基于空间重构和文本引导的室内开放域3D目标检测方法

樊怡麟[1],杨方正[1],臧银珠[1],李 辉[1*],刘治宇[2],孔祥振[3]
(1. 青岛科技大学数据科学学院,山东青岛 266061;2. 中国海洋大学信息科学与工程学部,山东青岛 266100;
3. 荷兰埃因霍芬理工大学工业工程学院,荷兰 5612 AE)

摘 要: 3D目标检测是提升智能系统在复杂室内环境下自主决策与精细化感知能力的关键。传统视觉方法受室内物体长尾分布影响显著,且缺乏文本语义的引导,导致模型难以对室内场景中不断涌现的未见类别进行有效泛化。针对室内复杂场景下目标类别繁杂且分布不均导致的多尺度语义混叠问题,以及跨模态异构空间中由于背景噪声引发的表征分布失配挑战,本文提出一种基于空间重构和文本引导的室内开放域3D目标检测方法。在视觉表征前端,本文设计层级边缘聚焦网络,利用分流注意力机制过滤通道冗余并增强跨尺度语义一致性,构建层级融合模块生成自适应层间注意力,从而有效解决由尺度变化引发的检测歧义。在多模态融合阶段,本文构建基于空间重构的跨模态特征融合模块,依据统计特性将异构特征空间解耦为信息流与冗余流,通过提纯高价值视觉特征缓解室内杂乱环境引起的信噪比差异。在解码阶段,本文建立语义增强的文本查询交互机制,将对比语言-图像预训练(Contrastive Language-Image Pre-training,CLIP)文本嵌入作为先验注入解码器,赋予查询向量目标导向的感知能力,降低细粒度类别的识别误差。在SUN RGB-D数据集上的实验结果表明,该方法在新颖类别上的检测精度达11.33%,全类别平均检测精度达19.30%,较基线方法分别提升了1.32%和1.15%,证明了该方法在室内开放域场景下具有优越的3D目标检测性能。

关键词: 3D目标检测;室内开放域;跨模态融合;空间重构;层级边缘聚焦;文本嵌入

基金项目: 国家重点研发计划(No.2023YFF0612100);山东省自然科学基金(No.ZR2024MF023);中国高校产学研创新基金“智能驾驶及智能座舱教育专项”(No.2024HT030)

中图分类号: TP391;TP183 **文献标识码:** A **文章编号:** 0372-2112(XXXX)XX-0001-12

电子学报URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20260414

Indoor Open Domain 3D Object Detection Method Based on Spatial Reconstruction and Text Guidance

FAN Yilin[1], YANG Fangzheng[1], ZANG Yinzhu[1], LI Hui[1*], LIU Zhiyu[2], KONG Xiangzhen[3]
(1. *School of Data Science, Qingdao University of Science and Technology, Qingdao, Shandong* 266061, *China*;
2. *Faculty of Information Science and Engineering, Ocean University of China, Qingdao, Shandong* 266100, *China*;
3. *Faculty of Industrial Engineering, Eindhoven University of Technology, Eindhoven* 5612 *AE, The Netherlands*)

Abstract: 3D object detection is crucial for enhancing the autonomous decision-making and refined perception capabilities of intelligent systems in complex indoor environments. Traditional visual methods are significantly affected by the long-tail distribution of indoor objects and lack textual semantic guidance, making it difficult for models to effectively generalize to the constantly emerging unseen categories within indoor scenes. To address the challenges of multi-scale semantic aliasing caused by the complex and uneven distribution of object categories in complex indoor scenes, and the representational distribution mismatch caused by background noise in cross-modal heterogeneous spaces, this paper proposes an indoor open-domain 3D object detection method based on spatial reconstruction and textual guidance. At the visual representation front end, this paper designs a hierarchical edge-focusing network, utilizing a shunt attention mechanism to filter channel redundancy and enhance cross-scale semantic consistency. This paper constructs a hierarchical fusion module to generate adaptive inter-layer attention, effectively resolving detection ambiguities caused by scale variations. In the multimodal fusion stage, this paper constructs a cross-modal feature fusion module based on spatial reconstruction. Based on statistical characteristics, the heterogeneous feature space is decoupled into information flow and redundant flow, and high-value visual features are purified to alleviate the signal-to-noise ratio differences caused by the

收稿日期:2026-07-01;录用日期:2026-07-13;责任编委:丛润民
*通信作者:李辉

cluttered indoor environment. In the decoding stage, this paper establishes a semantically enhanced text query interaction mechanism, which injects contrastive language-image pre-training (CLIP) text embeddings into the decoder as priors, endowing the query vectors with target-oriented perception capabilities and reducing recognition errors for fine-grained categories. Experimental results on the SUN RGB-D dataset show that the proposed method achieves a detection accuracy of 11.33% for novel categories and an average detection accuracy of 19.30% for all categories, representing improvements of 1.32% and 1.15% respectively compared to the baseline method. This demonstrates the superior 3D object detection performance of the proposed method in indoor open-domain scenes.

Keywords: 3D object detection; indoor open-domain; cross-modal fusion; spatial reconstruction; hierarchical edge-focusing; text embeddings

Foundation Item(s): National Key Research and Development Program (No.2023YFF0612100); Shandong Provincial Natural Science Foundation (No.ZR2024MF023); "Intelligent Driving and Intelligent Cabin Education Special Project" of China University Industry-University-Research Innovation Fund (No.2024HT030)

0 引言

在物联网与智慧家居生态日益成熟的趋势下,家庭服务机器人展现出巨大的应用潜能。为确保机器人在高度非结构化的室内环境中执行复杂任务,具备对场景内多类别目标的3D精确感知与语义理解能力是关键。室内3D目标检测技术[1-2]作为环境感知的底层支撑,赋予了系统对空间实体及动态属性的实时提取与解析能力,为提升室内场景的综合智能化水平奠定了基础。

传统的3D目标检测范式依赖预定义的闭集标签进行定位与分类,但面对室内环境极其显著的类别长尾分布以及不断涌现的未见实体时,其封闭性导致模型泛化能力严重受限。为突破这一瓶颈,旨在精准感知训练集外任意类别实体的开放词汇目标检测[3-5]迅速发展。近年来,伴随着大规模视觉-语言预训练[6]模型的驱动,二维开放词汇检测发展迅速。然而,由于3D几何数据固有的稀疏性及高维语义的缺失,直接将2D先验知识向3D空间平滑迁移面临着严峻挑战。当前主流的开放词汇3D检测方法多将对比语言-图像预训练(Contrastive Language-Image Pre-training,CLIP)等模型局限于末端分类器的定位,仅在网络输出阶段将提取的3D区域特征与文本嵌入进行简单的余弦相似度匹配。这种应用方式割裂了多模态信息的深度交互,使文本语义未能作为先验知识前置参与到特征提取与区域生成的关键环节中。

同时,除了高层文本语义的利用受限,室内场景固有的物理复杂性也对底层的视觉表征与跨模态对齐提出了严峻考验。室内空间中物体尺度跨度极大,传统视觉主干网络在进行特征降采样时,极易产生感受野内的多尺度语义混叠,导致模型在应对尺度剧烈变化时陷入检测歧义。此外,室内环境通常具有高度杂乱、遮挡密集的非结构化特点,这在引入多源异构数据时会裹挟海量的背景噪声。跨模态特征在进行空间对齐时,往往存在严重的信噪比失衡,导致高价值的目标特征被冗余的背景信号淹没。

针对上述挑战,本文提出一种基于空间重构和文本引导的室内开放域3D目标检测方法,以空间重构后的特征提纯与文本语义先验深度交互为核心主线,将二维图像的语义与语言大模型的先验知识注入到3D检测框架中,从而提升模型在复杂室内场景下对新类别的泛化与感知能力。为了验证所提方法的有效性,本文在大型室内场景数据集SUN RGB-D上进行了实验评估。实验结果表明,本文方法相比基线方法在各项评价指标上均有提升,证明了所提方法在复杂室内场景中的有效性和鲁棒性。本文的贡献总结如下:

(1)提出层级边缘聚焦网络用于视觉表征前端,该网络通过引入分流注意力机制与自适应层间注意力,增强跨尺度语义一致性,有效过滤视觉通道冗余并对齐多尺度特征,从特征提取端解决了室内场景中由目标尺度多变引发的语义混叠问题。

(2)提出基于空间重构的跨模态特征融合机制,利用数据的统计分布特性将异构特征空间解耦为纯净信息流与干扰冗余流,实现了跨模态交互中视觉特征的高效提纯,大幅缓解了室内杂乱背景引起的信噪比失衡难题。

(3)构建语义增强的文本查询交互机制,将CLIP文本嵌入作为高阶语义先验前置注入解码网络,打破了传统晚期融合的被动对齐局限,赋予查询向量目标导向的感知能力,显著降低细粒度类别的识别误差。

1 相关工作

1.1 跨模态目标检测

结合图像与点云数据的跨模态特征融合技术,能够实现两者优势互补,从而提取更具辨识度和细节丰富的目标特征,这种方法有效打破了单一模态在3D目标检测中的性能瓶颈,显著提升了整体的检测精

度。为打破单阶段融合局限,周治国等人[7]设计了包含点云色彩编码与候选框联合优化的多层检测网络,实现了特征深度交互。Xu等人[8]设计了PointFusion网络,将3D点视作空间锚点,依托全局与密集融合模块实现双模态特征的深度交互。Qi等人[9]以VoteNet为基础构建了ImVoteNet,通过将视觉特征附加至点云中来增强模型输入。为缓解多模态数据的语义差异,葛同澳等人[10]提出了体素与网格级双融合框架,并引入自适应鸟瞰图融合与Transformer编码器,有效增强了小目标感知与全局上下文提取能力。在特征拼接方面,Vora等人[11]利用传感器标定参数,实现了图像语义分割特征与3D点云特征的精准结合。为了降低3D检测的误检率,Zhang等人[12]基于不同数据特性,分别构建了点注意力与密集注意力融合机制。Tan等人[13]提出通过自适应注意力机制提取跨模态特征,并结合RoI池化融合模块强化局部信息。在优化计算效率上,Wang等人[14]创新性地将PointNet与对应二维区域的3D视锥空间相聚合,提取视锥特征用于边界框回归,大幅削减了计算开销。针对弱纹理目标,藏传方等人[15]引入表面法向量和球形邻域划分以增强几何表征,并在高维空间内自适应融合颜色、几何与法向信息。尽管上述跨模态融合方法在空间对齐与交互机制上取得了显著进展,但在复杂的室内场景下,二维图像往往伴随着复杂的背景信息,被动的交互方式易将二维背景噪声引入3D几何空间,导致真实目标的有效信号被冗余干扰淹没,引发严重的信噪比失衡。

1.2 开放域目标检测

得益于CLIP等视觉-语言预训练模型的突破,开放域目标检测技术迎来了飞跃式发展。Gu等人[16]提出的ViLD方法通过视觉-语言知识蒸馏,将预训练教师模型的分类知识迁移至两阶段检测器,通过编码候选框的图文特征并实现检测框嵌入与教师模型推断的语义对齐。Minderer等人[17]通过整合预训练模型与Vision Transformer架构,采用对比式文本-图像联合预训练获取跨模态表征,并实施端到端检测微调策略实现开放词汇目标检测。然而,这些视觉语言模型的训练主要聚焦于图像-文本层级的对齐,导致其缺乏区域-文本层级的精准匹配能力。针对缺乏文本语义的局限,樊琳等人[18]设计了多模态网络TMM-Net,利用图文交叉注意力机制,显著提升了诊断精度与可解释性。Cheng等人[19]开发的YOLO-World利用视语路径聚合网络及区域-文本对比损失,大幅加深了双模态信息的交互。Liu等人[20]的Grounding DINO框架整合了语言引导查询与跨模态注意力解码等模块,出色地完成了语言先验与视觉表征的语义对齐。面向更复杂的室内外环境,Wang等人[21]推出了OV-Uni3DETR多模态架构,借助2D与3D间的知识循环传播机制,保障了全场景、多模态下的高效检测。尽管开放域3D目标发展迅速,但现有的大多数开放域检测工作仍局限于将CLIP等模型仅作为末端分类器,使得检测网络在核心解码阶段缺乏高阶语义先验的有效引导。

2 本文方法

基于空间重构与文本引导的室内开放域3D目标检测算法框架如图1所示。该框架以图像-点云-文本多模态输入,主要由层级边缘聚焦网络、基于空间重构的跨模态特征融合模块,以及语义增强的文本-查询交互机制三部分组成。

2.1 层级边缘聚焦网络

在多模态协同感知的室内开放域3D目标检测任务中,图像特征的提取质量直接决定了跨模态对齐的上限。尽管常规的多尺度视觉表征范式能够构建基础的层级语义,但其主要依赖于粗放的跨层聚合与简单的逐元素交互,缺乏对通道维度的精细化筛选与语义校准,导致输出特征仍存在混叠效应与关键语义信息的模糊。为此,本文设计了层级边缘聚焦网络作为视觉特征重构的核心前端。该网络摒弃了传统的被动式特征交互,首先引入分流注意力融合模块,在单一特征层内构建主干与辅助双流路径以解决特征冗余问题。同时,构建层级融合模块生成自适应层间注意力,通过相邻层级的加权交互为多尺度特征空间提供精确的语义一致性约束,从而有效解决由尺度变化引发的检测歧义。

在具体实现上,该网络接收骨干编码器提取的初始多层级特征集合作为输入。首先,针对各层级的输入特征$\boldsymbol{X}$,利用分流注意力融合模块进行特征重构,其具体结构如图2所示。该模块首先将输入特征按通道维度平均划分为主干特征$\boldsymbol{X}_a$与辅助特征$\boldsymbol{X}_b$。这一设计的原理在于特征空间内部通常存在显著的通道冗余,通过在通道维度进行等分降维,能够实现特征流的初步解耦,进而通过后续的非对称网络结构对特征进行约束和引导。其中,主干分支旨在提取深层语义,通过连续的卷积层与ReLU激活提取特征,并引入通道注意力机制。具体地,利用全局平均池化提取全局上下文向量,进而结合双层卷积结构计算通道间的相关性权重,从而实现特征的特征增强;辅助分支则保持轻量化,以最大程度保留原始图像的底层几何结构与边缘细节。具体为

$$\boldsymbol{F}_a = F_{\text{trunk}}(\boldsymbol{X}_a) \odot \sigma\left(\boldsymbol{M}\left(\text{gap}\left(F_{\text{trunk}}(\boldsymbol{X}_a)\right)\right)\right) \tag{1}$$

其中,F_{trunk}代表主干提取网络,$\odot$表示逐元素相乘,σ为

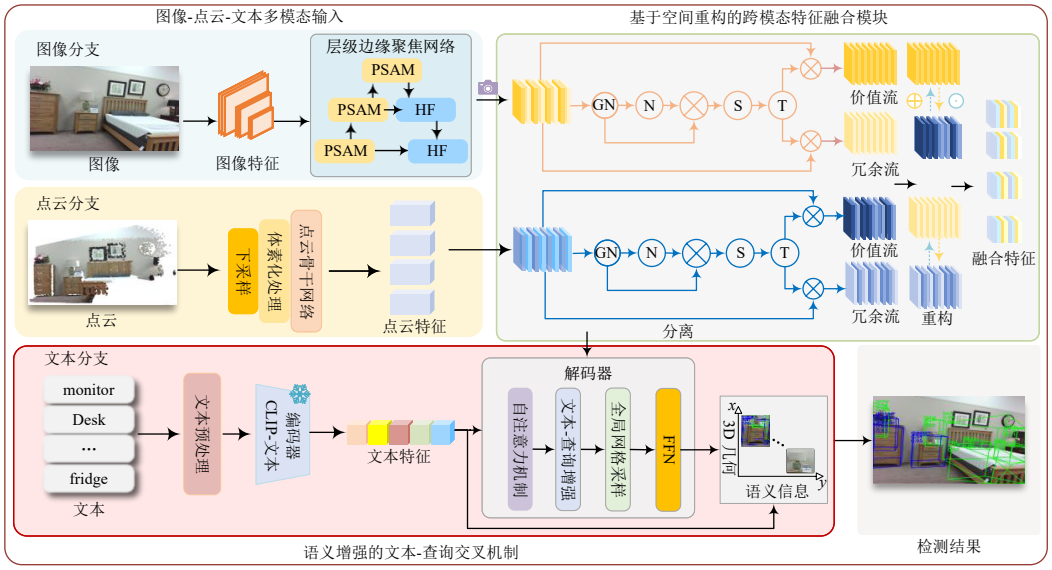

图1 提出方法的网络结构

Figure 1 Network architecture of the proposed method

Sigmoid激活函数，gap表示沿空间维度执行的全局平均池化操作，旨在聚合全局上下文信息，$\boldsymbol{M}$表示由双层卷积结构计算后生成的通道间的相关性权重矩阵。接着，为了充分利用不同频段的信息，PASM设计了双路交叉注入融合机制。如图2右侧所示，利用1×1卷积作为投影变换，分别将辅助分支$\boldsymbol{X}_b$映射至主干分支$\boldsymbol{F}_a$的通道空间，以及将主干分支$\boldsymbol{F}_a$映射至辅助分支$\boldsymbol{X}_b$的通道空间。两路特征分别进行拼接与3×3融合卷积，生成两个独立的融合路径输出。最后经过通道聚合与1×1卷积，得到当前层级的精炼特征。这一过程旨在利用轻量级的辅助分支保留原始几何细节，同时利用主干分支引入高层语义，实现特征的互补。具体公式如下：

$$\boldsymbol{Y}_1 = F_{\text{fuse1}}\left(\text{Concat}[\boldsymbol{F}_a, \phi_{b\to a}(\boldsymbol{X}_b)]\right) \tag{2}$$

$$\boldsymbol{Y}_2 = F_{\text{fuse2}}\left(\text{Concat}\left[\phi_{a\to b}(\boldsymbol{F}_a), \boldsymbol{X}_b\right]\right) \tag{3}$$

$$\text{Out}_{\text{pasm}} = \boldsymbol{F}_{\text{out}}\left(\text{Concat}[\boldsymbol{Y}_1, \boldsymbol{Y}_2]\right) \tag{4}$$

其中，F_{fuse}表示用于特征跨尺度聚合的融合卷积，Concat表示通道维度上的拼接操作，ϕ表示通道投影变换，$\boldsymbol{F}_{\text{out}}$为最终输出映射。

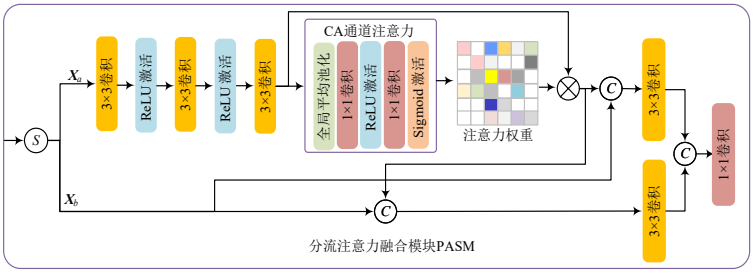

图2 分流注意力模块

Figure 2 Parallel attention shunt module

在完成层内特征重构后，仅依赖单一层级的感受野难以应对室内场景中物体尺度的剧烈变化。为此，本文引入层级融合模块作为相邻尺度的语义桥梁，其结构如图3所示。该模块接收来自PASM处理后的当前层特征$\boldsymbol{P}_l$与经过上采样的高层特征$\boldsymbol{P}_{l+1}$作为输入。HF模块采用双路并行的通道注意力机制，分别对两路输入特征进行全局上下文建模与权重计算。为了防止深层网络的梯度消失并保持特征的原始分布，在注意力加权后引入内部残差连接。加权并增强后的两路特征在通道维度进行拼接，并输入至3×3融合卷积层，用于学习生成跨尺度的语义描述，实现尺度空间的平滑过渡。具体公式如下：

$$\boldsymbol{U}_i^{\text{HF}} = \boldsymbol{U}_i \odot \text{CA}(\boldsymbol{U}_i) + \boldsymbol{U}_i \tag{5}$$

$$\boldsymbol{U}_{i+1}^{\text{HF}} = \boldsymbol{U}_{i+1}^{\text{UP}} \odot \text{CA}(\boldsymbol{U}_{i+1}^{\text{UP}}) + \boldsymbol{U}_{i+1}^{\text{UP}} \tag{6}$$

$$\boldsymbol{V}_i = F_{\text{fuse}}\left(\text{Concat}[\boldsymbol{U}_i^{\text{HF}}, \boldsymbol{U}_{i+1}^{\text{HF}}]\right) \tag{7}$$

其中，CA表示通道注意力操作，$\odot$表示特征图间的逐

元素相乘,$\boldsymbol{U}_i$和$\boldsymbol{U}_{i+1}^{\mathrm{UP}}$分别代表当前层级特征与经双线性插值上采样的高层级特征,$\boldsymbol{U}^{\mathrm{HF}}$代表经由注意力-残差结构优化后的特征表示。Concat表示沿通道维度的拼接操作,F_{fuse}表示用于特征跨尺度聚合的融合卷积,$\boldsymbol{V}_i$为该模块输出的最终精炼融合特征。为了实现特征增强的鲁棒性,本文最后设计了全局残差连接的自适应输出机制。将经过PASM和HF处理后的精炼特征$\boldsymbol{V}_i$与原始多尺度特征直接相加。该全局残差机制不仅实现了多阶段信息的有效回流,更充当了特征学习的自适应稳定器。

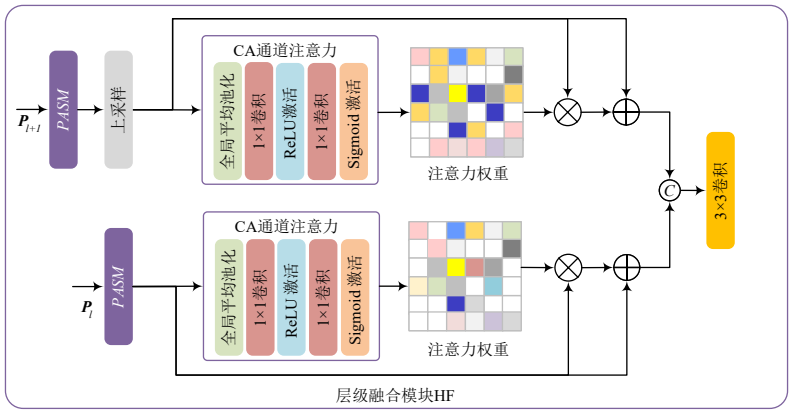

图3 层级融合模块

Figure 3 Hierarchical fusion module

2.2 基于空间重构的跨模态融合模块

在面向室内场景的3D开放域目标检测任务中,现有的特征融合方法往往忽略了深层特征中信噪比的空间差异性,简单地对全局特征进行逐元素相加或拼接。为此,本文提出基于空间重构的跨模态特征融合模块。该模块遵循分裂-变换-聚合的设计思想,首先基于通道响应自适应地将特征解耦为信息部分与冗余部分;然后,通过双流交互机制增强显著特征的跨模态一致性,并对冗余特征进行压缩提纯,从而在抑制噪声的同时保留上下文信息。基于空间重构的跨模态特征融合模块网络结构如图4所示。

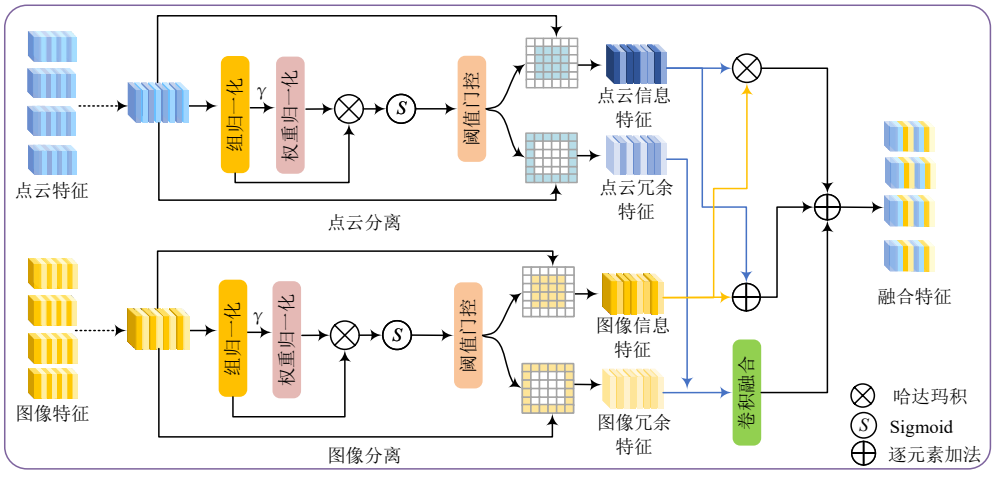

图4 基于空间重构的跨模态融合模块

Figure 4 Cross-modal fusion module based on spatial reconstruction

该模块的输入为体素化的点云特征$\boldsymbol{F}_{\mathrm{pts}}$与二维图像特征。为实现跨模态的三维空间对齐,本文基于相机内外参矩阵将非空体素中心投影至二维图像平面,并利用双线性插值进行特征采样,以此重构出与点云空间分辨率一致的体素化图像特征$\boldsymbol{F}_{\mathrm{imgs}}$。完成上述空间对齐后,该模块的具体处理流程如下:首先,为了准确评估特征空间中不同区域的重要性,本文利用组归一化的统计特性来生成空间注意力图。具体而言,考虑到GN层中的可学习仿射权重γ能够反映通道的激活程度,本文提取该权重并进行归一化

处理，随后将其与归一化后的特征图相乘。通过Sigmoid激活函数将特征响应映射到(0,1)区间，生成反映体素重要性的评分图$\boldsymbol{A}$，该过程旨在利用特征自身的统计分布来挖掘显著区域，具体权重归一化与评分生成公式为

$$\boldsymbol{w}_g=\frac{\gamma}{\sum\gamma+\xi} \tag{8}$$

$$\boldsymbol{A}=\sigma\left(\mathrm{GN}(\boldsymbol{F})\cdot\boldsymbol{w}_g\right) \tag{9}$$

其中，γ代表组归一化层中的可学习仿射权重，其数值大小直接反映了对应通道的特征显著性。ξ为一个极小的常数，用于保证数值计算的稳定性。$\boldsymbol{W}_g$表示经归一化处理后的通道显著性权。σ代表Sigmoid激活函数，负责将特征响应映射至(0,1)区间。

接着，针对室内开放域检测中背景噪声引发的语义匹配问题，本文基于评分图$\boldsymbol{A}$引入了显式特征分裂机制。通过引入阈值门控机制，设定阈值λ将特征空间二值化，生成信息掩码$\boldsymbol{M}$。利用该掩码将原始特征$\boldsymbol{F}$解耦为包含物体核心结构的信息特征$\boldsymbol{F}_{\text{value}}$和包含背景噪声的冗余特征$\boldsymbol{F}_{\text{red}}$。具体分裂操作如下：

$$\boldsymbol{M}=\begin{cases}1, & \boldsymbol{A}>\lambda\\ 0, & \boldsymbol{A}<\lambda\end{cases} \tag{10}$$

$$\boldsymbol{F}_{\text{value}}=\boldsymbol{F}\cdot\boldsymbol{M},\boldsymbol{F}_{\text{red}}=\boldsymbol{F}\cdot(1-\boldsymbol{M}) \tag{11}$$

其中，$\cdot$代表逐元素相乘。经过此步，点云与图像特征均被划分为显式划分为信息特征与冗余特征。

在传统的跨模态交互中，往往因缺乏对特征性质的区分而导致融合模糊。为解决该问题，本文对筛选出的信息特征$\boldsymbol{F}_{\text{value}}$执行双路增强交互。首先通过卷积层将图像特征映射至点云特征空间以消除维度差异，随后分别执行逐元素互乘与逐元素相加操作。互乘操作旨在挖掘几何与语义的共现一致性，仅当两模态在同一体素均有高响应时才激活，从而有效过滤由于单模态噪声引发的伪激活；相加操作则作为残差补充，确保在单模态信息缺失时特征的完整性。与此同时，对于被分离出的冗余特征$\boldsymbol{F}_{\text{red}}$本文并未直接丢弃，而是将其在通道维度拼接后输入至轻量级瓶颈层。通过1×1×1卷积对冗余特征进行通道压缩与提纯，具体公式如下：

$$\boldsymbol{F}_{\text{refine}}=\mathcal{H}_{1\times1\times1}\left(\mathrm{Concat}(\boldsymbol{F}_{\text{red}}^{\text{pts}},\boldsymbol{F}_{\text{red}}^{\text{img}})\right) \tag{12}$$

该设计在降低计算开销的同时，有效保留了有助于场景理解的全局上下文信息。最终，将交互增强后的信息特征与提纯后的背景特征进行多流聚合，通过特征重构层输出最终的融合特征。

2.3 文本-查询交互机制

在室内3D场景中，视觉特征的模糊性及局部纹理相似性常导致严重的类别混淆。尽管CLIP等预训练视觉语言模型在多模态对齐任务中表现优异，但在面对视觉特征模糊及局部纹理高度相似的场景时，其特征判别力显著下降。这一局限性在处理室内场景中外观高度近似的细粒度物体时尤为突出，仅在网络输出端依赖特征对齐，难以纠正前序特征提取阶段的语义偏差，从而严重制约了室内开放域检测的精度表现。

现有的Transformer检测器多依赖随机初始化的查询点与视觉特征进行交互。由于缺乏显式语义先验，这种盲目的几何探索极易受背景噪声干扰，难以在复杂场景中精准提取目标特征。虽然利用文本直接增强视觉特征已成为开放域感知的常规范式，但在三维目标检测中，若将高维密集的文本语义强行注入稀疏的三维特征空间，极易淹没点云脆弱的局部几何拓扑结构，引发严重的特征级几何退化。为此，本文提出语义增强的文本-查询交互机制，如图5所示。该机制突破了将CLIP仅作为末端分类器的传统范式，同时在解码器层级引入文本模态作为查询向量的语义先验，利用预训练视觉语言模型开放词汇的语义表征，在查询点与视觉特征交互之前，先对其进行类别特征的软注入，从而赋予查询点明确的寻找目标。

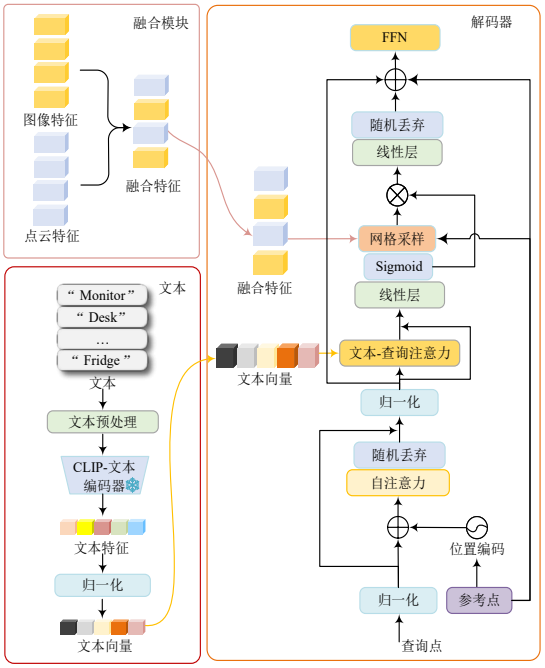

图5 文本查询交互机制

Figure 5 Text query interaction mechanism

具体而言，首先构建涵盖所有待检测类别的文本提示集合，并经由冻结参数的CLIP文本编码器提取原始的高维语义特征。鉴于检测器解码器的嵌入维度D与CLIP输出维度不一致，为了实现特征空间的

对齐,本文引入了一个可学习的线性映射层,将 $\boldsymbol{F}_{\text{text}}$ 投影至 D 维空间,并进行归一化处理,最终形成文本向量 $\boldsymbol{P}\in\mathbb{R}^{K\times D}$。

本文在解码器前向传播阶段,并未采用可能导致查询点空间分布退化的全量交叉注意力机制,而是设计了一种基于门控残差的文本注入模块。不同于常规检测器中查询向量直接与视觉特征交互的范式,可学习查询在经过自注意力层建模实例间的依赖关系后,引入文本-查询注意力模块。在该模块中,可学习查询作为 Query,而处理后的文本向量作为 Key 和 Value,通过多头注意力机制捕捉查询点与特定类别语义之间的亲和度。为了防止强烈的文本语义特征破坏查询点原有的几何位置信息,本文引入了一个可学习的门控系数 φ 来动态调节语义注入的强度。文本引导后的语义增强查询 $\Delta\boldsymbol{Q}$ 计算过程如下:

$$\Delta\boldsymbol{Q}=\text{MHAttn}(\text{LN}(\boldsymbol{Q}),\text{LN}(\boldsymbol{T}),\text{LN}(\boldsymbol{T})) \tag{13}$$

$$\boldsymbol{Q}_{\text{sem}}=\boldsymbol{Q}+\gamma\cdot\Delta\boldsymbol{Q} \tag{14}$$

其中,MHAttn 代表多头注意力,$\boldsymbol{Q}$ 表示经过自注意力更新后的原始物体查询,$\boldsymbol{T}$ 代表 CLIP 生成的文本原型特征,LN 表示层归一化操作,$\Delta\boldsymbol{Q}$ 为计算得到的语义增量。φ 初始化为一个较小的数值,允许模型在训练初期保留几何探索能力,随训练进行自适应地引入语义先验。

经过上述文本语义注入后,语义增强查询 $\boldsymbol{Q}_{\text{sem}}$ 被赋予了针对特定类别的感知倾向,并携带了经过修正的几何参考点信息。随后,语义增强的查询向量被输入至网络采样模块,旨在以明确的类别语义为导向,与融合模块输出的多模态特征进行深度交互。为了实现特征的自适应选择,本文利用线性投影层从 $\boldsymbol{Q}_{\text{sem}}$ 中解码出针对采样点的注意力权重 $\boldsymbol{W}_T$。随后,依据 $\boldsymbol{P}_{\text{ref}}$ 在多模态特征图 $\boldsymbol{V}$ 上执行双线性插值采样。利用语义增强后的参考点进行定点采样,并利用前述生成的语义注意力权重对采样后的视觉特征进行加权聚合。最后,通过输出投影层与残差连接更新查询向量。具体公式为

$$\boldsymbol{W}_T=\sigma(\boldsymbol{W}_{\text{attn}}\cdot\boldsymbol{Q}_{\text{sem}}) \tag{15}$$

$$\boldsymbol{F}_{\text{sample}}=\text{GridSample}\left(\boldsymbol{V},\boldsymbol{\Phi}(\boldsymbol{P}_{\text{ref}})\right) \tag{16}$$

$$\boldsymbol{F}_{\text{agg}}=\sum_{k=1}^{K}\boldsymbol{W}_T\odot\boldsymbol{F}_{\text{sample},k} \tag{17}$$

$$\boldsymbol{Q}_{\text{out}}=\text{Dropout}(\boldsymbol{W}_{\text{out}}\cdot\boldsymbol{F}_{\text{agg}})+\boldsymbol{Q}+\text{PE}(\boldsymbol{P}_{\text{ref}}) \tag{18}$$

其中,σ 表示 Sigmoid 激活函数,$\boldsymbol{W}_{\text{attn}}$ 为注意力权重生成层的可学习参数,$\boldsymbol{V}$ 代表多模态融合特征图,$\boldsymbol{\Phi}$ 表示将归一化参考点 $\boldsymbol{P}_{\text{ref}}$ 投影至体素坐标系的变换操作,GridSample 代表双线性插值网格采样,$\boldsymbol{F}_{\text{sample}}$ 代表依据参考点采样得到的视觉特征,$\boldsymbol{W}_T$ 为解码出的针对采样点的注意力权重,$\boldsymbol{F}_{\text{agg}}$ 表示经过 $\boldsymbol{W}_T$ 加权聚合后的特征;K 表示采样点数,$\boldsymbol{W}_{\text{out}}$ 为输出投影矩阵,PE(·)为位置编码生成器,$\boldsymbol{Q}$ 为进入该模块前的原始输入。最终,在分类阶段,本节将解码器输出的物体特征与 CLIP 文本原型进行余弦相似度计算。

3 实验结果及分析

为验证所提室内 3D 目标检测方法的有效性,实验在 SUN RGB-D 标准数据集上展开。本文采取了长尾分布下的任务设定,保持将训练样本数量最多的前 10 个类别被视为已知类别,其余 36 个类别作为新颖类别;本文以图像、点云双模态作为输入,实验中使用 AP 和 AR 作为评估指标,其中分为所有类别、基础类别、新颖类别三部分。关于评估指标,采用 IoU 阈值为 0.25 时的 AP 和 AR 作为验证集性能指标。

3.1 实验环境设置

如表 1 所示,本文使用 GPU 为 NVIDIA RTX 4090 的服务器进行训练和测试。

表 1 实验配置

Table 1 Experiment configuration

操作系统	Ubuntu 20.04.4 LTS
CPU	Intel(R) Xeon(R) Gold 6430
GPU	NVIDIA GeForce RTX 4090
Python	3.8.10
Pytorch	1.11.0+cu113
Torchvision	0.12.0+cu113
CUDA	11.3
CUDNN	8.2.0

3.2 实验实施细节

本文基于 MMDetection3D 框架实现所提出的 3D 目标检测模型,所提方法以 OV-Uni3DETR 作为基线网络。在网络架构方面,图像分支利用 ResNet50 结合特征金字塔网络提取多尺度视觉特征;点云分支则沿用 SparseEncoderHD 进行体素化特征编码,体素尺寸设定为 0.02 m。针对开放词汇检测任务,本文集成利用 CLIP 文本嵌入向量初始化检测头。在训练策略上,本文选用 AdamW 优化器,基础学习率设为 2×10^{-4},权重衰减为 0.01。为保护预训练特征的稳定性,对图像骨干网络、点云编码器等预训练分支设置了 0.1 的学习率缩减系数。整体训练周期设定为 40 个 Epoch。同时,由于基线方法 OV-Uni3DETR 官方并未开源其带有 DCNv2 模块的专属预训练权重以及特定的 CLIP 类别嵌入文件,本文在进行对比实验时,采用通用的 ResNet50 预训练权重替代进行主干网络初始化,并依托开源 CLIP 模型自主生成文本嵌入向量以完成基线复现。

3.3 消融实验

为了验证所提方法中各核心模块的有效性，通过逐步嵌入层级边缘聚焦网络（HEFN）、跨模态特征融合模块（SR-CFFM），以及语义增强的文本-查询交互机制（TQI），表2是各个模块对检测的实验结果。实验分析表明，随着各模块的协同加入，模型性能呈现出稳步上升的态势，最终达到了最优的检测精度。

表2 各模块有效性实验

Table 2 Effectiveness experiment of each module

基线	HEFN	SR-CFFM	TQI	AP_{novel}	AP_{base}	$AP_{average}$	AR_{novel}	AR_{base}	$AR_{avergae}$	推理时延 / ms
√				10.01	47.47	18.15	47.46	81.66	54.89	348.15
√	√			10.25	47.58	18.36	47.92	81.70	55.26	360.65
√		√		10.64	47.63	18.68	48.21	81.78	55.51	356.25
√			√	10.30	47.81	18.45	48.17	80.86	55.27	380.39
√		√	√	10.91	47.89	18.95	**48.86**	**80.98**	**55.85**	**388.49**
√	√	√	√	**11.33**	**48.01**	**19.30**	48.78	80.74	55.73	400.99

实验结果表明，相比于基线方法，层级边缘聚焦网络（HEFN）的加入有效解决了室内物体尺度差异剧烈导致的语义偏差。实验证明，该模块通过分流注意力与自适应层间加权，对图像分支的通道冗余进行了深度过滤，确保了在不同分辨率下边缘信息的锐度与语义的一致性。与SR-CFFM协同工作时，其对复杂遮挡目标的分类精度提升尤为明显。

单独引入基于空间重构的跨模态特征融合模块（SR-CFFM）后，AP_{novel}提升了0.63%，同时在全类别召回率上展现出稳步增益。这表明该模块通过特征解耦机制，成功将异构模态中的信息流与冗余流进行了显式分裂，在抑制室内杂乱背景噪声的同时，大幅提纯了高价值的视觉表征，为后续的精确对齐奠定了高质量的特征基础。语义增强的文本-查询交互机制（TQI）对新颖类别指标AP_{novel}的贡献最为显著。当SR-CFFM与TQI联合使用时，AP_{novel}从基线的10.01%跃升至10.91%。这验证了将CLIP文本嵌入作为先验软注入解码器的优越性，这种目标导向的感知方式缩短了视觉表征与语义锚点之间的映射距离，使得查询向量在交互阶段即具备了对细粒度新物体的判别先验，有效缓解了室内场景中常见的类别混淆现象。

当同时引入三个模块时，模型在指标上均取得最优结果。实验结果充分证明了本文设计的科学性：HEFN负责尺度校准，SR-CFFM负责特征筛选，而TQI负责语义制导。三大模块在特征空间、尺度空间与语义空间形成了闭环协同，共同突破了现有室内开放域3D检测在异构特征聚合与跨模态对齐方面的性能瓶颈。

同时，单独加入HEFN、SR-CFFM和TQI后模型的推理时延仅分别增加12.50 ms、8.10 ms和32.24 ms，完整方法的单帧推理时延控制在400.99 ms，约2.5 FPS，总参数量仅85.25 M，推理峰值显存低至0.79 GB。证明了本文方法在显著提升开放域感知效能的同时，兼顾了高实时性低硬件开销。

（1）层级边缘聚焦网络有效性

为了深度剖析HEFN对多尺度特征建模的贡献，针对其核心组件分流注意力融合模块（PASM）与层级融合模块（HF）开展了消融实验。实验结果如表3，在仅引入分流注意力融合模块（PASM）后，$AP_{average}$提升至18.28%。这证明了PASM通过主干与辅助分支的双流交互机制，能够有效抑制FPN输出特征在通道维度的冗余。通过分流注意力机制，模型能够自动强化具备显著语义边缘的通道响应，从而为后续的3D融合提供了更具判别力的视觉表征。单独引入层级融合模块（HF）同样带来了性能增益。HF模块通过自适应层间加权，建模了相邻特征层级间的语义关联。这有效解决了室内场景中由于物体尺寸剧烈变化导致的尺度感知歧义。实验数据表明，HF模块能够显著提升模型对多尺度目标的捕获能力，减少了因语义混叠导致的定位偏差。

表3 层级边缘聚焦网络消融实验

Table 3 Ablation study of the hierarchical edge-focusing network

方法	AP_{novel}	AP_{base}	$AP_{average}$
基线方法	10.01	47.47	18.15
+PASM	10.16	47.52	18.28
+HF	10.12	47.50	18.25
+HEFN	**10.25**	**47.58**	**18.39**

（2）基于空间重构的跨模态特征融合有效性

基线方法仅通过逐元素相加融合映射后的3D图像特征与点云特征，虽然3D图像特征提供了前景点的图像语义特征，但融合方式粗糙，没有充分融合不同模态的信息。为此，本文在基线基础上增加HEFN模块的消融实验，结果如表4。首先，引入基于简单注意力的融合策略，通过为不同模态生成体素级的注

意力权重图,实现特征的自适应加权融合。实验数据显示,该策略使模型平均精度提升了0.15%,证明了利用注意力权重挖掘显著区域能初步缓解杂乱背景对语义对齐的影响。然后,在仅保留信息流的实验中,模型仅利用特征分离核心结构特征并执行双路增强交互,而完全舍弃解耦出的冗余部分。实验显示模型通过双路增强交互机制,实现了3D点云的几何结构与二维图像语义的深度耦合,使$AP_{average}$显著提升至18.52%。证明了逐元素互乘操作通过挖掘空间几何结构与视觉语义的共现一致性,有效地抑制了由单一模态干扰引起的伪激活响应;而残差相加操作则发挥了模态补偿作用,确保在点云数据稀疏或几何结构破碎的区域,图像语义先验仍能为目标识别提供稳健的特征引导。最终,在引入冗余特征提纯机制后,模型相较于仅保留信息特征获得了0.16%的精度提升。这证明了通过对冗余分量进行通道压缩,能有效提取出被过滤的全局上下文信息,从而在抑制噪声的同时兼顾了场景语义的连贯性。

表4 基于空间重构的跨模态特征融合模块消融实验

Table 4 Ablation study of the cross-modal feature fusion module based on spatial reconstruction

融合策略	AP_{novel}	AP_{base}	$AP_{average}$	AR_{novel}	AR_{base}	$AR_{average}$
基线方法	10.01	47.47	18.15	47.46	81.66	54.89
基于简单注意力	10.17	47.59	18.30	47.75	81.58	55.10
仅保留信息特征	10.49	47.54	18.52	48.16	81.74	55.48
完整空间重构	**10.64**	**47.63**	**18.68**	48.21	81.78	55.51

在基于空间重构的跨模态特征融合模块中,特征分裂阈值λ是决定信息特征与冗余特征去耦效果的关键超参数。若λ过小,模块难以有效过滤背景噪声;若λ过大,则可能导致核心几何结构的丢失。为此,本文针对不同阈值设定开展了敏感性实验,结果如表5所示。

从表5的实验数据可以观察到,模型精度与召回率对阈值λ的变化呈现出非单调波动的趋势。当阈值$\lambda=0.5$时,模型在平均精度指标上达到最好,且在新颖类别上的表现最为突出,这表明中等强度的过滤策略能够最有效地平衡噪声抑制与特征保留。尽管召回率指标$\lambda=0.5$在左右达到全局最优,展现出更强的目标捕捉能力,但随着阈值进一步升高至0.7,过为剧烈的特征分裂开始破坏基础类别的稳定性,导致综合精度出现下滑。实验结果有力验证了SR-CFFM模块通过显式重构特征空间,能够显著改善模型在处理未知物体时的语义一致性。综上所述,为了兼顾检测精度与召回率的综合表现,本文在后续所有实验中将空间重构阈值设定为0.5。

表5 参数敏感性分析

Table 5 Parameter sensitivity analysis

阈值λ	AP_{novel}	AP_{base}	$AP_{average}$	AR_{novel}	AR_{base}	$AR_{avergae}$
0.3	10.44	47.50	18.49	47.90	81.45	55.19
0.4	10.18	47.44	18.28	48.55	82.04	55.83
0.5	**10.64**	**47.63**	**18.68**	48.21	81.78	55.51
0.6	10.54	47.44	18.56	**48.81**	81.83	**55.99**
0.7	10.48	47.33	18.49	48.26	**82.20**	55.64

(3)文本-查询交互机制有效性

为了验证提出的文本-查询交互机制中各关键组件的有效性,本文在基础多模态Transformer检测器上进行了逐步消融实验。具体实验结果如表6所示。首先,仅在网络末端使用CLIP进行分类对齐,缺乏提取阶段的语义引导,其在新颖类别上的检测精度仅为10.01。本文尝试将CLIP文本特征直接与查询向量拼接,结果发现虽然新颖类精度略有上升,但基础类别精度下降至47.24。这表明强硬的文本特征注入破坏了查询点原有的几何空间分布退化。然后在完整模块中,引入了基于门控残差的文本注入模块。通过可学习的门控系数自适应调节语义注入强度,模型不仅恢复了基础类的精度,还将AP_{novel}提升至10.30,证明了门控机制在保留几何探索能力与引入语义先验之间的平衡,实现了在密集遮挡与细粒度相似场景下对目标特征判别力的显著增强。

表6 文本-查询交互机制消融实验

Table 6 Ablation study of the text-query interaction mechanism

方法	AP_{novel}	AP_{base}	$AP_{average}$	AR_{novel}	AR_{base}	$AR_{avergae}$
无语义引导	10.01	47.47	18.15	47.46	81.66	54.89
直接拼接	10.24	47.24	18.28	47.65	80.62	54.92
完整模块	**10.30**	**47.81**	**18.45**	**48.17**	**80.86**	**55.27**

3.4 对比实验

为证实本文所提方法的先进性,本文将其与基于3DETR生成伪检测框的3D-CLIP、Det-CLIP2、Det-PointCLIP、Det-PointCLIPv2以及当前最前沿的室内开放域检测框架INHA、CoDAv2进行对比。

如表7所示,引入图像信息的双模态方法整体优于仅以点云为输入的单模态方法。本文通过空间重构模块有效缓解多模态融合中的信息冗余,在基础类别AP上显著超过单模态的CoDA和CoDAv2,表明精确的特征对齐与提纯使图像模态能够为点云提供关键的语义补充,从而突破单模态感知瓶颈。与同样采用双模态的GLRD相比,本文方法虽在AP指标上略逊,但推理时延显著更低,同时取得了更高的AR指标,召回性能更优。此外,相较于单模态方法,本文仅以小幅增加的推理开销便实现了显著的精度

提升，单帧推理时延控制在400.99 ms。总体而言，本文方法通过端到端的轻量化感知架构，在保持高效推理的同时，依然取得了极具竞争力的综合感知性能，在精度与效率之间实现了良好的平衡。

表7 与其他先进方法对比实验

Table 7 Comparison experiments with other advanced methods

方法	输入	AP_{novel}	AP_{base}	$AP_{average}$	AR_{novel}	AR_{base}	$AR_{average}$	推理时延/ms
Det-PointCLIP[22]	点云	0.09	5.04	1.17	21.98	65.03	31.33	262.4
Det-PointCLIPv2[23]	点云	0.12	4.82	1.14	21.33	63.74	30.55	258.7
Det-CLIP2[24]	点云	0.88	22.74	5.63	22.21	65.04	31.52	271.6
3D-CLIP[6]	双模态	3.61	30.56	9.47	21.47	63.74	30.66	324.2
OV-Uni3DETR[21]	图像	2.06	16.50	5.41	27.79	62.92	35.96	301.5
CoDA[25]	点云	6.31	36.21	12.81	35.65	64.80	41.98	286.3
INHA[26]	点云	8.91	42.17	16.18	**51.34**	78.65	**57.23**	312.8
CoDAv2[27]	点云	9.17	42.04	16.31	43.16	71.64	49.35	307.5
OV-Uni3DETR[21]	双模态	10.01	47.47	18.15	47.45	**81.66**	54.89	348.15
GLRD[28]	双模态	12.96	49.40	20.88	47.74	80.93	54.96	786.2
本文方法	双模态	11.33	48.01	19.30	48.78	80.74	55.73	400.99

3.5 实验结果可视化及分析

为了更直观地验证本文所提算法在复杂室内环境下的检测效能，本节选取了SUN RGB-D数据集中6个典型复杂场景的检测结果进行可视化展示，并与基线网络进行了对比分析。图中红色检测框代表模型训练中见过的基础类别，蓝色检测框代表未见过的新颖类别。

如图6所示，在目标分布密集且相互遮挡严重的区域，基线网络因缺乏有效的语义引导，对部分被遮挡的目标实例产生了明显漏检。相比之下，本文方法通过语义增强的文本-查询交互机制引入了文本语义先验，利用预训练语言模型赋予查询点针对特定类别的目标导向感知能力，从而成功召回了更多被遮挡的基础类实例以及基线完全无法识别的新颖类物体，例如场检测覆盖率。从整体可视化效果来看，本文方法生成的3D检测框在朝向角和包围框贴合度上普遍优于基线网络，这主要归功于基于空间重构的跨模态特征融合模块对背景冗余信息的有效过滤。通过显式解耦信息流与冗余流，模型能够更聚焦于物体核心区域，降低了环境杂波对目标定位的干扰。

然而，本文方法在极端场景下仍存在局限性。如图7的失败案例所示，在面对严重物体遮挡、光照不足或几何结构极度相似的极端情况时，模型对部分长尾新颖类别仍偶发漏检或边界框定位不准。这表明当双模态输入源信息严重缺失或存在歧义时，跨模态特征的解耦与重构能力仍面临挑战。未来研究将探索引入多帧时序融合机制或大语言模型的常识推理能力，利用时空上下文和逻辑推理单帧感知信息的不足，进一步提升算法在极端环境下的检测上限。

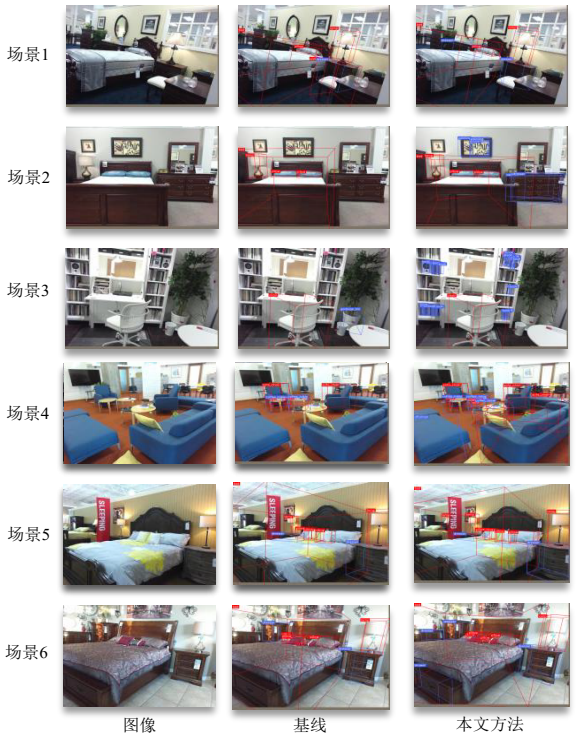

图6 对比实验结果可视化

Figure 6 Visualization of comparison experiment results

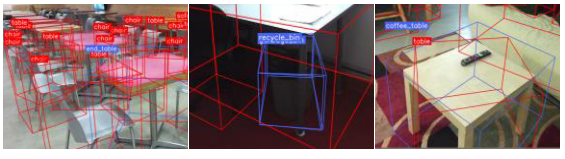

图7 失败案例

Figure 7 Failure cases

4 结论

本文提出了一种空间重构与文本引导的室内开放域3D目标检测方法。设计层级边缘聚焦网络,通过分流注意力机制过滤图像通道维度的冗余信息,并利用层级融合模块建模自适应层间关联,为3D检测分支提供了具有边缘增强和多尺度语义对齐的高质量视觉特征;同时,通过引入基于空间重构的跨模态特征融合模块,利用组归一化的统计特性将特征显式解耦为信息流与冗余流,通过双流交互机制增强显著特征并有效抑制背景噪声,实现了高价值特征的精准聚焦;此外,设计了语义增强的文本-查询交互机制,利用门控残差模块将文本语义先验提前注入解码器,使查询向量在交互前具备类别感知能力,突破了传统CLIP模型仅作为末端分类器的局限。实验结果表明,所提出的方法在SUN RGB-D数据集上展现出显著优势,增强了模型在室内复杂环境下的泛化能力与鲁棒性。

参考文献

[1] Lazarow J, Griffiths D, Kohavi G, et al. Cubify anything: Scaling indoor 3D object detection[C]//2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2025: 22225-22233.

[2] Wang Jiangyi, Zhao Na. Uncertainty meets diversity: A comprehensive active learning framework for indoor 3D object detection[C]//2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2025: 20329-20339.

[3] Zhu Chaoyang, Chen Long. A survey on open-vocabulary detection and segmentation: Past, present, and future[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 46(12): 8954-8975.

[4] Li Kaiyu, Cao Xiangyong, Deng Yupeng, et al. DynamicEarth: How far are we from open-vocabulary change detection [J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2026, 40(8): 6279-6287.

[5] Hu Yupeng, Ding Changxing, Sun Chang, et al. Bilateral collaboration with large vision-language models for open vocabulary human-object interaction detection[C]//2025 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2025: 20126-20136.

[6] Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision[C]//International Conference on Machine Learning. 2021: 8748-8763.

[7] 周治国, 马文浩. 一种多层多模态融合3D目标检测方法[J]. 电子学报, 2024, 52(3): 696-708.
Zhou Zhiguo, Ma Wenhao. 3D object detection based on multilayer multimodal fusion [J]. Acta Electronica Sinica, 2024, 52 (3): 696-708. (in Chinese)

[8] Xu Danfei, Anguelov D, Jain A. PointFusion: Deep sensor fusion for 3D bounding box estimation[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 244-253.

[9] Qi C R, Chen Xinlei, Litany O, et al. ImVoteNet: Boosting 3D object detection in point clouds with image votes[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2020: 4403-4412.

[10] 葛同澳, 李辉, 郭颖等. 基于双融合框架的多模态3D目标检测算法[J]. 电子学报, 2023, 51(11): 3100-3110. DOI: 10.12263/DZXB.20230414.
GE Tongao, LI Hui, GUO Ying, et al. A Multimodal 3D Object Detection Method Based on Double-Fusion Framework[J]. ACTA ELECTRONICA SINICA, 2023, 51(11): 3100-3110. DOI: 10.12263/DZXB. 20230414. (in Chinese)

[11] Vora S, Lang A H, Helou B, et al. PointPainting: Sequential fusion for 3D object detection[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2020: 4603-4611.

[12] Zhang Zehan, Zhang Ming, Liang Zhidong, et al. MAFF-net: Filter false positive for 3D vehicle detection with multi-modal adaptive feature fusion[C]//2022 IEEE 25th International Conference on Intelligent Transportation Systems. Piscataway: IEEE, 2022: 369-376.

[13] Tan Xun, Chen Xingyu, Zhang Guowei, et al. MBDF-net: Multi-branch deep fusion network for 3D object detection[C]//Proceedings of the 1st International Workshop on Multimedia Computing for Urban Data. New York: ACM, 2021: 9-17.

[14] Wang Zhixin, Jia Kui. Frustum ConvNet: Sliding Frustums to aggregate local point-wise features for amodal 3D object detection[C]//2019 IEEE/RSJ International Conference on Intelligent Robots and Systems. Piscataway: IEEE, 2019: 1742-1749.

[15] 臧传方, 党建武, 雍玖. 自适应融合多模态特征的6D物体位姿估计方法[J]. 激光与光电子学进展, 2025, 62(4): 0415002.
Zang Chuanfang, Dang Jianwu, Yong Jiu. Adaptive multimodal-feature fusion for 6D object position estimation[J]. Laser & Optoelectronics Progress, 2025, 62(4): 0415002. (in Chinese)

[16] Gu Xiuye, Lin Tsung-Yi, Kuo Weicheng, et al. Open-vocabulary object detection via vision and language knowledge distillation[PP/OL]. V3. arXiv (2022-05-12)[2026-07-01]. https://doi.org/10.48550/arXiv.2104.13921.

[17] Minderer M, Gritsenko A, Stone A, et al. Simple open-vocabulary object detection[C]//Computer Vision - ECCV

2022. Cham: Springer, 2022: 728-755.
[18] 樊琳, 龚勋, 郑岑洋. 基于文本引导下的多模态医学图像分析算法[J]. 电子学报, 2024, 52(7): 2341-2355.
Fan Lin, Gong Xun, Zheng Cenyang. A multi-modal medical image analysis algorithm based on text guidance [J]. Acta Electronica Sinica, 2024, 52 (7): 2341-2355. (in Chinese)
[19] Cheng Tianheng, Song Lin, Ge Yixiao, et al. YOLO-world: Real-time open-vocabulary object detection[C]// 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2024: 16901-16911.
[20] Liu Shilong, Zeng Zhaoyang, Ren Tianhe, et al. Grounding DINO: Marrying DINO with Grounded pre-training for Open-set object detection[C]//Computer Vision - ECCV 2024. Cham: Springer, 2025: 38-55.
[21] Wang Zhenyu, Li Yali, Liu Taichi, et al. OV-Uni3DETR: Towards unified open-vocabulary 3D object detection viaCycle-modality propagation[C]//Computer Vision - ECCV 2024. Cham: Springer, 2025: 73-89.
[22] Zhang Renrui, Guo Ziyu, Zhang Wei, et al. PointCLIP: Point cloud understanding by CLIP[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2022: 8542-8552.
[23] Zhu Xiangyang, Zhang Renrui, He Bowei, et al. PointCLIP V2: Prompting CLIP and GPT for powerful 3D open-world learning[C]//2023 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2023: 2639-2650.
[24] Zeng Yihan, Jiang Chenhan, Mao Jiageng, et al. CLIP2: Contrastive language-image-point pretraining from real-world point cloud data[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2023: 15244-15253.
[25] Cao Yang, Zeng Yihan, Xu Hang, et al. CoDA: Collaborative novel box discovery and cross-modal alignment for open-vocabulary 3D object detection[C]// Advances in Neural Information Processing Systems 36. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2023: 71862-71873.
[26] Jiao Pengkun, Zhao Na, Chen Jingjing, et al. Unlocking textual and visual wisdom: Open-vocabulary 3D object detection enhanced by comprehensive guidance from text and image[C]//Computer Vision - ECCV 2024. Cham: Springer, 2025: 376-392.
[27] Cao Yang, Zeng Yihan, Xu Hang, et al. Collaborative novel object discovery and box-guided cross-modal alignment for open-vocabulary 3D object detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2025, 47(11): 10475-10489.
[28] Peng Xingyu, Liu Si, Gao Chen, et al. GLRD: Global-local collaborative reason and debate with PSL for 3D open-vocabulary detection[PP/OL]. V1. arXiv (2025-03-26)[2026-07-01]. https://doi.org/10.48550/arXiv.2503.20682.

作者简介

樊怡麟 男,2002年5月出生于山东省淄博市。2026年毕业于青岛科技大学数据科学学院。现为软控股份有限公司算法工程师。主要研究方向为大模型应用、开放词汇多目标检测。
E-mail: fylqust@163.com

李 辉 男,1984年3月出生于河南省平顶山市。现为青岛科技大学数据科学学院副教授、硕士生导师。主要研究方向为计算机视觉,多目标检测与跟踪。
E-mail: lihui@qust.edu.cn

杨方正 男,2003年7月出生于山东省滨州市。现为青岛科技大学硕士研究生。主要研究方向为多目标检测与跟踪。
E-mail: fzyang@mails.qust.edu.cn

刘治宇 男,1988年3月出生于黑龙江省七台河市。现为中国海洋大学博士后。主要研究方向为多模态特征编码、AI高性能推理计算等方向。
E-mail: lzy1280@ouc.edu.cn

臧银珠 女,2003年3月出生于山东省日照市。现为青岛科技大学硕士研究生。主要研究方向为计算机视觉、三维建模。
E-mail: zangyinzhu@mails.qust.edu.cn

孔祥振 男,1990年12月出生于山东省聊城市。现为荷兰埃因霍芬理工大学工业工程学院博士后。主要研究方向为图像处理和目标检测。
E-mail: x.kong@tue.nl