在获得的多尺度特征$\{\boldsymbol{F}_2, \boldsymbol{F}_3, \boldsymbol{F}_4\}$基础上，逐尺度执行无参的 SimAM 重标定。首先依据通道内均值与方差构造能量函数$e_{\min}^{t}$，度量单个位置与同通道其余位置的可分性；接着将由能量函数得到的响应值$\boldsymbol{W}$经 Sigmoid 映射为权重张量$\tilde{\boldsymbol{W}}$，最后与原特征逐元素相乘完成显著性增强。该过程零参数、仅含元素级运算，能够增强关键区域与远距小目标响应并抑制背景噪声，提升跨视角语义稳健性。过程数学公式如下。

$$e_{\min}^{t}=\frac{4(\tilde{\sigma}^2+\lambda)}{(t-\mu_t)^2+2\tilde{\sigma}_t^2+2\lambda}, \tag{4}$$

$$\tilde{\boldsymbol{W}}=Sigmoid\left(\frac{1}{\boldsymbol{E}}\right)\odot\boldsymbol{W}, \tag{5}$$

式中：μ_t、$\tilde{\sigma}_t^2$和λ分别表示在排除目标神经元后的均值、方差和数值稳定常数；t为神经元；$\boldsymbol{E}$为所有神经元能量值组合的张量。

2.3 点云分支

鉴于 LiDAR 点云在空间分布上高度不均且整体稀

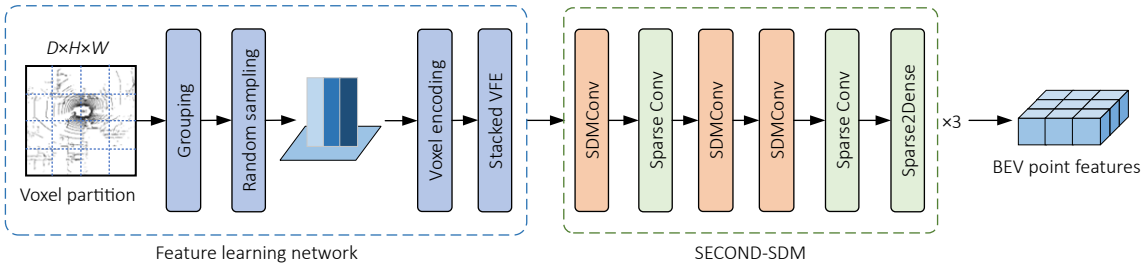


图 3 SDMConv 增强的点云分支结构

Fig. 3 Architecture of the point cloud branch enhanced by SDMConv

疏，标准 3×3 稀疏卷积对核内各相对位置等权聚合，易在远距稀疏、遮挡与噪声条件下产生“均值化”效应，进而削弱窄带边界与线状几何细节的判别性。为此，本文在点云分支设计 FLN (feature learning network)→SECOND-SDM 的二级路径，其结构如图 3 所示。其中，$D\times H\times W$ 表示体素化后特征张量的空间尺寸，D 表示竖直方向的离散层数，H 和 W 分别表示 BEV 平面两个方向上的离散网格数。

具体而言，首先将原始点云按规则网格进行体素化，并在每个体素内通过随机采样获得定长点集，随后经 Voxel Encoding 和 Stacked VFE 将点级属性逐层聚合与压缩，构建稀疏四维体素张量并输入 SECOND 主干。其中，主干由三阶段 SECOND-SDM 顺序组成，各阶段以空间密度调制卷积 (spatial density-modulated convolution, SDMConv) 为核心，通过位置密度调制抑制等权聚合导致的信息稀释，逐级提炼特征并维持合理的尺度组织；阶段之间通过 *stride*=2 的稀疏卷积完成尺度切换与感受野扩展[24]。最终，Sparse2Dense 沿高度维聚合得到 BEV 平面表征并输入融合模块。该路径在网络深度、通道配置与下采样倍率上保持稳定层级，确保从点级信息到 BEV 表示的高效映射与可复现实验流程。

在此总体路径中，SDMConv 作为默认的 3×3 空间算子贯穿主干残差/堆叠块：前向阶段预置固定、中心对称、秩为 1 的位置先验 $\boldsymbol{\Phi}$，并以 Hadamard 逐元素方式对卷积核实施位置依赖的权重调制，形成“中心增强—外围递减”的响应分布。由此，在保持计算可比性的同时，有效抑制远距稀疏与遮挡噪声引发的等权“均值化”，显著强化 BEV 平面上细窄边界与局部几何回波的保真度，提升几何判别性、小目标可分性与整体检测稳健性。SDMConv 数学公式如下。

$$(\boldsymbol{I}*\boldsymbol{W}_{\phi})_{ij}^{f}=\sum_{a=1}^{K}\sum_{b=1}^{K}\boldsymbol{\Phi}_{a,b}\cdot\omega_{a,b}^{f}\cdot\boldsymbol{I}_{i+a-K_a+1,j+b-K_b+1}\,, \tag{6}$$

式中：$\boldsymbol{I}\in\mathbb{R}^{R\times C}$ 为输入信号；(i,j) 为网格坐标；$\boldsymbol{W}_{\phi}\in\mathbb{R}^{K_a\times K_b\times F}$ 为卷积核张量；$\omega_{a,b}^{f}$ 为第 f 个核在相对位置 (a,b) 处的权重；离散密度函数 $\boldsymbol{\Phi}\in\mathbb{R}^{K_a\times K_b}$ 为空间位置先验；元素 $\boldsymbol{\Phi}_{ab}$ 为调制核内位置 (a,b) 贡献。

2.4 模态融合

在多模态 3D 目标检测中，尽管将图像与激光雷达特征统一至 BEV 空间有助于发挥两者在语义理解与几何建模方面的互补优势，但两种模态在信息来源、感知维度与采样密度上的差异往往导致 BEV 特征语义不一致与空间错位，简单拼接或均值融合不仅难以缓解这一矛盾，反而易将噪声引入最终表征，从而削弱检测性能[25]。为此，本文设计门控对齐融合模块 (gated alignment fusion module, GAFM)，通过轻量化空间注意力机制在 BEV 网格逐位置学习空间相关的门控权重，使关键区域突出、冗余区域抑制，从而提高融合表征的一致性与判别性[26]。具体而言，本文在 BEV 网格位置 (x,y) 上将对检测贡献更显著且更稳定的局部响应区域定义为关键区域，其特征为该位置至少存在一种模态提供相对可靠的语义或几何证据；而背景噪声占优、有效响应较弱，或两模态在该位置出现不一致表征的区域定义为冗余区域。GAFM 在 BEV 网格上为两模态学习逐位置的空间门控权重，以选择性加权的方式进行融合，使空间上呈现“优势模态主导、弱证据抑制”的聚合特性，从而避免简单拼接将冗余信息一并引入检测头。

如图 4 所示，GAFM 采用分离-聚合的总体范式，先完成模态内显著性重标定，再进行跨模态互导校正，最后实施空间门控聚合操作。具体而言，先在各自模态

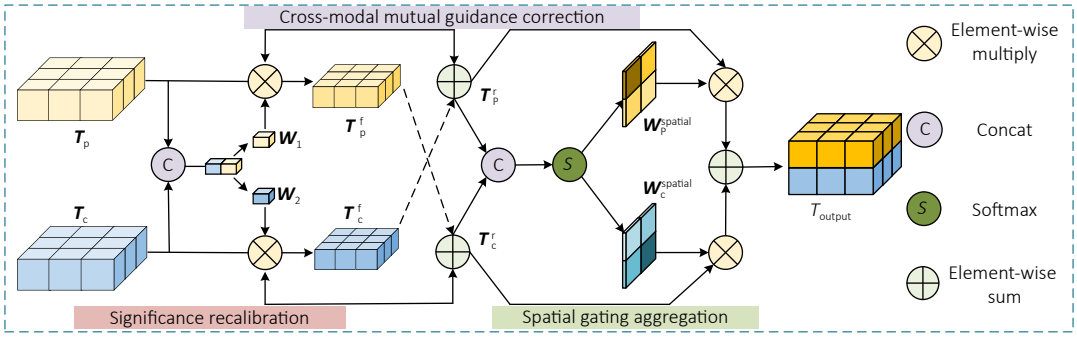


图 4 门控对齐融合模块

Fig. 4 Gated alignment fusion module

内部做针对性增强，使图像侧更聚焦全局语义与类别判别、点云侧更突出几何边界与局部结构；继而在两模态之间进行轻量互导以削弱系统性偏差并互补细节；最后在模态维实施可解释的门控加权，输出语义完整且空间一致的融合 BEV 特征，用作检测头的高质量输入。在此过程中，首先，BEV 图像特征$\boldsymbol{T}_{\mathrm{p}}$和点云特征$\boldsymbol{T}_{\mathrm{c}}$分别经通道注意力强化全局上下文信息表达，得到增强后特征$\boldsymbol{T}_{\mathrm{p}}^{\mathrm{f}}$、$\boldsymbol{T}_{\mathrm{c}}^{\mathrm{f}}$。该过程本质上是通道维的显著性重标定，增强与类别语义判别及目标整体轮廓和结构表征相关的通道响应，同时抑制背景噪声与点云稀疏性导致的激活不稳定，从而在模态内先行净化特征表示，为后续跨模态互导校正与空间门控聚合提供更具判别性的输入。其数学表达式分别为

$$\boldsymbol{T}_{\mathrm{p}}^{\mathrm{f}}=\sigma(\boldsymbol{W}_2\cdot ReLU(\boldsymbol{W}_1\cdot Gap(Concat(\boldsymbol{T}_{\mathrm{p}},\boldsymbol{T}_{\mathrm{c}}))))\cdot\boldsymbol{T}_{\mathrm{p}}\,, \tag{7}$$

$$\boldsymbol{T}_{\mathrm{c}}^{\mathrm{f}}=\sigma(\boldsymbol{W}_2\cdot ReLU(\boldsymbol{W}_1\cdot Gap(Concat(\boldsymbol{T}_{\mathrm{p}},\boldsymbol{T}_{\mathrm{c}}))))\cdot\boldsymbol{T}_{\mathrm{c}}\,, \tag{8}$$

式中：σ表示 Sigmoid 激活函数；$Gap(\cdot)$表示全局平均池化操作；$\boldsymbol{W}_1$、$\boldsymbol{W}_2$为两层全连接层的权重，用于生成通道加权向量。

其次，为增强模态间信息协同而引入残差校正策略，构建点云对图像的引导关系，对图像特征进行自适应校准以有效去除干扰信息，获得优化后的图像特征表示$\boldsymbol{T}_{\mathrm{p}}^{\mathrm{r}}$。同时，利用图像丰富的语义与细节信息对点云特征进行补充，增强其空间结构表达能力,得到互补点云特征$\boldsymbol{T}_{\mathrm{c}}^{\mathrm{r}}$，进一步提升融合特征的表达能力。需要强调的是，残差校正除跨模态互补外，还承担局部消噪与一致性约束：在错配位置削弱不可靠分量、保留一致证据，从而提升关键区域的有效响应并抑制冗余干扰。其数学表达式分别为

$$\boldsymbol{T}_{\mathrm{p}}^{\mathrm{r}}=\boldsymbol{T}_{\mathrm{c}}^{\mathrm{f}}+\boldsymbol{T}_{\mathrm{p}}\,, \tag{9}$$

$$\boldsymbol{T}_{\mathrm{c}}^{\mathrm{r}}=\boldsymbol{T}_{\mathrm{p}}^{\mathrm{f}}+\boldsymbol{T}_{\mathrm{c}}\,, \tag{10}$$

然后，通过 1×1 卷积计算每种模态特征的空间注意力权重，并利用 Softmax 函数归一化空间权重，确保在每个 BEV 位置上融合权重的动态分配，从而在目标证据显著处由更可靠模态主导，而背景或冲突区域的贡献被自动压低，从而避免噪声放大并提升融合判别性。最后，依据注意力权重对校准后的图像与点云特征进行加权求和，生成语义完整、空间一致的融合特征表示$\boldsymbol{T}_{\mathrm{output}}$，实现关键区域突出与冗余区域抑制，为后续目标检测提供高质量的 BEV 融合特征输入。其数学表达式分别为

$$\boldsymbol{W}_{\mathrm{p}}^{\mathrm{spatial}}=Softmax\left(Conv1\times1\left(Concat\left(\boldsymbol{T}_{\mathrm{c}}^{\mathrm{f}},\boldsymbol{T}_{\mathrm{p}}^{\mathrm{f}}\right)\right)\right), \tag{11}$$

$$\boldsymbol{W}_{\mathrm{c}}^{\mathrm{spatial}}=Softmax\left(Conv1\times1\left(Concat\left(\boldsymbol{T}_{\mathrm{c}}^{\mathrm{f}},\boldsymbol{T}_{\mathrm{p}}^{\mathrm{f}}\right)\right)\right), \tag{12}$$

$$\boldsymbol{T}_{\mathrm{output}}=\boldsymbol{W}_{\mathrm{p}}^{\mathrm{spatial}}\cdot\boldsymbol{T}_{\mathrm{p}}^{\mathrm{f}}+\boldsymbol{W}_{\mathrm{c}}^{\mathrm{spatial}}\cdot\boldsymbol{T}_{\mathrm{p}}^{\mathrm{f}}\,, \tag{13}$$

式中：$\boldsymbol{W}_{\mathrm{p}}^{\mathrm{spatial}}$、$\boldsymbol{W}_{\mathrm{c}}^{\mathrm{spatial}}$为动态空间加权权重；$Conv1\times1$为 1×1 卷积层操作。

3 实验结果与分析

3.1 数据集与评价指标

本文基于大规模多模态自动驾驶数据集 nuScenes 开展实验。该数据集包含约 1000 个真实城市驾驶场景，配备 6 个环视相机获取环境纹理信息，1 个 32 线 LiDAR 感知环境深度信息，以及 5 个毫米波雷达提供速度与运动特征信息。同时提供 23 类精细目标标注与高清地图，覆盖多样交通参与体与复杂环境，是自动驾驶领域最具代表性的大规模多模态 3D 感知数据集之一,为本文方法提供了有效性验证。本文仅使用环视相机与 LiDAR 两种模态数据进行训练与评估，严格遵循官方训练集和验证集划分与评测协议。采用 nuScenes 官方指标体系，使用平均精度均值 (mean average precision, mAP) 和 nuScenes 检测分数 (nuScenes detection score, NDS) 作为主要评价指标。评价指标的定义如下：

1) mAP 衡量预测结果在不同类别上的整体检测性能，其定义为所有类别平均精度的均值，值越高表示模型检测效果越优，公式如下:

$$mAP=\frac{1}{C}\sum_{c=1}^{C}\int_{0}^{1}P_{\mathrm{c}}(r)\mathrm{d}r\,, \tag{14}$$

式中：C 表示类别总数；$P_{\mathrm{c}}(r)$表示在召回率 r 下的精确率。

2) NDS 在 mAP 的基础上进一步综合预测目标检测分数，数值越高表示模型整体检测性能越优公式如下：

$$NDS=\frac{1}{10}\left(5\cdot mAP+\sum_{m\in M}(1-\min(1,m))\right), \tag{15}$$

式中：$M=\{mATE,mASE,mAOE,mAVE,mAAE\}$，其中每个误差度量限制在 [0,1]。具体而言，$mATE$ 为平均位移误差、$mASE$ 为平均尺度误差、$mAOE$ 为平均方向误差、$mAVE$ 为平均速度误差和 $mAAE$ 为平均属性误差。

3.2 实验环境与训练策略

本文基于 Ubuntu 22.04 LTS 环境开展仿真验证实验，使用的软件栈主要包括 Python 3.8、PyTorch 1.10.0、

CUDA 11.3 与 cuDNN 8.6，并基于 MMDetection3D 1.3.0 实现模型训练与推理。硬件平台采用 Intel 至强 (Xeon) Platinum 8336C 处理器 (32 核，主频 2.3~3.4 GHz) 与 NVIDIA RTX 4090D×2，为大规模多模态数据处理与并行训练提供算力保障。

训练策略方面，选用 AdamW 优化器，初始学习率设为 2.5×10^{-5}，权重衰减 0.01，配合余弦退火策略在训练过程中自适应降低学习率，以提升收敛稳定性与泛化能力。整体训练历时 26 个 epoch，其中前 20 个 epoch 以 LiDAR 分支为主进行预训练，用于稳固几何表征；随后 6 个 epoch 进行多模态联合训练，在保持 LiDAR 表征稳定的基础上促进跨模态特征对齐与融合性能提升。为保证结果可复现，训练过程固定随机种子，并采用常规梯度稳定化配置。

网络与超参数设置如下：点云体素化分辨率为 (0.075 m, 0.075 m, 0.2 m)，处理空间范围设为 $x\in[-54, 54]$ m，$y\in[-54, 54]$ m，$z\in[-5, 3]$ m，覆盖主要可驾驶区域并兼顾远距目标。数据增强策略与 BEVFusion 基线保持一致，点云进行整体随机缩放、旋转、平移与点顺序打乱，以降低采样偏差并增强姿态鲁棒性；图像采用旋转与翻转等几何增强。所有几何变换均通过显式记录并传递对应的内外参与同构变换矩阵，确保经由 LSS 视图转换映射后的图像特征与点云 BEV 表征在空间上严格对齐。

3.3 损失函数

为在稀疏远距与类别不均衡条件下同时保证中心定位稳健性和几何回归精度，本文在 TransFusion 框架下构建以中心热力图和多分支回归为核心的检测损失 L。训练时，先将三维标注投影到 BEV 平面生成类别中心热力图 H，并把每个真值框编码为回归向量 $(\Delta x, \Delta y, \Delta z, w, l, h, \theta, v_x, v_y)$。随后在 LiDAR 坐标系下对预测查询与真值执行 Hungarian 一对一匹配，联合分类代价 (focal loss cost)、BEV 几何 L1 代价 (BBox BEV-L1 Cost) 与三维 IoU 代价 (IoU3D Cost) 确定正负样本。得到配对后，在 BEV 网格上以 Gaussian Focal Loss 计算中心热力图损失 $\lambda_{\rm hm}$，对全部查询以 Focal Loss 计算分类损失 $\lambda_{\rm cls}$，对匹配到的查询以加权 L1 计算回归损失 $\lambda_{\rm box}$ (对速度通道赋予较小权重以抑制低速噪声)；三项分别按有效样本数归一化，并以 $\lambda_{\rm hm}$、$\lambda_{\rm cls}$、$\lambda_{\rm box}$ 加权求和得到总损失 L。最终以 L 反向传播，使“中心命中”与“框体质量”(位置、尺度、朝向与速度) 协同优化，从而提升 mAP 与 NDS。

总体损失定义为

$$L = \lambda_{\rm hm}L_{\rm hm} + \lambda_{\rm cls}L_{\rm cls} + \lambda_{\rm box}L_{\rm box}\,, \tag{16}$$

式中：$L_{\rm hm}$ 是 BEV 中心热力图损失函数；$L_{\rm cls}$ 为查询分类损失函数，用于解码查询的类别判别；$L_{\rm box}$ 为几何与运动回归损失函数；$\lambda_{\rm hm}, \lambda_{\rm cls}, \lambda_{\rm box}$ 为三项损失的权重系数。并对各通道施加权重 $[1,1,1,1,1,1,1,1,0.2,0.2]$ ($\lambda_{\rm box}=0.25$)，以降低速度项的不确定性干扰。正负样本一一匹配由 Hungarian Assigner 3D 在 LiDAR 坐标系下完成，综合 Focal Loss Cost、BBox BEV-L1 Cost 与 IoU3D Cost 进行联合度量，从而在语义置信、BEV 平面几何与体积重叠度之间取得平衡，确保监督信号稳定且与几何一致。

3.4 对比试验

为系统评估所提方法的有效性，实验遵循 nuScenes 官方评价体系，报告总体指标与细粒度指标两部分结果。首先，在本地实现 BEVFusion 基线复现实验，其 NDS 与 mAP 分别为 69.1% 与 64.3%。在相同的本地硬件与训练配置下，本文方法 mAP 提升 2.8%、NDS 提升 1.7% (见表 1、表 2)。表明本文 MEPFusion 方法在检测精度与综合感知能力在同等算力预算下均有实质提升。其中“C.V.”、“T.L.”、“M.T.”、“Ped.”和“T.C.”分别代表施工车辆、拖车、摩托车、行人和交通锥。

表 2 展示各目标类别的检测精度，本文 MEPFusion 方法在小目标与复杂几何目标上取得稳定增益。小体积高语义类别方面，Pde.、T.C.、Barrier 与 M.T.分别提升 1.0%、5.7%、3.0% 与 6.5%，表明在多目标密集且局部遮挡并存的路口场景中，行人轮廓与交通锥边界的响应更集中、漏检明显减少。大型交通工具方面，虽 Trailer

表 1　nuScenes 测试集类别目标检测精度对比 (单位：%)

Table 1　Comparison of detection performance across categories on the nuScenes test set (unit: %)

Method	mAP	Car	Truck	C.V.	Bus	T.L.	Barrier	M.T.	Bike	Pde.	T.C.
BEVFusion	64.3	87.0	59.7	25.9	73.5	43.2	68.2	70.1	58.6	86.3	71.9
Ours	67.1	88.7	63.4	28.7	75.7	42.8	71.2	76.4	58.3	87.3	77.6
Gain	+2.8	+1.7	+3.7	+2.8	+2.2	−0.4	+3	+6.5	−0.3	+1.0	+5.7

表 2 nuScenes 测试集细粒度指标对比 (单位：%)

Table 2 Comparison of fine-grained evaluation metrics on the nuScenes test set (unit: %)

Method	NDS↑	mATE↓	mASE↓	mAOE↓	mAVE↓	mAAE↓
BEVFusion	69.1	28.4	25.5	31.7	26.5	18.6
Ours	70.8	28.1	25.1	30.3	25.3	18.3
Gain	+1.7	−0.3	−0.4	−1.4	−1.2	−0.3

略低于基线 (−0.4%)，但 Car、Truck 与 Bus 分别提升 1.7%、3.7% 与 2.2%，体现出在复杂交通流与视角变化下对大型车辆的更强稳健性，并显著抑制基线中的漏检与框漂移。上述增益源于多模态协同带来的对小尺寸、细长几何与远距稀疏回波上的综合改善。SimPVT 在金字塔多尺度建模基础上引入显著性重标定，提升小体积目标的语义判别与边界响应，使行人与交通锥在密集与局部遮挡场景下更易被凸显；SDMConv 依据点云密度实施位置加权，强化稀疏远距与细窄结构的几何连续性；GAFM 在 BEV 空间完成跨模态动态对齐与门控融合，在遮挡条带或远距回波弱时优先几何证据、在纹理混淆时提高语义权重，抑制框漂移与漏检，在大型交通工具上取得稳定收益。

三者协同形成功能分工，最终在小目标、远距与遮挡场景下同时提升检出与定位精度，基线存在的漏检与框漂移在本方法中明显缓解。

为更全面评估目标检测的回归质量，除 mAP 与 NDS 外，本文依据 nuScenes 官方评测协议进一步报告五类误差：平移误差 mATE (m)、尺度误差 mASE (1−*IoU*)、朝向误差 mAOE (rad)、速度误差 mAVE (m/s) 与属性误差 mAAE (1−*acc*)。与基线方法的细粒度指标对比见表 2：MEPFusion 在上述误差项上均呈现改善，其中对跨模态对齐更为敏感的朝向与速度误差下降更为显著，mAOE 与 mAVE 分别降低 1.4% 和 1.2%。这表明所提出的选择性融合与互导校正能够有效削弱错配与冗余干扰对回归分支的影响，从而提升融合表示的空间一致性与回归稳定性；此外，mATE、mASE 与 mAAE 的小幅下降亦说明定位、尺度与属性估计得到稳步优化。

同时，为系统检验 MEPFusion 在远距小目标场景下的检测增益，本文在 nuScenes 验证集上补充了基于距离分段的细粒度评测结果。表 3 将评测范围划分为 0~20 m (近距)、20~30 m (中距) 与 30~50 m (远距) 三个区间。结果表明，MEPFusion 在各距离段均取得一致提升，且在更具挑战的 30~50 m 远距段仍能获得稳定增益，mAP/NDS 分别提升 1.4% 和 0.8%。鉴于远距场景更易受到目标尺度缩小、遮挡累积以及点云回波稀疏等因素影响，基线性能在该区间显著退化，而 MEPFusion 仍能保持稳定改进，说明所提方法在远距离条件下具备更强的目标可检出性与综合感知质量。

此外，为定位增益来源并聚焦“远距小目标”这一关键情形，表 4 给出分距离段类别 AP，并选取 Pde.、M.T.作为典型小型障碍物代表，Car、Truck 作为大型目标代表进行对比。结果显示，在远距区间内，小型障碍物的增益更为显著，M.T.的 AP 由 26.2% 提升至 30.4% (+4.2%)，Pde.也有小幅提升；同时，大型目标亦保持稳定提升 (Car+1.2%，Truck+2.6%)。

上述结果表明，MEPFusion 的性能增益并非主要来源于近距易检出样本的累积贡献，而是在远距段面对更

表 3 nuScenes 验证集距离分段性能对比 (单位：%)

Table 3 Comparison of detection performance across distance ranges on the nuScenes validation set (unit: %)

Method	0-20 m		20-30 m		30-50 m	
	mAP	NDS	mAP	NDS	mAP	NDS
BEVFusion	77.8	77.2	61.0	67.4	30.4	46.0
Ours	78.6	78.0	62.6	68.4	31.8	46.8
Gain	+0.8	+0.8	+1.6	+1.0	+1.4	+0.8

表 4 nuScenes 验证集距离分段类别 AP 对比 (单位：%)

Table 4 Comparison of category-wise ap across distance ranges on the nuScenes validation set (unit: %)

Method	0-20 m				20-30 m				30-50 m			
	Pde.	M.T.	Car	Truck	Pde.	M.T.	Car	Truck	Pde.	M.T.	Car	Truck
BEVFusion	92.0	87.2	96.7	74.5	82.3	66.7	88.7	63.3	43.4	26.2	73.1	47.3
MEPFusion/Ours	92.4	89.9	96.7	79.3	83.5	68.8	88.8	63.5	43.7	30.4	74.3	49.9
Δ*AP*	+0.4	+2.7	+0.0	+4.8	+1.2	+2.1	+0.1	+0.2	+0.3	+4.2	+1.2	+2.6

具挑战性的小型障碍物时，仍能显著降低漏检并提升判别精度，验证了本文方法在远距小目标检测方面的有效性与稳健性。该优势源于多模态表征与对齐的协同：SimPVT 增强远距小目标语义与边界响应，SDMConv 提升稀疏回波下的几何连续性，GAFM 通过动态对齐与门控融合抑制远距错位，从而获得更稳健的跨模态互补增益。

进一步，将本文方法与不同技术路径的代表性三维目标检测方法进行对比评估。表 5 汇总了相机 (C)、激光雷达 (L) 与多模态 (C+L) 方法在 nuScenes 的公开结果 (均引自原论文或官方基准，含 Validation 与 Test)，用于量级参照。与相机组对比，本文 MEPFusion 方法在 Test 集达到 67.1% mAP 和 70.8% NDS，相对该组最优的 BEVDepth 分别领先 16.8% 和 10.8%；在 Validation 集达到 67.1% mAP 和 70.9% NDS，相对该组公开最高的 BEVFormer 分别领先 25.5% 和 19.2%。这表明在仅依赖视觉时常见的尺度与深度不确定、遮挡敏感等问题，经由跨模态补偿后得到根本性缓解。与激光雷达组对比，本文 MEPFusion 方法在 Test 集，MEPFusion 相比 TransFusion-L 仍有 1.6% mAP 和 0.6% NDS 的增益；在 Validation 集，相比 TransFusion-L，本文方法领先了 2.0% mAP 和 0.8% NDS。说明在保持高质量几何先验可达水准的同时，通过更稳健的语义、几何互证与对齐机制，MEPFusion 在方向与动态估计上进一步压缩误差。与多模态组对比，MEPFusion 在与 BEVFusion 保持相同算力与实现设置下，于 Test/Validation 分别实现+2.8%/+1.9% mAP 与+1.7%/+1.2% NDS 的提升，并在 Validation 取得全表最高 NDS (70.9%)。同时，相比 GraphAlign、LIFT、MVP 等近期方法，Test 集在 mAP 上分别领先+0.6%、+2.0%、+0.7%，NDS 亦小幅占优；在 Validation 集与 MVP 的 mAP 并列第一 (67.1%)，而 NDS 居首位 (70.9%)。上述结果说明，在不增加计算预算的前提下，跨模态对齐与信息互补被更有效地发挥。

综上，本文方法在精度、效率与鲁棒性之间实现更优平衡，体现出面向复杂交通场景的工程可用性与推广价值。在不增加计算预算的前提下，本文 MEPFusion 方法相较基线不仅在总体指标与关键类别、关键动态量上实现协同提升。并且，三项改进 SimPVT 强语义判别、SDMConv 稀疏几何增强与 GAFM 跨模态动态对齐分工明确且相互补偿，有效缓解小目标语义弱、稀疏导致的几何退化、跨模态对齐偏差的问题，使系统在遮挡、远距与复杂背景下保持更高的检测稳定性与工程可用性。

3.5 消融实验

为系统评估所提出的独立增益与协同效应，本文在 nuScenes 上开展控制变量的消融实验。为避免外部干扰，除被替换模块外，其余训练策略、数据增强、视图变换与检测头均与基线 BEVFusion 保持一致。实验结果如表 6 所示。基线在测试集上取得 64.3% mAP、69.1% NDS、6.8 f/s、39.80 M 参数，可以得出，SimPVT、

表 5 nuScenes 上三维目标检测性能对比结果 (单位：%)

Table 5 Comparison of 3D object detection performance on the nuScenes dataset (unit: %)

Method	Modality	Test		Validation	
		mAP	NDS	mAP	NDS
BEVDet	C	42.2	48.2	39.3	47.2
BEVFormer	C	48.1	56.9	41.6	51.7
BEVDepth	C	50.3	60.0	35.1	47.5
PointPillars	L	30.5	45.3	–	–
CenterPoint	L	58.0	65.5	50.3	60.2
TransFusion-L	L	65.5	70.2	65.1	70.1
PointPainting[27]	C+L	46.4	58.1	–	–
MVP	C+L	66.4	70.5	67.1	70.8
LIFT[28]	C+L	65.1	70.2	–	–
mmFusion[29]	C+L	–	–	65.4	69.8
GraphAlign[30]	C+L	66.5	70.6	62.8	68.5
BEVFusion	C+L	64.3	69.1	65.2	69.7
Ours	C+L	67.1	70.8	67.1	70.9

表 6 在 nuScenes 测试集上的消融实验对比结果

Table 6 Comparison of ablation study results on the nuScenes test set

Method	SimPVT	SDMConv	GAFM	mAP/%	NDS/%	FPS/%	Params/M
Bevfusion	–	–	–	64.3	69.1	6.8	39.80
E1	√	–	–	65.5	69.7	6.6	37.03
E2	–	√	√	64.6	69.5	6.9	39.08
E3	√	–	√	65.4	69.8	6.9	36.31
E4	√	√	–	66.2	70.0	7.0	37.03
E5/Ours	√	√	√	67.1	70.8	6.8	36.31

SDMConv 与 GAFM 在不同组合下均能稳定带来性能增益，并在三者联合时取得最优。

首先，从单模块出发，仅替换图像主干并引入无参注意力的 E1 (SimPVT) 即可将 mAP/NDS 提升至 65.5%/69.7%，提升 1.2% 和 0.6%，同时模型参数量下降至 37.03 M。需要强调的是，参数量的下降主要归因于 SimPVT 主干自身的轻量化特性，表明更高效的全局建模与显著性重标定能够在几乎不改变实时性水平的前提下有效释放视觉分支潜力；其次，转向双模块组合，可以观察到三条互补路径：其一，E2 (SDMConv+GAFM) 在不改动图像分支的前提下，将 mAP/NDS 提升至 64.6%/69.5%，并将参数降至 39.08 M，说明密度调制卷积对稀疏点云的几何刻画与门控对齐对跨模态稳定融合均具备正向作用；其二，E3 (SimPVT+GAFM) 的 mAP/NDS 达到 65.4%/69.8%，提升 1.1% 和 0.7%，参数进一步降至 36.31 M，显示语义增强与门控对齐之间存在互补；其三，E4 将 SimPVT 与 SDMConv 结合时，性能提升至 66.2%/70.0%，且推理速度提升至 7.0 FPS，同时参数维持 37.03 M，因而在“精度—速度—参数”三维度上取得较为均衡的改进。最后，在三模块联合的情形下，E5 在保持 6.8 f/s 实时性的同时，取得 67.1% mAP 与 70.8% NDS 的最佳结果，相对基线提升+2.8% 和+1.7%。参数降至 36.31 M，较基线减少约 8.8%。值得注意的是，E3 与 E5 的参数量 (Params) 保持一致，表明以 SDMConv 替换标准稀疏卷积不会引入额外的参数开销。如表 7 所示，SDMConv 作用于 BEV 点云几何表征特征图时，其平均计算量由 6.71×10^{-3} GFLOPs 增至 1.34×10^{-2} GFLOPs，绝对增量为 6.71×10^{-3} GFLOPs；对应推理时间仅由 3.5×10^{-2} ms 增至 3.8×10^{-2} ms，仅增加 3×10^{-3} ms。尽管 GFLOPs 与单模块推理延迟存在小幅上升，但其增量处于毫秒与 10^{-3}GFLOPs 量级，对整体实时推理链路的影响较为有限。更重要的是，SDMConv 通过引入空间密度调制机制，将传统等权聚合转化为对点云稀疏分布的自适应加权聚合，从而强化 BEV 空间的几何结构建模能力。综合来看，SDMConv 在不增加参数量的前提下，以可控的计算与时延代价换取更显著的表征与性能收益，有助于提升模型在复杂场景下的鲁棒性与检测精度。

表 7 模块计算开销与推理延迟统计

Table 7 Statistics of computational overhead and inference latency of different modules

Module	Params/M	GFLOPs	Inference time/ms
Conv	2.7×10^{-5}	6.71×10^{-3}	3.5×10^{-2}
SDMConv	2.7×10^{-5}	1.34×10^{-2}	3.8×10^{-2}

综上，SimPVT 主要强化视觉语义的全局一致性与关键区域显著性，SDMConv 面向稀疏与远距场景提升点云几何细节的可分性，GAFM 则在 BEV 空间实现跨模态对齐与互补交互；三者在效率与精度维度形成“可分—可对齐—可互补”的链式增益，使模型在参数规模更小、实时性不降的前提下，显著提高检测鲁棒性与上限。这表明本文的结构设计在保持实时性的同时具备更优的参数效率与检测性能。

3.6 可视化分析

为直观评估所提方法的有效性，图 5~图 8 给出典型场景下的实验结果可视化对比。本文方法和 BEVFusion 基线每组样例均按上下排列，左侧多视角图像与右侧 BEV 点云以同色虚线同步圈注关注区域，便于在纹理语义与几何回波之间进行一一对应的对比。这表明本文检测结果与 LiDAR 证据相互印证，性能提升来源于更稳健的跨模态对齐与几何表征增强，而非由视觉噪声诱发的语义误检。

首先，从城市白天主干道场景中观察到 (见图 5)，基线方法在远距小车与细长护栏处易出现漏检或边界残缺，而本文方法在相同圈注区域给出更完整的外接框并保持更稳定的朝向估计，在多车相互遮挡处，框体分离更清晰，边界与 BEV 点云的高密度回波贴合度更高。

图 5 城市白天主干道环境目标检测结果可视化

Fig. 5 Visualization of object detection results in an urban daytime arterial road environment

图 6 复杂交通环境目标检测结果可视化

Fig. 6 Visualization of object detection results in complex traffic environments

该现象主要得益于 SimPVT 对关键区域响应的强化以及视图变换阶段更稳定的深度分配，同时 SDMConv 对稀疏边界与细长结构的几何敏感性更强，从而提升了远距与细粒度目标的可检出性与定位一致性。

其次，在多目标密集、尺度跨度大与远距小目标并存的路口/干道场景下 (见图 6)，基线更易出现尺寸低估、姿态漂移与局部漏检等问题。本文方法在远距车辆与施工障碍物上维持连贯框体与正确朝向，圈注区域与点云轮廓高度一致；长边物体 (如拖车侧面、长护栏) 呈现更好的贴边性与姿态稳定，体现 SDMConv 在复杂遮挡与多目标干扰条件下对稀疏/细长几何结构的建模优势与 GAFM 的跨模态错位抑制能力。

此外，为验证所提算法在夜间低照度与阴雨天气等困难工况下的检测能力，除可视化对比外，本文进一步在 nuScenes 验证集上构建 Night 与 Rain 子集开展单独评测，并按官方评价协议统计 mAP、NDS 及误差分解指标以验证鲁棒性。其中，Night 子集包含 602 个样本，Rain 子集包含 1088 个样本。

在夜间低照度与强反射干扰条件下 (见图 7 与表 8)，强反光、车灯眩光与低照度易诱发基线高置信度误检与框体漂移。相比之下，本文方法在圈注处明显降低虚警，并在暗部区域保持对车辆、路障等关键目标连续检出，

图 7 夜间主干道环境目标检测结果可视化

Fig. 7 Visualization of object detection results in a nighttime arterial road environment

图 8 阴雨天环境目标检测结果可视化

Fig. 8 Visualization of object detection results in a rainy weather environment

同时框体在 BEV 中仍能稳定对齐稀疏但结构明确的几何回波。与此一致，表 8 显示本文方法在 Night 子集上 mAP/NDS 分别提升+2.0%/+0.8%，且多项误差指标整体下降，表明在低照度与强反射干扰下仍能提升检测稳定性与定位质量。其收益主要来自 SimPVT 的背景纹理抑制与 GAFM 的逐位置门控，使融合结果对噪声与反射更不敏感。

同时，在阴雨场景中 (见图 8 与表 9)，雨滴散射与路面反光会造成成像对比度下降与边缘纹理退化，并伴随点云有效回波减少，从而加剧框体抖动与局部漏检的风险。本文方法在圈注区域能够维持更稳定的框体尺度与位置一致性，并对受反光影响较大的目标类型表现更稳健。表 9 进一步表明，本文方法在 Rain 子集上 mAP/NDS 分别提升+2.3%/+0.8%，且速度/属性等误差项同步降低，说明其对降雨导致的感知退化与点云扰动具有更强的鲁棒性。

最后，综合图 5~图 8 的可视化对比与表 8~表 9 的子集定量结果，本文方法在小目标、远距稀疏、强遮挡以及夜间/阴雨强干扰等困难工况下，均较 BEVFusion 展现出更高的边界贴合度、更稳定的朝向与更低的假阳

表 8 nuScenes 验证集 Night 子集鲁棒性定量评测结果 (单位：%)

Table 8 Quantitative robustness evaluation results on the night subset of the nuScenes validation set (unit: %)

Method	mAP↑	NDS↑	mATE↓	mASE↓	mAOE↓	mAVE↓	mAAE↓
BEVFusion	37.8	43.6	48.9	45.1	43.1	57.3	61.1
Ours	39.8	44.4	47.6	44.5	42.8	57.0	60.7
Gain	+2.0	+0.8	−1.3	−0.6	−0.3	−0.3	−0.4

表 9 nuScenes 验证集 Rain 子集鲁棒性定量评测结果 (单位：%)

Table 9 Quantitative robustness evaluation results on the rain subset of the nuScenes validation Set (unit: %)

Method	mAP↑	NDS↑	mATE↓	mASE↓	mAOE↓	mAVE↓	mAAE↓
BEVFusion	64.7	70.0	29.5	26.9	33.1	20.0	13.8
Ours	67.0	70.8	29.2	26.8	33.1	19.6	12.7
Gain	+2.3	+0.8	−0.3	−0.1	0	−0.4	−1.1

性率，验证了所提方法在复杂环境下的稳健性与泛化能力。圈注区域在点云 BEV 上的几何证据与最终检测结果相互支撑，与定量结果 (mAP/NDS 提升) 保持一致，验证了 SimPVT 语义判别、SDMConv 几何建模与 GAFM 精细对齐与冗余抑制的协同作用。

4 结 论

针对多模态 3D 目标检测中图像语义信息受限、点云稀疏以及跨模态特征对齐不稳定等关键问题，本文提出的 MEPFusion 方法有效强化了多模态特征的表达与交互能力，显著提升了检测精度与系统鲁棒性。在 nuScenes 数据集上的实验验证了该框架的优越性，其 mAP 与 NDS 指标相较于 BEVFusion 基线分别提升了 2.8% 与 1.7%。进一步消融研究揭示，各模块在语义建模、几何结构增强与跨模态对齐方面具有独立增益，并通过协同融合进一步放大了整体性能。本文方法为多模态 3D 感知任务的深度融合与高效建模提供了可行思路。未来工作将重点探索 MEPFusion 在极端天气条件、超远距离小目标检测等复杂场景下的适应性与稳定性，并进一步研究面向边缘计算设备的轻量化部署方案，以推动高精度 3D 感知技术在实际自动驾驶与智能交通系统中的广泛应用。

作者贡献

李旭：实验设计与完成，论文的分析与写作；刘鸿恩、康婷：指导论文选题、论文修改；黄晟：研究资料的收集。

利益冲突

所有作者声明无利益冲突

参考文献

1. Li W L, Yu F, Shi X H, et al. Research on target detection algorithm for fusion of LiDAR and monocular camera under BEV features[J]. *Comput Eng Appl*, 2024, **60**(11): 182−193.
李文礼, 喻飞, 石晓辉, 等. BEV 特征下激光雷达和单目相机融合的目标检测算法研究[J]. 计算机工程与应用, 2024, **60**(11): 182−193.
2. Chen S, Ma Y, Qiao Y, et al. M-BEV: Masked bev perception for robust autonomous driving[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2024, 38(2): 1183-1191. https://doi.org/10.1609/aaai.v38i2.2788
3. Lang B, Li X, Chuah M C. BEV-TP: end-to-end visual perception and trajectory prediction for autonomous driving[J]. *IEEE Transactions on Intelligent Transportation Systems*, 2024, **25**(11): 18537−18546
4. Ge C J, Chen J S, Xie E Z, et al. MetaBEV: solving sensor failures for 3D detection and map segmentation[C]//*Proceedings of 2023 IEEE/CVF International Conference on Computer Vision*, 2023: 8687–8697. https://doi.org/10.1109/ICCV51070.2023.00801.
5. Wang W Y, Xu Z F, Qu C Y, et al. BEV space 3D object detection algorithm based on fusion of infrared camera and LiDAR[J]. *Acta Photonica Sin*, 2024, **53**(1): 0111002.
王五岳, 徐召飞, 曲春燕, 等. 基于红外与激光雷达融合的鸟瞰图空间三维目标检测算法[J]. 光子学报, 2024, **53**(1): 0111002.
6. Ji Y Z, Chen Y J, Yang L Q, et al. Spatial-channel attention multi-sensor fusion based on Bird's-Eye View[J]. *Pattern Recognit Artif Intell*, 2023, **36**(11): 1029−1040.
吉宇哲, 陈奕洁, 杨柳青, 等. 基于鸟瞰图的空间-通道注意力多传感器融合[J]. 模式识别与人工智能, 2023, **36**(11): 1029−1040.
7. Philion J, Fidler S. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3D[C]//*Proceedings of the 16th European Conference on Computer Vision*, 2020: 194–210. https://doi.org/10.1007/978-3-030-58568-6_12.
8. Huang J J, Huang G, Zhu Z, et al. BEVDet: high-performance multi-camera 3D object detection in bird-eye-view[Z]. arXiv: 2112.11790, 2022. https://doi.org/10.48550/arXiv.2112.11790.
9. Li Y H, Ge Z, Yu G Y, et al. BEVDepth: acquisition of reliable depth for multi-view 3D object detection[C]//*Proceedings of the 37th AAAI Conference on Artificial Intelligence*, 2023: 1477–1485. https://doi.org/10.1609/aaai.v37i2.25233.
10. Li Z Q, Wang W H, Li H Y, et al. BEVFormer: learning bird's-eye-

view representation from multi-camera images via spatiotemporal transformers[Z]. arXiv: 2203.17270, 2022. https://doi.org/10.48550/arXiv.2203.17270.

11. Chen X D, Han L L, Xiao Y H, et al. 3-D object detection for point clouds based on Voxel-Keypoint Feature Aggregation Network[J]. *Laser Technol*, 2026, **50**(2): 291−299.
陈旭东, 韩丽丽, 肖英豪, 薛保国, 马留洋. 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): 291−299.
12. Lang A H, Vora S, Caesar H, et al. PointPillars: fast encoders for object detection from point clouds[C]//*Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019: 12689–12697. https://doi.org/10.1109/CVPR.2019.01298.
13. Yan, Y, Mao Y X, Li B. SECOND: sparsely embedded convolutional detection[J]. *Sensors*, 2018, **18**(10): 3337.
14. Wang H, Tao L, Cai Y F, et al. CenterPoint-SE: a single-stage anchor-free 3-D object detection algorithm with spatial awareness enhancement[J]. *IEEE Trans Intell Transp Syst*, 2023, **24**(10): 10760−10773.
15. Cai H Y, Yang Z Q, Cui Z Y, et al. Image-guided and point cloud space-constrained method for detection and localization of abandoned objects on the road[J]. *Opto-Electron Eng*, 2024, **51**(3): 230317.
蔡怀宇, 杨朝乾, 崔子扬, 等. 图像引导和点云空间约束的公路洒落物检测定位方法[J]. 光电工程, 2024, **51**(3): 230317.
16. Bai X Y, Hu Z Y, Zhu X G, et al. Transfusion: robust LiDAR-camera fusion for 3D object detection with transformers[C]//*Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022: 1080–1089. https://doi.org/10.1109/CVPR52688.2022.00116.
17. Liu Z J, Tang H T, Amini A, et al. BEVFusion: multi-task multi-sensor fusion with unified bird's-eye view representation[C]//*Proceedings of 2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023: 2774–2781. https://doi.org/10.1109/ICRA48891.2023.10160968.
18. Yin T W, Zhou X Y, Krähenbühl P. Multimodal virtual point 3d detection[C]//*Proceedings of the 35th International Conference on Neural Information Processing Systems*, 2021: 1261.
19. Chen Z H, Li Z Y, Zhang S Q, et al. Deformable feature aggregation for dynamic multi-modal 3D object detection[C]//*Proceedings of the 17th European Conference on Computer Vision*, 2022: 628–644. https://doi.org/10.1007/978-3-031-20074-8_36.
20. Li Y W, Chen Y L, Qi X J, et al. Unifying voxel-based representation with transformer for 3D object detection[C]//*Proceedings of the 36th International Conference on Neural Information Processing Systems*, 2022: 1340.
21. Wang W H, Xie E Z, Li X, et al. PVT v2: improved baselines with pyramid vision transformer[J]. *Comput Vis Media*, 2022, **8**(3): 415−424.
22. Yang L X, Zhang R Y, Li L D, et al. SimAm: a simple, parameter-free attention module for convolutional neural networks[C]//*Proceedings of the 38th International Conference on Machine Learning*, 2021: 11863–11874.
23. Cammarasana S, Patanè G. Optimal weighted convolution for classification and denosing[Z]. arXiv: 2505.24558, 2025. https://doi.org/10.48550/arXiv.2505.24558.
24. Luan Q L, Chang X Y, Wu Y, et al. PAW-YOLOv7: algorithm for detection of tiny floating objects in river channels[J]. *Opto-Electron Eng*, 2024, **51**(4): 240025.
栾庆磊, 常昕昱, 吴叶, 等. PAW-YOLOv7: 河道微小漂浮物检测算法[J]. 光电工程, 2024, **51**(4): 240025.
25. Xia J Z, Sun H M, Hu S H, et al. 3D laser point cloud clustering method based on image information constraints[J]. *Opto-Electron Eng*, 2023, **50**(2): 220148.
夏金泽, 孙浩铭, 胡盛辉, 等. 基于图像信息约束的三维激光点云聚类方法[J]. 光电工程, 2023, **50**(2): 220148.
26. Zhang S W, Shao Y. Modified U-Net based ARSI object detection[J]. *Electron Opt Control*, 2025, **32**(9): 1−7.
张善文, 邵彧. 一种改进 U-Net 的 ARSI 目标检测方法[J]. 电光与控制, 2025, **32**(9): 1−7.
27. Vora S, Lang A H, Helou B, et al. PointPainting: sequential fusion for 3D object detection[C]//*Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020: 4603–4611. https://doi.org/10.1109/CVPR42600.2020.00466.
28. Zeng Y H, Zhang D, Wang C W, et al. LIFT: learning 4D LiDAR image fusion transformer for 3D object detection[C]//*Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022: 17151–17160. https://doi.org/10.1109/CVPR52688.2022.01666.
29. Cui L C, Li X X, Meng M, et al. MMFusion: a generalized multi-modal fusion detection framework[C]//*Proceedings of 2023 IEEE International Conference on Development and Learning (ICDL)*, 2023: 415–422. https://doi.org/10.1109/ICDL55364.2023.10364441.
30. Song Z Y, Wei H Y, Bai L, et al. GraphAlign: enhancing accurate feature alignment by graph matching for multi-modal 3D object detection[C]//*Proceedings of 2023 IEEE/CVF International Conference on Computer Vision*, 2023: 3335–3346. https://doi.org/10.1109/ICCV51070.2023.00311.

作者简介

刘鸿恩 (1987-)，男，博士，副教授，研究方向为智能交通系统融合感知、自学习特征建模和协同优化控制方法研究。

E-mail：len3610@126.com

【通信作者】李旭 (2001-)，男，硕士研究生，主要研究方向为机器学习、多模态感知等。

E-mail：885239570@qq.com

BEV object detection method with multi-layer perceptual enhancement and gated alignment fusion

Liu Hongen[1,2], Li Xu[1*], Huang Sheng[1], Kang Ting[3]

[1] Jiangxi Provincial Key Laboratory of Maglev Rail Transit Equipment, Ganzhou, Jiangxi 341000, China;
[2] Institute of Permanent-Magnet Maglev Technology and Rail Transit, Jiangxi University of Science and Technology, Ganzhou, Jiangxi 341000, China;
[3] School of Electrical Engineering and Automation, Jiangxi University of Science and Technology, Ganzhou, Jiangxi 341000, China

Abstract

Objective: Multimodal three-dimensional object detection combines camera semantics with LiDAR geometry and plays a central role in autonomous driving perception. Yet performance often degrades in complex traffic due to unstable semantics, sparse geometry, and cross-modal misalignment. Camera features become unreliable under glare, low illumination, motion blur, and cluttered backgrounds, introducing uncertainty into bird's-eye-view (BEV) projection. LiDAR point clouds are sparse at long range or in adverse weather, weakening local geometry and causing boundary blur and missed small objects. In BEV fusion, misalignment between modalities introduces conflicting cues and redundant responses, reducing confidence and accuracy. This study proposes an efficient BEV detector that enhances camera semantics, stabilizes sparse geometry, and achieves adaptive cross-modal fusion for improved accuracy and robustness without excessive computational cost.

Methods: The proposed framework, MEPFusion (multi-layer perception enhancement and gated alignment fusion), improves camera semantics, stabilizes sparse-geometry modeling, and enables gated alignment fusion in BEV space. For camera encoding, SimPVT enhances global context and salient-region representation with low overhead. A pyramid vision transformer backbone extracts multi-scale features to capture global and local dependencies. A parameter-free saliency reweighting unit amplifies target-related activations and suppresses background distraction, improving semantic separability. To reduce computation, the highest-resolution stage is removed while preserving three-scale outputs for projection. Features are lifted to BEV space via a lift–splat–shoot transformation supervised by depth estimation, stabilizing the depth distribution and reducing projection noise to improve spatial consistency. This camera branch aims to provide more reliable semantics under appearance variations, reducing texture-induced false positives and improving recall for visually small instances.

For LiDAR encoding, raw point clouds are voxelized and processed through a SECOND-style sparse convolution backbone. Standard sparse convolutions are replaced with spatial density-modulated convolution (SDMConv) to address equal-weight aggregation in sparse regions. SDMConv injects a position-dependent density prior into convolution responses via multiplicative modulation, producing a center-enhanced and periphery-decayed pattern within the receptive field. The modulation mitigates feature dilution and preserves boundary cues and shape continuity, especially in long-range or partially occluded areas, without increasing parameter count. The resulting sparse 3D features are compressed along height to produce LiDAR BEV features. In practice, SDMConv acts as a plug-in operator that strengthens local geometry when point evidence is weak, benefiting thin structures and small obstacles that are easily under-sampled.

Cross-modal fusion occurs in BEV space through a gated alignment fusion module (GAFM) designed to suppress misalignment while maintaining complementarity. Each modality undergoes channel recalibration to emphasize discriminative channels. A residual mutual-guidance path performs bidirectional correction: LiDAR features refine camera BEV features in geometry-sensitive regions, while camera semantics complement LiDAR features where geometric evidence is weak. Spatial gates are predicted for each BEV cell and normalized via softmax to assign adaptive fusion weights, allowing the more reliable modality to dominate locally while reducing inconsistent responses. The fused BEV features are decoded by a multi-scale neck and a transformer-based detection head to produce 3D bounding boxes and attributes. This strategy improves optimization stability and encourages reliability-aware weighting rather than uniform mixing.

Results and Discussions: Extensive experiments on the nuScenes benchmark demonstrate the effectiveness of MEPFusion. With six cameras and one LiDAR input, MEPFusion achieves 67.1% mAP and 70.8% NDS, surpassing the BEVFusion baseline by 2.8% and 1.7%,

Foundation item: National Natural Science Foundation of China (62463011), Sub-project of the National Key R&D Program of China (2023YFB4302104-2), and Jiangxi Provincial Natural Science Foundation, General Program (20224BAB202025)

*Correspondence: Li X, E-mail: 885239570@qq.com

respectively. Regression errors in translation, scale, orientation, velocity, and attribute are reduced, confirming improved spatial alignment and motion estimation. Category-level analysis shows notable gains for small or geometry-sensitive objects (e.g., traffic cones, barriers, motorcycles, and pedestrians) while maintaining improvements for major vehicle classes. These trends align with module contributions: SimPVT strengthens semantics in cluttered scenes, SDMConv alleviates sparsity-induced degradation, and GAFM reduces cross-modal interference through adaptive gating and complementary fusion. Qualitative inspection suggests fewer duplicated boxes and cleaner spatial responses in crowded regions, consistent with suppressing redundant activations caused by misalignment.

Distance-stratified evaluation further verifies consistent improvements across near, mid, and far ranges. The far range, where point sparsity and scale ambiguity are most severe, still achieves clear gains, highlighting the benefits of density-aware encoding and local gating. Robustness tests under adverse conditions also validate generalization. On the night subset, MEPFusion improves mAP and NDS while reducing error terms, indicating stable semantics and selective fusion under low illumination and glare. On the rain subset, performance again improves, confirming resilience to reduced contrast and weaker LiDAR returns. Ablation studies reveal both independent and synergistic effects: SimPVT alone enhances accuracy with fewer parameters; SDMConv and GAFM each bring additional gains; combining all three yields the best overall performance. Efficiency analysis shows SDMConv introduces minimal overhead, and the full model maintains comparable throughput with reduced total parameters, indicating a favorable accuracy–efficiency trade-off for on-vehicle inference.

Conclusions: In summary, MEPFusion integrates semantic enhancement, sparse-geometry reinforcement, and gated alignment fusion within a unified BEV framework. The method achieves higher accuracy, improved regression stability, and consistent robustness across distance and environmental variations while retaining deployment-friendly efficiency. Stable semantics before view transformation, density-modulated LiDAR encoding, and adaptive BEV fusion jointly mitigate misalignment and suppress redundant noise. MEPFusion thus offers an effective and reliable multimodal perception solution for urban autonomous driving.

Keywords: autonomous driving; multimodal fusion; BEV representation; multi-layer perception enhancement; gated alignment fusion

Liu H, Li X, Huang S, et al. BEV object detection method with multi-layer perceptual enhancement and gated alignment fusion[J]. *Opto-Electron Eng*, 2026, **53**(4): 250340; DOI: 10.12086/oee.2026.250340

Opto-Electronic Journals Group
www.oejournal.org

扫描二维码，获取 PDF 全文