像-点云-文本多模态融合的室内三维目标检测方法

樊怡麟[1]，季雨昂[1]，秦修功[2]，杨方正[1]，李 辉[1†]

(1. 青岛科技大学 数据科学学院，山东 青岛 266100；2. 北京机械工业自动化研究所有限公司，北京 100120)

摘 要: 三维目标检测对于提升智能系统在复杂室内环境中的感知与理解能力具有重要意义. 然而, 现有基于单模态点云的检测方法普遍存在语义信息不足、泛化能力受限等问题, 难以有效应对室内场景中新类别目标的检测需求. 针对上述问题, 提出一种图像-点云-文本多模态融合的室内三维目标检测方法. 该方法首先引入密集深度图引导的图像-点云早期融合策略, 通过深度约束将图像语义特征精确映射至三维空间, 有效增强点云的语义表达能力并缓解遮挡带来的空间错位问题; 然后, 设计混合查询引导的室内 Transformer 检测器, 采用几何查询与可学习查询相结合的双分支查询机制, 在兼顾局部目标精细建模的同时强化场景级上下文建模能力; 最后, 提出动态解耦 3D-IoU 损失增强策略, 通过解耦空间梯度并根据目标尺度动态调整权重, 提高新物体候选框的定位质量与发现能力. 在 SUN-RGBD 数据集上的实验结果表明, 所提出方法在多项评价指标上均优于现有先进方法, 验证了其在室内开放域三维目标检测任务中的有效性与鲁棒性.

关键词: 多模态融合；室内；三维目标检测；密集深度图；混合查询；动态解耦

中图分类号: TP391 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2025.1112

引用格式: 樊怡麟, 季雨昂, 秦修功, 等. 像-点云-文本多模态融合的室内三维目标检测方法 [J]. 控制与决策

Image-point cloud-text multimodal fusion method for indoor 3D object detection

FAN Yi-lin[1]，*JI Yu-ang*[1]，*QIN Xiu-gong*[2]，*YANG Fang-zheng*[1]，*LI Hui*[1†]

(1. College of Data Science，Qingdao University of Science & Technology，Qingdao 266100，China；2. Beijing Institute of Mechanical Industry Automation Co.，Ltd.，Beijing 100120，China)

Abstract: 3D object detection holds significant importance in enhancing the perception and understanding capabilities of intelligent systems in complex indoor environments. However, existing detection methods based on single-modality point clouds generally suffer from issues such as insufficient semantic information and limited generalization ability, making it difficult to effectively address the detection needs of new categories of objects in indoor scenes. To address these issues, this paper proposes an indoor three-dimensional object detection method that integrates image-point cloud-text multimodal fusion. Firstly, the method introduces an early fusion strategy for image-point cloud based on dense depth maps, accurately mapping image semantic features to three-dimensional space through depth constraints, effectively enhancing the semantic expression ability of point clouds and alleviating spatial misalignment issues caused by occlusion. Secondly, a hybrid query-guided indoor Transformer detector is designed, utilizing a dual-branch query mechanism combining geometric queries and learnable queries, which simultaneously considers fine-grained modeling of local objects and strengthens scene-level context modeling capabilities. Finally, a dynamic decoupling 3D-IoU loss enhancement strategy is proposed, which decouples spatial gradients and dynamically adjusts weights based on object scale, improving the localization quality and detection ability of new object candidate boxes. Experimental results on the SUN-RGBD dataset demonstrate that the proposed method outperforms existing state-of-the-art methods in multiple evaluation metrics, validating its effectiveness and robustness in indoor open-domain three-dimensional object detection tasks.

Keywords: multimodal fusion；indoor；3D object detection；dense depth maps；hybrid query；dynamic decoupling

收稿日期: 2025-10-24；**录用日期:** 2026-01-27.

基金项目: 国家重点研发计划项目 (2023YFF0612100)；山东省自然科学基金项目 (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”项目 (2024HT030).

责任编委: 王琦.

[†]通信作者. E-mail: lihui@qust.edu.cn.

0 引 言

随着人工智能的迅速发展, 智慧家庭已经成为现代家庭的新潮流, 家庭服务机器人已逐渐成为家庭中的一个重要组成部分. 然而, 要让家庭服务机器人真正融入人类的生活环境, 具备对家庭空间中的各种目标的准确识别和理解能力是重要前提. 室内目标检测技术[1-3], 可以实现对家庭环境中的人员、物体和行为等目标进行实时监测和识别, 它的发展进一步提升了家庭空间的智能化水平.

传统三维目标检测旨在预定义类别集合上实现三维边界框的定位与分类, 近年来在室外场景已取得进展. 然而在室内环境中, 类别呈长尾分布且存在大量训练阶段未见的新物体, 使得封闭类别设定难以满足实际应用需求. 因此, 开放词汇三维目标检测在近年来迅速发展. 开放词汇三维目标检测是计算机视觉领域的重要研究方向[4-7], 旨在检测出训练阶段未见过的任意类别三维物体, 该研究方向对于提升智能系统在复杂室内环境中的感知能力具有重要意义. 随着大规模视觉-语言预训练模型 CLIP[8] 的快速发展, 二维开放词汇检测取得了显著进展. 然而, 受限于三维数据的稀疏性和语义信息缺失, 直接将二维场景的成功经验迁移至三维场景仍面临巨大挑战, 例如, 在复杂室内场景中新类别物体的准确定位与分类十分困难. 现有的开放词汇三维目标检测方法主要利用大规模预训练的 2D 开放词汇物体检测模型 (OV-2DDet) 作为语义先验, 在二维图像中定位潜在的新颖物体区域, 再通过几何映射或多视角投影为对应目标生成伪三维边界框标签, 这类方法借助 2D 模型来定位大量 2D 新颖物体框[9], 随后为相应的 3D 新颖物体生成伪 3D 物体框标签,其在一定程度上缓解了三维场景中语义信息不足的问题. 同时, 三维目标检测正在向多模态方向发展, 其中通过特征投影、拼接或注意力交互等方式引入图像语义信息已被广泛采用, 多模态数据的深度协同与互补已成为克服单模态感知局限、增强模型泛化能力的核心途径.

尽管国内外在三维目标检测、多模态融合以及开放域目标检测等方向已取得一定进展, 但现有研究在室内开放域三维目标检测任务中仍存在明显不足. 一方面, 部分方法依赖外部二维开放词汇检测器或伪标签生成策略, 限制了模型在三维空间中的端到端学习能力与空间一致性; 另一方面, 现有图像与点云融合方法多采用简单投影或特征拼接方式, 缺乏深度约束与遮挡建模, 难以实现高精度的跨模态对齐; 此外, 室内场景目标密集、尺度差异大, 对全局上下文建模与候选框定位精度提出了更高要求, 而现有查询机制与损失建模策略在兼顾全局感知与局部精细定位方面仍存在局限.

为解决上述问题, 本文提出一种基于多模态融合的开放域三维目标检测方法, 以图像–点云–文本三模态融合对齐为核心主线, 将二维图像的语义表达与语言先验通过跨模态对齐注入到稀疏点云表示中, 从而提升三维目标检测的语义判别能力及对新类别的泛化能力. 针对外部依赖问题, 设计端到端的联合学习策略, 无需依赖外部二维检测器即可实现新物体的自主发现与分类; 本文仅将 CLIP 作为冻结的图像-文本编码器, 为 3D 候选框提供语义打分, 而不引入任何具有检测头的 2D 开放词汇检测器模型, 从而避免了 2D 到 3D 伪标注带来的误差传递. 针对跨模态融合局限, 提出密集深度图引导的图像-点云早期融合策略, 通过深度引导将图像语义特征精确映射至体素空间, 实现全局语义信息的有效传递; 针对室内检测场景, 提出混合查询引导的室内 Transformer 检测器, 借助几何查询与可学习查询的混合引导机制, 结合自适应注意力层结构, 增强对室内场景中三维目标的空间几何与语义特征的精准感知与表达能力; 最后, 设计动态解耦 3D-IoU 的损失增强策略, 通过对三维交并比在 xy 平面与 z 方向的解耦优化并动态更新融合权重, 提升对高质量新物体候选框的发现能力. 实验结果表明, 本文方法相比基线方法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) 提出密集深度图引导的早期融合模块, 该模块通过生成密集深度图, 将图像语义特征精准映射至三维空间, 有效缓解室内场景中点由遮挡引起的跨模态空间错位问题. (2) 提出一种混合查询引导的室内 Transformer 三维目标检测器, 通过几何查询和可学习查询的双分支查询机制, 有效兼顾室内的全局信息和局部特征. (3) 提出动态解耦 3D-IoU 损失增强, 通过在类无关检测器中引入动态解耦 3D-IoU 损失, 切断两个空间的梯度耦合, 避免了不同方向信息之间的相互干扰.

1 相关工作

1.1 点云目标检测

近年来, 点云目标检测方法经历了从投票方法到基于 Transformer 和体素方法的演变. 其中, Qi 等人[10] 提出的 VoteNet 开创了基于点投票机制进行三维检测的先河, 采用投票和聚类方法生成近似目标

中心点的集群, 将集群送入区域生成网络 (RPN) 处理, 取得了不错的效果. BRNet[11]、H3DNet[12] 和 RBGNet[13] 不断优化点组生成和特征聚合过程, 进一步提升了检测精度. 然而, 投票方法容易受到室内杂乱背景点的干扰, 导致预测的中心点不准确. 为了更有效地利用三维数据, Zhou 等人[14] 利用体素特征编码层将体素内的点特征编码为高维的特征数据, 再利用 RPN 生成检测框, 虽然该方法取得了较好的检测结果, 但在室内场景下体素特征编码层中的三维卷积显著增加了模型的计算量. FCAF3D[15] 等体素方法利用稀疏三维卷积, 实现了更高效的特征提取, 但受到体素化处理中量化误差的影响, 检测精度受到一定限制. 随着 Transformer 架构在计算机视觉领域的广泛应用, GroupFree[16] 和 3DETR[17] 等方法通过迭代更新目标查询位置, 减少了对领域特定超参数的依赖, 真正实现了端到端训练.

1.2 跨模态目标检测

为了充分利用图像和点云数据各自的优势以获取目标更加丰富和判别性的特征, 跨模态特征融合图像和点云信息, 能够克服图像或点云单模态检测方法的局限, 有效地提升三维目标检测性能. Qi 等人[18] 在 VoteNet 的基础上提出 ImVoteNet, 增加图像作为网络输入, 将图像特征附加到点云上. Xu 等人[19] 提出 PointFusion, 以输入的三维点作为空间锚点, 分别利用全局融合和稠密融合模块处理图像和点云上的目标特征融合. Vora 等人[20] 借助传感器映射参数将三维点云特征与获取的图像分割特征完成拼接融合. 于等人[21] 提出使用全局感知模块提取图像语义信息, 使用深度去噪模块滤除深度噪声, 采用非对称跨模态融合方法弥补图像和深度图之间的差异性. Zhang 等人[22] 针对图像和三维点云数据的特性分别设计稠密注意力融合和点注意力融合模块, 有效地降低三维检测时的假阳性. 考虑到图像和点云数据的差异, Tan 等人[23] 首先利用自适应注意力融合机制从单模态特征中生成跨模态融合的特征, 之后利用池化的 RoI 融合模块生成增强的局部特征, 但该方法对小目标的检测效果较差. Wang 等人[24] 设计 PointNet 与二维区域对应的三维锥体空间进行聚合得到融合的锥体特征, 并使用锥体特征进行预测框回归, 大大减少了模型计算量. 佟等人[25] 提出了一种名为 VSIL-SLAM 的融合框架, 通过将聚类后的激光点云依据投影关系映射至视觉语义检测框内, 实现语义物体的构建, 有效弥补了原始点云数据中特征稀疏的不足.

1.3 开放域目标检测

随着 CLIP 等视觉语言预训练模型的出现, 基于视觉语言预训练技术实现突破性进展, 开放域目标检测领域实现了跨越式发展. Kuo[26] 等提出一种基于冻结视觉语言模型 F-VLM 的开放词汇目标检测新范式, 解耦视觉-语言特征学习与检测任务适配, 规避了知识蒸馏和定制化检测预训练的技术依赖. Wu 等[27] 提出 CORA 框架, 通过区域特征提示和类别感知锚点预匹配机制, 将 CLIP 集成到开放词汇检测中, 分别缓解区域分布差异并增强目标定位泛化能力. 此外, Li 等[28] 通过自适应上下文归一化与时序特征建模提升了模型对未知类别和跨域场景的泛化能力, 为开放词汇目标检测提供了启示. Qorbani 等[29] 提出通过检索与融合 LoRA 适配器实现无需训练的测试时域自适应, 有效提升了开放词汇模型在跨域场景下的泛化能力. Gupta 等[30] 提出的 OW-DETR 框架通过显式建模多尺度上下文特征和跨类别知识迁移通道, 实现已知类到未知类的知识迁移. Zhao[31] 等提出了分层跨模态对齐框架, 通过联合建模局部目标与全局场景语义, 有效缓解了视觉语言模型在开放词汇三维目标检测中场景上下文缺失的问题. 最近, RegionCLIP[32] 通过构建区域感知的跨模态对齐机制, 将 CLIP 模型的图像级语义对齐能力解耦并延伸至细粒度区域层级, 实现了区域视觉特征与文本语义空间的精准动态映射. Cheng 等[33] 提出 YOLO-World, 通过引入视觉语言路径聚合网络, 结合区域-文本对比损失函数, 有效促进了视觉与语言信息的深度交互. Liu 等人[34]提出一种多模态协同检测框架 Grounding DINO, 通过构建视觉特征增强模块、语言引导查询选择机制、多模态融合解码器与跨模态注意力解码器, 有效实现了语言先验与视觉表征的语义对齐. Wang 等人[35] 提出了适用于室内和室外场景的多模态架构 OV-Uni3DETR, 通过二维和三维模态间的知识循环传播机制, 实现了多种模态和不同场景下高效检测.

2 本文方法

本文多模态 3D 室内开放域目标检测算法框架如图 1 所示. 主要包括密集深度图引导的早期融合、混合查询引导的室内 Transformer 检测器、基于动态解耦 3D-IoU 的新物体发现策略三部分. 首先, 通过边缘感知深度补全模块生成密集深度图, 基于深度信息辅助增强 2D 图像特征对遮挡区域的感知能力并将 2D 图像特征升维至 3D 空间, 生成与点云分辨

率严格匹配的 3D 图像特征, 实现双模态特征的空间对齐; 其次, 引入混合查询引导的室 Transformer 检测器, 采用可学习查询和几何查询的双分支查询设计, 在实现聚焦物体局部特征的精准捕捉的同时扩大对室内的感受野范围, 实现场景级上下文信息的整合, 同时在编码器结构中引入自适应注意力层在不额外增加参数的基础上, 增强室内特征的全局语义信息. 最后, 在新物体发现策略中引入动态解耦 3D-IoU 损失切断两个空间的梯度耦合并动态独立更新, 为新物体发现提供更高质量的语言先验.

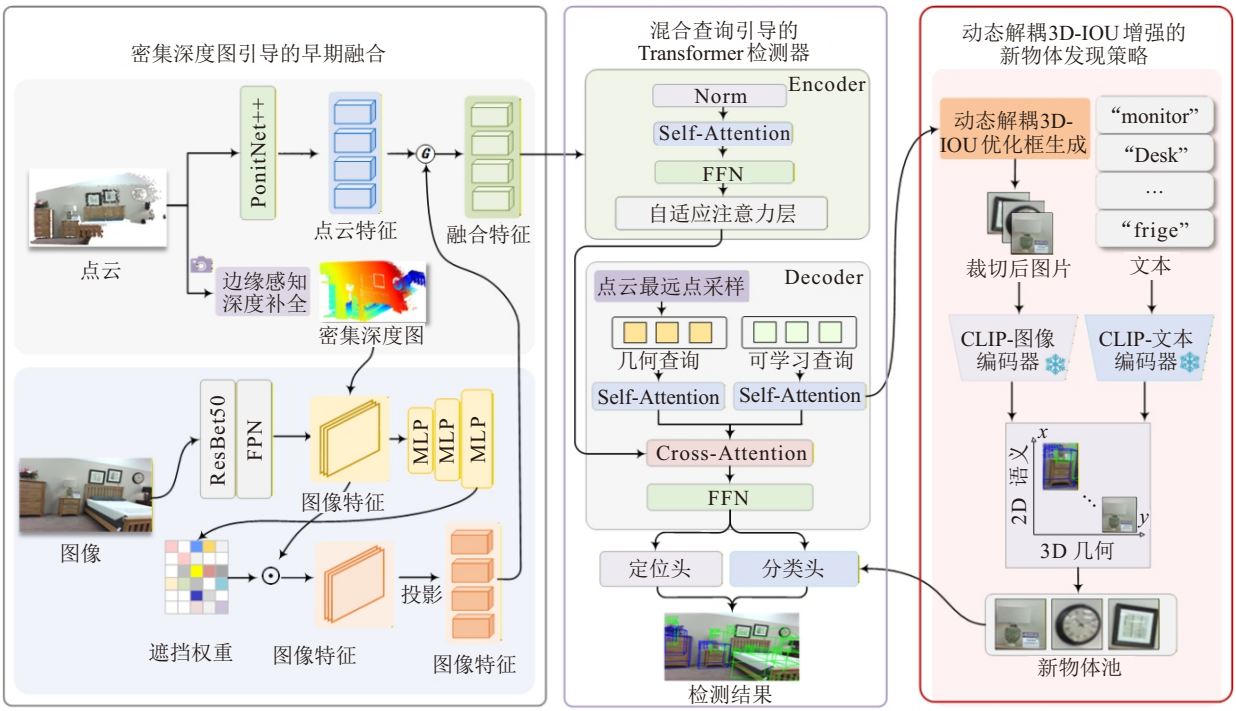


图1 提出方法的网络结构

2.1 密集深度图引导的早期融合模块

在多模态协同学习框架下, 图像特征所蕴含的高层语义表征可为点云注入丰富的先验知识. 然而, 现有的常规融合方法多依赖于稀疏点云的直接投影或简单的特征拼接, 缺乏对三维几何结构的深度理解, 容易导致特征空间错位. 因此, 本文设计一种密集深度图引导的早期融合模块. 不同于传统的直接映射, 本模块首先引入边缘感知机制构建密集深度图以解决稀疏性问题; 更关键的是, 利用几何信息生成自适应遮挡权重, 通过像素级的加权过滤为图像特征提供精确的几何约束, 从而有效解决由物体遮挡引发的语义歧义. 密集深度图引导的早期融合模块网络如图 2 所示.

该网络的输入为点云和图像数据. 具体处理流程如下: 首先, 利用 PointNet++提取点云特征F_{point}; 同时, 根据相机内参将点云坐标映射到二维平面生成深度图. 具体而言, 通过三维点云坐标P_{3d}与相机内参矩阵 K, 将三维点投影至图像平面P_{2d}得到对应的图像坐标 (u_i,v_j), 同时提取并归一化对应的深度信息d_{ij}. 基于投影坐标和深度信息, 构建深度图D_{map}和法向量图N_{map}. 其中, 法向量图N_{map}通过深度计算的空间梯度近似局部表面法向量, 具体公式为:

$$D_{map}[u_i,v_j]=d_{ij}, \tag{1}$$

$$N_{map}[u_i,v_j]=\nabla d_i=[\frac{\partial d}{\partial u},\frac{\partial d}{\partial v}](u_i,v_i). \tag{2}$$

接着, 对生成的深度图D_{map}进行边缘检测. 利用 Sobel 算子在x、y方向分别进行边缘检测, 得到水平边缘响应和垂直边缘响应. 通过计算梯度幅值生成边缘强度图, 再经过阈值处理得到二值化边缘掩码M. 基于边缘掩码M, 采用滑动窗口进行深度补全. 具体操作如下: 对于深度图中缺失的像素 (u,v), 在窗口内搜索有效深度值, 并通过距离加权平均计算补全值. 具体公式如下:

$$\begin{aligned}S=\{(u',v')\in W|&D_{map}(u',v')\neq 0\text{且}\\&M(u',v')=1\},\end{aligned} \tag{3}$$

$$D_{complete}(u,v)=\frac{\sum\limits_{(u',v')\in S}\omega(u',v')\cdot D_{map}(u',v')}{\sum\limits_{(u',v')\in S}\omega(u',v')}. \tag{4}$$

其中W代表以(u,v)为中心的窗口, 大小为 7×7, S为窗口内的有效深度点集合, $\omega(u',v')$为每个有效

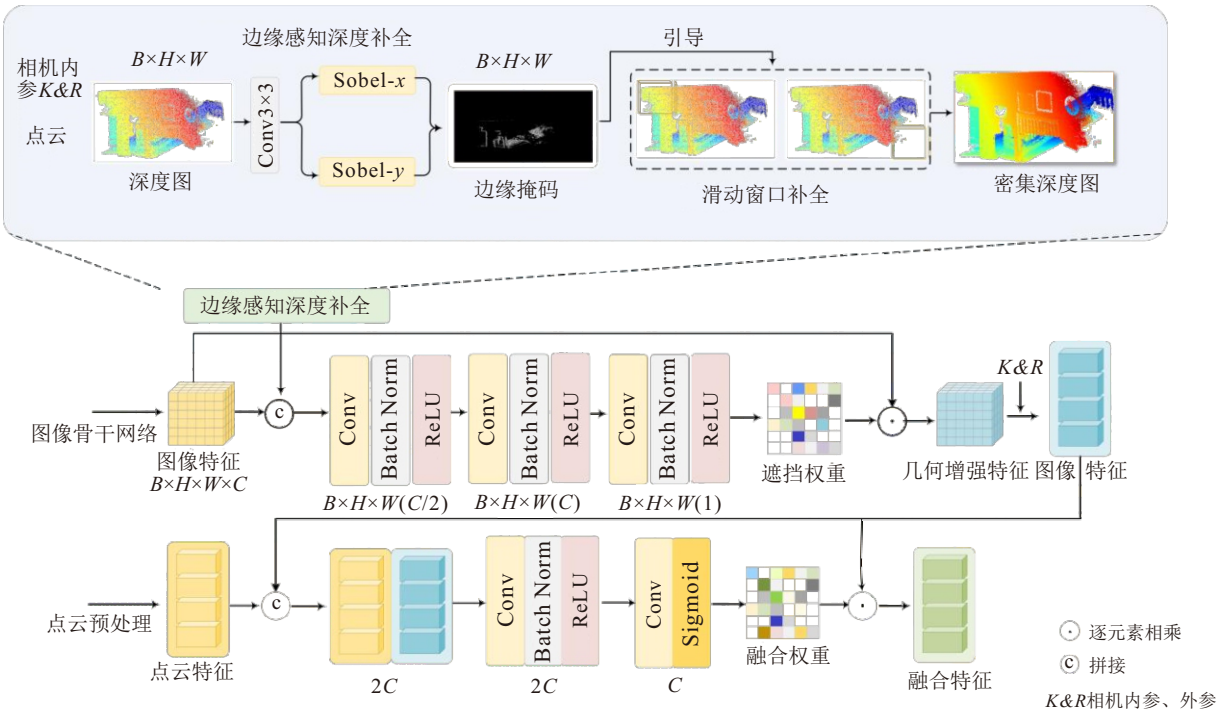


图2　密集深度图引导的早期融合模块

深度点到中心的欧式距离. 经过滑动窗口补全后, 得到密集深度图$D_{complete}$.

在传统点云–图像融合中, 仅依赖相机内参将三维点投影至二维平面会引入固有的视锥投影歧义, 即位于同一投影射线上的前景与背景点云会映射到相同像素位置, 从而在存在遮挡时导致背景点云错误地引入前景语义. 为缓解该问题, 本文利用生成的密集深度图作为可见性几何约束, 并联合法向量图与图像特征在通道维度进行融合. 本文的图像特征提取模块采用 ResNet50 和 FPN 提取图像多尺度特征, 通过双线性插值上采样进行逐层融合得到图像特征F_{2d}. 随后, 将得到的密集深度图$D_{complete}$和法向量图N_{map}与图像特征F_{2d}在通道维度进行拼接. 拼接后的特征被输入至一个轻量级降维模块, 用于学习生成像素级遮挡权重W_{img}, 实现几何一致性的软门控. 然后将该权重与图像特征F_{2d}逐像素相乘, 得到增强图像特征F_{img}. 该模块通过结合深度结构与边缘信息, 能够有效感知深度突变区域并识别潜在遮挡. 当投影点与其对应像素在几何上不一致时, 网络自适应地产生较低权重, 从而抑制由遮挡引起的错误语义传递. 最终, 经权重调制后的图像特征仅保留与点云空间位置一致的语义信息, 实现更加准确的跨模态对齐.

为了实现图像与点云特征的深度融合, 本文设计了门控线性单元的自适应融合机制. 将点云特征F_{point}与图像特征F_{img3d}拼接后输入到由二维卷积、归一化、激活函数组成的门控机制, 自适应调节点云和图像的权重进行融合. 预测一个取值范围在 (0,1) 的软门控融合权重W_{fuse}, 用于动态调节模态间的贡献度. 该门控融合机制不仅实现了多模态信息的有效互补, 更充当了特征质量的自适应滤波器. 在室内复杂场景中, 当由于遮挡或传感器噪声导致深度图补全质量不佳时, 门控网络能够通过端到端的学习降低该区域图像特征的融合权重, 从而避免低质量的深度先验对点云原始几何特征造成干扰, 有效防止了特征融合的退化问题.

2.2　混合查询引导的室内 Transformer 检测器

现有基于 Transformer 的 3D 检测器通常仅采用可学习查询点, 但针对室内场景物体密集分布、全局上下文信息丰富的特性, 单一查询机制难以同时满足全局场景覆盖与局部物体精准定位的需求. 单使用可学习查询易受局部特征主导而忽略全局物体关联, 而只使用固定采样查询则缺乏对物体细节的自适应捕捉. 针对室内场景物体密集堆叠且遮挡严重的特性, 本文构建了几何覆盖-语义精修的混合查询机制. 利用几何查询的广义空间分布来捕捉被遮挡或新出现的物体候选, 同时利用可学习查询聚焦于物体中心的语义特征对齐. 同时研究发现在编码器中引入自适应注意力层, 在无需额外增加参数的基础上, 可以增强对室内场景中全局语义信息的提取, 提高检测性能. 检测器具体结构如图 3 所示.

在室内场景中, 可学习查询在训练过程中会逐渐更新并聚焦于物体实例附近. 其初始化为服从正态分布的 3D 坐标, 并嵌入维度为 256 的特征向量.

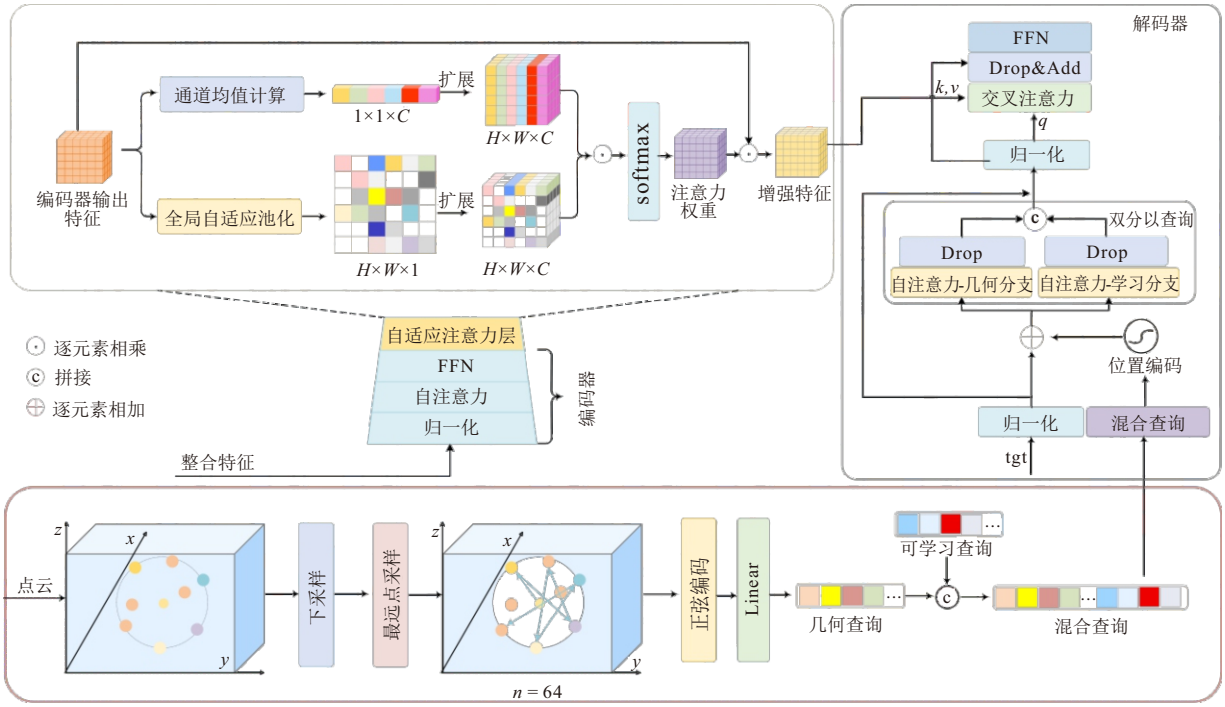


图3 混合查询引导的室内 Transformer 检测器

在 Transformer 解码器中, 可学习查询通过自注意力机制建模查询间的关联性, 同时通过交叉注意力机制与编码器输出的特征进行交互, 迭代优化自身的位置与特征. 每一层解码器通过预测查询点相对于物体中心的偏移量$(\Delta x, \Delta y, \Delta z)$, 逐步将可学习查询引导至物体局部特征显著区域, 从而提升对物体边界的定位精度. 但在室内场景中, 物体范围与场景尺寸近似, 仅依赖局部信息会忽略全局结构. 如果仅使用固定的非可查询点会导致无法自适应物体变化, 所以本文提出了一个基于几何查询和可学习查询的混合查询.

几何查询通过最远点采样 (FPS) 从点云中生成, 直接采样会导致内存过载. 设原始点云为$P = \{p_1, p_2, \ldots, p_n\}$, 先对输入点云进行随机下采样, 得到下采样后的点云$P_{down} = \{p_1, p_2, \ldots, p_{20000}\}$. 基于下采样点$P_{down}$, 随机初始化一个点, 通过计算所有剩余点到已采样点的最小欧式距离, 选择距离最大的点加入采样集, 生成固定数量的几何查询点P_{geo}. 为了使几何查询点适配 Transformer 的特征交互, 对采样过后的几何查询点P_{geo}进行位置编码与特征嵌入, 通过正弦位置编码 PE 将 3D 坐标映射为高维位置特征, 再通过线性层嵌入为与可学习查询Q_{Learn}同维度的特征向量,具体公式如下:

$$PE(q) = \left[\sin(\frac{q_x}{\theta})\cos(\frac{q_x}{\theta})\sin(\frac{q_y}{\theta})\cos(\frac{q_y}{\theta})\sin(\frac{q_z}{\theta})\cos(\frac{q_z}{\theta})\right]^T \tag{5}$$

$$f_{geo} = Linear(PE(q_t)). \tag{6}$$

其中q_x, q_y, q_z为查询集合点的三维坐标, θ为缩放因子, 等于$10^{2d/D}$, $Linear()$为无偏执线性层, f_{geo}为生成的查询特征. 最终几何查询特征集为$Q_{geo} = \{f_{geo,1}, f_{geo,2}, \ldots, f_{geo,n}\}$. 为实现几何查询$Q_{geo}$的全局覆盖和可学习查询$Q_{Learn}$的局部精准互补, 设计自注意力分组交互机制, 避免双分支特征相互干扰, 同时通过与编码器自适应注意力层的交互, 增强全局语义提取. 将混合查询分为几何查询组与可学习查询组后, 仅在组内进行自注意力计算, 保持各自的信息特性, 具体为:

$$Q_L = SelfAttn(Q_{Learn}, Q_{Learn}, Q_{Learn}), \tag{7}$$

$$Q_G = SelfAttn(Q_{geo}, Q_{geo}, Q_{geo}), \tag{8}$$

$$Q = Contact(Q_L, Q_G). \tag{9}$$

其中$Contact()$为特征拼接操作. 与传统的双分支结构直接拼接查询并统一计算注意力不同, 分组交互策略将几何查询与可学习查询仅在组内进行自注意力更新. 这种设计通过物理隔离, 有效防止了基于 FPS 采样的全局几何分布特征被局部语义特征稀释, 从而保证了对室内长尾物体的召回能力.

在损失计算阶段, 为了解决异构查询在优化过程中的梯度冲突问题, 本文采用了独立匹配监督策略. 几何查询和可学习查询分别生成预测框, 通过匈牙利算法与 Ground Truth 独立匹配并计算损失; 最终预测时, 基于 3D-IoU 对两组预测框进行聚类去重, 取聚类内置信度最大值作为最终结果, 有效融合

全局覆盖与局部精准的双重优势, 解决室内密集物体的漏检与误检问题.

同时在编码器中引入自适应注意力层, 通过全局平均池化聚合全局语义, 使用通道均值生成注意力权重, 通过扩展将权重提升至特征尺寸并进行逐元素相乘, 得到对应的几何权重, 实现了在无需额外引入参数情况下通道-空间双重注意力协同增强. 既能动态抑制冗余通道、聚焦物体核心区域, 又能以轻量级计算提升特征与点云几何的语义一致性, 并行计算降低了通道和空间的相互干扰, 在跨模态任务中实现从通用特征提取到任务导向增强的升级, 有效提升复杂场景下的特征表达鲁棒性, 适合上下文信息丰富的室内场景. 从整体结构上看, 该自适应注意力层与 Squeeze-and-Excitation(SE)Block 在设计思路上具有一定相似性, 均通过全局统计信息对特征进行调制. 不同之处在于, SE Block 通常应用于卷积神经网络中, 主要用于通道维特征的重标定; 而本文的自适应注意力层被引入 Transformer 编码器结构, 面向室内三维检测任务, 在多模态特征编码阶段同时考虑通道与空间维度的全局语义信息. 通过这种方式, 模型能够在不显著增加计算或参数开销的前提下, 更好地利用场景级上下文信息, 从而提升特征表示在复杂室内环境中的稳定性与鲁棒性.

2.3　基于动态解耦 3D-IoU 的新物体发现策略

如图 4 所示, 动态解耦 3D-IoU 的新物体发现策略的核心目标是在训练过程中动态精确定位新颖物体. 不同于依赖 OV-2DDet 的方法, 本文利用基础类 3D 先验训练类无关 3D 检测器, 结合动态解耦 3D-IoU 直接在三维空间生成高质量候选框. 随后将候选框投影裁剪并输入冻结的 CLIP 实现语义对齐, 其中 CLIP 仅提供语言先验, 不参与检测与回归.

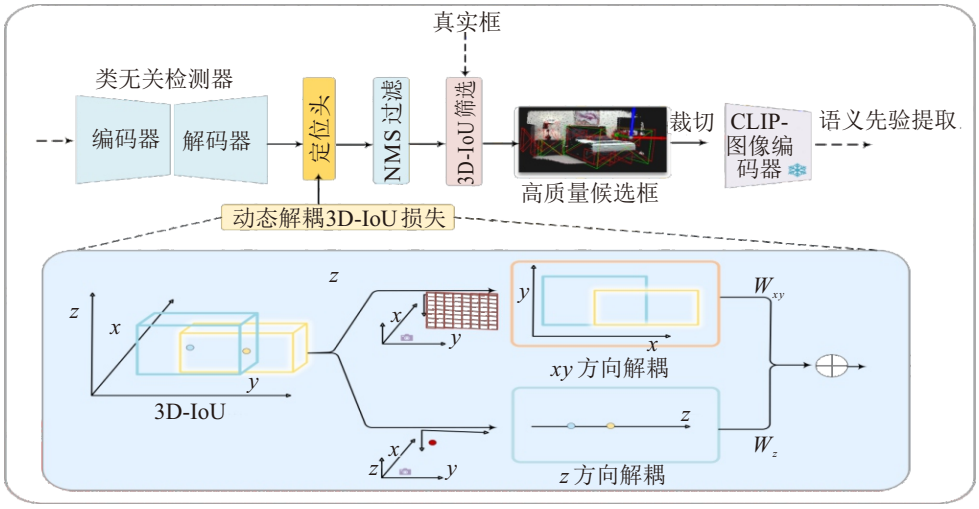


图4　基于动态解耦 3D-IoU 的新物体发现策略

首先, 在训练集上训练一个基于动态解 3D-IoU 监督的类无关检测器. 在训练过程中通过最小化检测器的回归损失和不考虑特定的分类损失, 筛选出物体候选框$B=\{b_1,b_2,\ldots,b_n\}$集合和对应的物体性概率p_g^i. 对候选物体框集 B 使用非极大值抑制去除掉冗余的物体框, 同时与训练集中的基础类别O_{base}的真实值使用 3D-IoU 阈值筛选出潜在的新物体框, 将新物体框依据相机参数映射至二维空间, 得到其二维中心点与尺寸, 然后将通过裁剪得到的对应的二维图像和经过预处理的 LVIS 的超类别列表输入到 CLIP 编码器, 分别得到相应的文本特征和图像特征, 通过计算图像特征与每个类别的语义概率p_s, 得到新最大语义置信度$p_s^{\max}$,筛选出的具体规则如下:

$$O_{novel}=\{O_j|O_j\notin C_{seen},\forall O_i'\in O_{base},\ IoU_{3d}(O_j,O_i')<0.25,p_g^j>\theta_g,p_s^{\max}>\theta_s\}. \quad (10)$$

其中C_{seen}是已发现的类别, θ_g为几何置信度阈值, θ_s为语义置信度阈值, 遵循 CoDA[4] 里的设置.

在训练类无关检测器的定位损失时, 传统 3D-IoU 采用乘法形式耦合xy平面与z轴回归, 使二者在反向传播过程中相互制约. 在室内场景中, 由于物体高度受空间结构约束, z轴回归在训练早期更易优化, 模型往往通过优先降低z轴误差来快速减小损失, 从而抑制了对更具挑战性的xy平面定位. 更进一步, 在开放词汇 3D 检测中, 新类别物体往往具有与基础类别显著不同的尺度与长宽比分布, 固定权重的 IoU 融合方式容易使模型过度依赖基础类别的几何先验, 导致对极端尺寸物体的泛化能力不足. 为此, 本文提出动态解耦 3D-IoU 策略, 根据物体真实体积

自适应调节xy平面与z轴的损失权重, 在缓解梯度耦合的同时显式建模跨类别的尺度分布差异, 从而有效提升了新物体的定位与发现能力.

传统的 3D-IoU 策略将xy平面的面积重叠与z轴的高度重叠直接相乘, 导致两个空间的梯度相互耦合, 导致在优化xy平面的定位时会干扰z轴的梯度更新, 从而造成训练不稳定, 在室内场景中存在物体密集且高度重叠的特点, 传统的 3D-IoU 的耦合梯度会导致模型难以同时优化xy平面的横向定位和z轴的高度定位. 对于缺乏先验几何知识的未见新物体, 极低的IoU_z往往会导致xy平面的梯度接近于零, 使得模型在定位初期因高度估计偏差而实际上停止学习平面位置. 同时室内场景中物体尺度差异大的特点逐渐成为定位精度的关键瓶颈, 且多伴随高度重叠. 在此基础上本文提出了解耦动态 3D-IoU, 切断两个空间的梯度耦合并独立更新, 有效缓解在两个方向梯度相互影响的特点. 随着训练的进行动态调整两个方向的权重. 为此基于物体真实尺度动态调整权重. 同时. 具体公式如下:

$$IoU_{xy} = \frac{Area_{overlapped}}{Area_1 + Area_2 - Area_{overlapped}}, \tag{11}$$

$$IoU_z = \frac{z_{overlapped}}{z_1 + z_2 - z_{overlapped}}, \tag{12}$$

$$IoU_{dyn-de} = w_{xy} \cdot IoU_{xy} + w_z \cdot IoU_z. \tag{13}$$

其中$Area_{overlapped}$为xy的重叠面积, $z_{overlapped}$为z轴的重叠高度, $Area_1$、$Area_2$与z_1、z_1分别为预测框和真实框在对应维度的面积或高度. w_{xy}和w_z分别为两个方向上的权重. 为确保训练初期模型对xy平面与z轴的定位能力均衡发展, 训练前期模型将w_{xy}、w_z的初始值均设为 0.5, 均等的权重可避免因监督偏向某一维度导致的检测能力失衡, 为后续动态调整奠定均衡的基础定位能力. 随着训练推进, 室内场景中物体尺度差异大的特点逐渐成为定位精度的关键瓶颈, 因此本文基于物体真实尺度动态调整权重. 以真实框体积V_{gt}为尺度指标, 定义尺度归一化系数β:

$$\beta = \frac{\log(V_{gt} + 1)}{\log(V_{\max} + 1)}. \tag{14}$$

其中$V_{\max}$为数据集中最大物体体积, 引入$\log()$函数可有效压缩极端尺度差异带来的数值跨度, 使β稳定落在 [0,1] 区间, 具体为:

$$\begin{aligned} \omega_{xy} &= 0.5 + \lambda \cdot (\beta - 0.5), \\ \omega_z &= 0.5 - \lambda \cdot (\beta - 0.5). \end{aligned} \tag{15}$$

其中λ为调节因子, 用于控制融合权重对物体尺度变化的敏感程度. 考虑到室内场景在中物体存在显著的尺度长尾分布特性, 需要一个适中的λ值以平衡大小物体的梯度贡献. 基于后续的参数敏感性实验, 本文将λ设定为 0.3. 通过使用动态权重机制, 让解耦动态 3D-IoU 不仅解决了梯度耦合问题, 更实现了训练前期依托相同的初始权重建立均衡定位能力, 中后期随β反馈不断调整方向, 为类无关检测器的稳定训练与新物体的精准定位提供了更优的监督信号.

3 实验与结果分析

为评估所提方法的有效性, 本文方法在 SUN RGB-D 标准数据集上进行实验, 该数据集中的数据由 3D 传感器和彩色摄像机采集得到, 包含 10335 张 RGB-D 图像. 在 SUN-RGBD 数据集上, 本文将官方数据集的训练数据分为 5285 个训练集用于网络模型的训练、5050 个验证集用于模型的验证. 每个训练样本拥有带方向的 3D 边界框标注, 涵盖 46 个物体类别, 每个类别均拥有超过 100 个训练样本. 对于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, 实验中使用平均精度 (Average precision, AP) 和平均召回率 (Average Recall, 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, 本文采用 IoU 阈值为 0.25 时的平均精度和平均召回率作为验证集性能指标.

3.1 实验环境设置

本文实验所使用的硬件设备和环境配置信息如表 1 所示, 在 GPU 为 NVIDIA RTX 4090 的服务器上训练和测试.

表1 实验配置

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

实验过程中, 所提方法以 CoDA[4] 作为基线网络并在 SUN-RGBD 数据上训练. 沿用基线网络的思想, 本文使用类无关蒸馏损失进行第一阶段检测器的训练, 总计 1080 个 epoch, 随后加入基于动态解耦 3D-IoU 的新物体发现策略和 CoDA 提出的跨模态对齐方法, 训练 200 个 epoch. 文本使用 PyTorch

实现模型. 批次大小为 6; 使用 AdamW 优化器, 初始学习率为 1.97×10^{-4}, 权重衰减为 0.1. 为了稳定训练, 本文应用了最大 L2 范数为 0.1 的梯度裁剪. 为了保证公平, 语义置信度和几何置信度均遵循 CODA 相同的设置. 所有比较方法均使用相同数量的 epoch. 训练过程中使用的超参数设置遵循默认的 3DETR 配置.

3.3 对比实验

为证实本文所提方法的先进性, 本文选取数据集 SUN-RGBD 作为测试基准, 将其与其他前沿算法展开对比分析. 需要指出的是, 开放世界词汇下的 3D 检测属于前沿研究领域, 目前相关研究在文献中极为有限. 导致现有的研究成果难以直接与本文的方法进行横向对比. 基于此, 本文决定沿用基线方法 CoDA 的思路, 将当前最新的开放词汇点云分类方法移植至本文的实验设定中, 并对其性能进行全面评估. 具体操作上, 本文采用 3DETR 算法生成伪检测框, 并依次运用 PointCLIP、PointCLIPv2 以及 DetCLIP2 这三种方法开展开放词汇对象检测任务, 为便于区分与后续分析, 本文将这三种组合分别命名为 Det-PointCLIP、Det-PointCLIPv2 和 Det-CLIP2. 除此之外, 本文还设计了一个融合 2D 颜色图像输入的检测流程: 首先, 同样利用 3DETR 生成伪 3D 检测框; 随后, 借助相机内参矩阵将生成的 3D 框投影转换至 2D 平面, 得到对应的 2D 框; 接着, 利用这些 2D 框裁剪出图像中的相应区域; 最后, 采用 CLIP 模型对裁剪出的 2D 区域进行分类识别. 通过这一系列操作, 本文成功生成了 3D 检测结果, 并将其标记为 3D-CLIP. 在表中可以看到, 本文方法的 AP_{Novel} 和 AR_{Novel} 显著高于其他方法, 进一步证明了本文的设计在新物体定位和分类方面的优越性. 在表 2 中, 本文展示了与 SUN-RGBD 上的其他方法的比较. 在此次评估中, 本文将 SUN-RGBD 数据集中训练样本数量最多的前 10 个类别视为基础类别. 如表 2 所示, 在 AP 和 AR 评估指标方面, 本文的方法在新类别和基础类别上均优于其他方法.

表2 与其他先进方法对比实验

方法	输入	AP_{novel}	AP_{base}	$AP_{average}$	AR_{novel}	AR_{base}	$AR_{average}$
Det-PointCLIP[36]	P	0.09	5.04	1.17	21.98	65.03	31.33
Det-PointCLIPv2[37]	P	0.12	4.82	1.14	21.33	63.74	30.55
Det-CLIP^2[38]	P	0.88	22.74	5.63	22.21	65.04	31.52
3D-CLIP[8]	P+I	3.61	30.56	9.47	21.47	63.74	30.66
OV-Uni3DETR[35]	I	2.06	16.50	5.41	27.79	62.92	35.96
CoDA[4]	P	6.31	36.21	12.81	35.65	64.80	41.98
Our method	P+I	**6.90**	**37.01**	**13.44**	**36.87**	**66.93**	**43.40**

3.4 消融实验

为了验证基于多模态融合的空间目标检测技术方法中各个模块的有效性和必要性, 本文节以 SUN RGB-D 为基准数据集, 针对密集深度图引导的早期融合模块 (FMGD)、混合查询引导的室内 Transformer 检测器 (MQ-DETR) 和动态解耦 3D-IoU 的损失增强进行了消融实验. 具体如下:

3.4.1 各模块有效性

为了验证每个创新点对模型检测效果的影响, 在基线方法的基础上, 依次增加密集深度图引导的早期融合模块 (FMGD)、混合查询引导的室内 Transformer 检测器 (MQ-DETR)、动态解耦 3D-IoU 的损失增强 (DD-3DIoU). 表 3 是各个模块对检测的实验结果. 从表 3 中可以看出, 每个模块都能不同程度的增加检测性能.

实验结果表明, FMGD 模块的加入提高了 AP_{novel}, 与基线方法相比, FMGD 模块的 AP_{novel} 上升了 0.24%, FMGD 弥补了单模态对跨模态互补信息的缺失, 让网络对新物体类与整体目标的检测精度更鲁棒. 值得注意的是, FMGD 的加入导致 AP_{base} 出现了微小的下降. 深入分析表明, 这并非单纯的权重干扰, 而是源于跨模态对齐中的几何噪声

表3 各模块有效性实验

基线方法	FMGD	MQ-DETR	DD-3DIoU	AP_{novel}	AP_{base}	$AP_{average}$	AR_{novel}	AR_{base}	$AR_{avergae}$	FPS	Memory(GB)
√				6.31	36.21	12.81	35.65	64.80	41.98	12	10.2
√	√			6.55	36.12	12.98	36.05	64.92	42.32	10.4	10.6
√		√		6.51	36.85	13.10	36.28	65.74	42.68	11.6	10.3
√	√	√		6.74	36.84	13.28	36.75	66.65	43.25	9.8	10.8
√	√	√	√	6.90	37.01	13.44	36.87	66.93	43.40	9.8	10.8

与特征权衡. 虽然密集深度图实现了图像特征的空间映射, 但边缘感知深度补全过程不可避免地在物体边界处引入了细微的平滑噪声, 这在一定程度上削弱了基础类别对高频几何结构的回归精度. 此外, 早期融合迫使模型在特征提取阶段就在视觉语义泛化与点云几何精度之间寻求平衡, 这种对齐代价虽然在几何特征完备的基础类别上表现为精度的轻微回落, 却成功打破了单模态的几何过拟合, 为提升新物体的识别能力构建了至关重要的跨模态语义基础.

通过引入 MQ-DETR 模块, AP_{base}提升了 0.64%, AP_{novel}提升了 0.2%. 与基线模型相比, 引入混合查询机制让 Transformer 检测器的查询更具有全局性, 一部分查询通过训练学习通用目标的特征表达, 另一部分直接从点云中全局信息中获取使得训练查询充分利用基础类别的充足标注, 提升了基础类别的定位精度, 又通过几何查询增强了对全局几何结构的感知能力, 混合查询的精准性降低了基础类别的漏检率, 使得AR_{base}进一步提升.

当同时引入 FMGD、MQ-DETR、DD-3DIoU 三个模块时, 模型在AP_{novel}(6.90%)、AP_{base}(37.01%)、$AP_{average}$(13.44%) 指标上均取得最优结果. 这验证了三个模块在“特征融合 (FMGD)-查询优化 (MQ-DETR)- 损失优化 (DD-3DIoU)”层面的互补性: FMGD 奠定高质量特征基础, MQ-DETR 提升查询匹配精度, DD-3DIoU 优化框定位损失, 三者共同解决了开放词汇 3D 检测中特征鲁棒性不足、查询匹配偏差、框定位不准的核心问题, 最终显著提升了新物体类与基础类别的检测性能. 同时AR_{novel}的增长直接证明了混合查询与动态解耦策略在几何定位上的有效性, 成功召回了大量被基线遗漏的新物体候选框, 提升开放域探索能力.

此外, 在模型推理与资源消耗方面, 随着三个模块的逐层引入, 计算成本仅有轻微增加. 与基线方法相比, 模型的推理速度从 12 降至 9.8, 显存占用仅从 10.2 GB 微增至 10.8GB. 这表明本文提出的方法在AP_{novel}显著提升与AP_{base}的同时, 并未给硬件带来过重的负担, 在精度大幅跃升与计算开销之间取得了良好的平衡, 仍具备较好的实际部署潜力.

3.4.2 密集深度图引导的早期融合模块

为进一步分析密集深度图来源及 FMGD 内部结构设计对早期融合质量与最终检测性能的影响, 本文在消融实验中额外引入了两种基于图像的可学习深度预测方法作为对比, 即 Det-FastDepth 与 Det-DenseDepth. 如表 4 所示, 表中“Det-*”方法表示在保持整体检测框架不变的前提下, 仅使用对应的单目深度估计模型从 RGB 图像中预测密集深度图, 并将其作为 FMGD 模块的深度引导输入, 以替代原有的非学习型密集深度生成方式.

表4 密集深度图引导的早期融合模块消融实验

方法	类型	窗口大小	AP_{novel}	AP_{base}	$AP_{average}$	FLOPS (ms)
Det-FastDepth[39]	学习型	—	6.64	36.75	13.18	2.9
Det-DenseDepth[40]	学习型	—	6.78	36.47	13.23	3.3
FMGD (1层卷积模块)	非学习型	7×7	6.44	35.53	12.76	2.4
FMGD (逐元素相加)	非学习型	7×7	6.42	35.67	12.77	2.1
FMGD(ALL)	非学习型	3×3	6.42	36.65	12.99	2.3
FMGD(ALL)	非学习型	11×11	6.65	36.70	13.18	2.3
FMGD(ALL)	非学习型	7×7	**6.90**	**37.01**	**13.44**	**2.3**

从检测性能上看, 采用学习型深度预测方法后, Det-FastDepth 与 Det-DenseDepth 在AP_{novel}上分别达到 6.64% 和 6.78%, 仍低于完整 FMGD 的 6.90%. 同时在基础类别上也有下降, 这些结果表明, 尽管学习型深度预测能够提供一定的几何先验信息, 但其在复杂室内场景中受到遮挡、尺度歧义及预测误差累积的影响, 限制了跨模态特征对齐的稳定性.

进一步结合 FMGD 内部结构的消融结果可以观察到, 窗口大小对融合效果具有显著影响. 当采用较小窗口 (3×3) 时, FMGD 的AP_{novel}下降至 6.42%, 表明过小的感受野不足以覆盖室内场景中常见的结构性上下文信息; 而当窗口增大至 11×11 时, A_{novel}虽有所回升至 6.65%, 但整体性能仍不及 7×7 设置, 说明过大的窗口会引入冗余背景信息, 从而削弱对关键几何区域的聚焦能力. 相比之下, 7×7 窗口在局部几何细节建模与全局上下文感知之间取得了更为合理的平衡.

同时, 当将卷积对齐模块从原有的 3 层二维卷积降维模块简化为单层卷积映射或移除自适应门控机制并采用逐元素相加时, 尽管计算量得到减少, 但是AP_{novel}分别下降至 6.44% 和 6.42%, 这一趋势与学习型预测方法的性能退化表现高度一致, 进一步说明仅依赖深度图本身或简单的融合方式难以充分建模模态间的几何错位与信噪比差异, 深度引导信息需要与合适的窗口尺度、多层特征对齐及自适应门控机制协同作用, 才能充分发挥其价值.

3.4.3 混合查询引导的室内 Transformer 检测器

为了验证混合查询引导的室内 Transformer 在融合方面的有效性, 本文分别在基线上引入几何查询方法、混合查询方法, 并增加了自适应注意力层. 消融实验结果如表 5 所示. 为了比较网络结构对特征

提取能力的提升, 并排除第二阶段自训练策略的影响, 本节实验仅记录了模型在第一阶段 (1 080 epoch) 训练结束后的性能表现. 本文使用基于类无关蒸馏的 3D-CLIP 进行第一阶段训练, 将此方法作为基线.

表5　混合查询引导的室内 Transformer 消融实验 (基于第一阶段训练结果)

方法	GQ数量	LQ数量	AP_{novel}	AP_{base}	$AP_{average}$	AR_{novel}	AR_{base}	$AR_{avergae}$
基线	0	128	2.88	33.56	9.55	19.62	63.52	29.16
基线+几何查询	128	0	2.44	32.98	9.07	16.42	59.15	25.70
基线+混合查询(1: 1)	64	64	2.97	34.11	9.73	19.93	64.19	29.55
基线+混合查询(3: 1)	96	32	2.62	33.24	9.27	16.87	61.24	26.51
基线+混合查询(1: 3)	32	96	2.92	33.64	9.60	19.19	63.75	28.87
基线+自适应注意力层	0	128	3.01	33.89	9.72	19.05	63.24	28.66
基线+混合查询(1: 1)自适应注意层	**64**	**64**	**3.09**	**34.30**	**9.87**	**19.74**	**64.04**	**29.37**

实验结果表明如果只使用几何查询会导致无法聚焦于目标, 导致整体检测精度下降. 使用混合查询可以在聚焦目标的同时兼顾全局上下文信息, 能有效提升室内目标检测的性能. 引入混合查询机制后, 模型性能得到显著改善. 通过对比不同混合比例发现, 1:1 的混合比例达到了最佳平衡, 其AP_{novel}提升至 2.97%, AP_{base}提升至 34.11%. 当同时引入 1:1 混合查询和自适应注意力层时, 模型在各项评价指标上均达到最佳效果, 证明了自适应注意力层能进一步增强特征的全局语义交互, 提升了室内目标检测的整体性能.

3.4.4　动态解耦 3D-IoU 消融实验

为了验证混合查询引导的室内 Transformer 在融合方面的有效性, 本文分别为 FMGD 模块和 MQ-DETR 模块引入解耦 3D-IOU 和动态解耦 3D-IOU, 对比结果如表 6 所示. 实验结果表明, 在原有模块上使用解耦 3D-IoU 能有效提升新物体的发现能力; 通过解耦两个方向的梯度计算, 可以显著增强检测器对新发现物体的检测能力. 动态解耦 3D-IoU 通过在训练中根据物体的尺寸调整解耦权重, 可以更好的帮助适应室内场景中多尺度物体的检测, 提高检测准确率.

表6　动态解耦 3D-IoU 消融实验

方法	AP_{novel}	AP_{base}	$AP_{average}$	AR_{novel}	AR_{base}	$AR_{avergae}$
FMGD+MQ-DETR	6.74	36.84	13.28	36.75	66.65	43.25
FMGD+MQ-DETR+解耦3D-IoU	6.79	**37.11**	13.38	**37.01**	66.82	**43.49**
FMGD+MQ-DETR+动态解耦3D-IoU	**6.90**	37.01	**13.44**	36.87	**66.93**	43.40

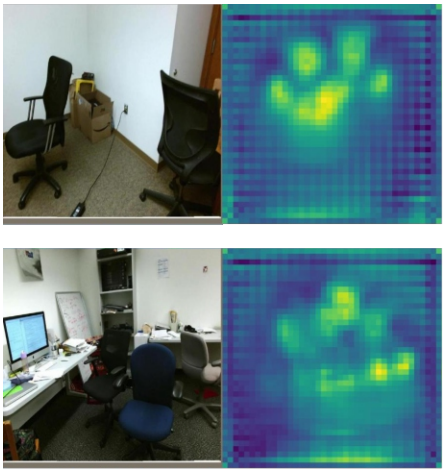
为了验证动态解耦 3D-IoU 模块中调节因子λ取值的合理性, 本文进行了参数敏感性实验. λ决定了权重分配偏离初始值 0.5 的幅度. 如表 7 所示, 本文在 [0.1, 0.5] 区间内对λ进行了测试. 实验结果表明, 模型对λ的变化表现出一定的鲁棒性, 但取值过大或过小均会影响检测性能. 当$\lambda = 0.1$时, 权重调整幅度较小, 难以充分应对极端尺度差异, 导致AP_{novel}仅为 6.74%; 当$\lambda = 0.5$时, 过激的权重分配导致训练波动, 使AP_{novel}下降至 6.64%. 当$\lambda = 0.3$时, 模型能够最有效地利用尺度先验平衡xy平面与z轴的解耦训练, 从而在AP_{novel}和AP_{base}上均取得最佳性能.

表7　参数敏感性分析

λ	AP_{novel}	AP_{base}	$AP_{average}$
0.1	6.74	36.84	13.28
0.2	6.82	36.93	13.36
0.3	**6.90**	**37.01**	**13.44**
0.4	6.86	36.84	13.37
0.5	6.64	36.54	13.36

4　实验结果可视化及分析

为了进一步验证本文提出的密集深度图引导的早期融合策略在特征层面的有效性, 本文对融合后的特征和补全后的深度图进行了可视化分析. 图 5 展示了在不同室内场景下的特征响应热力图.

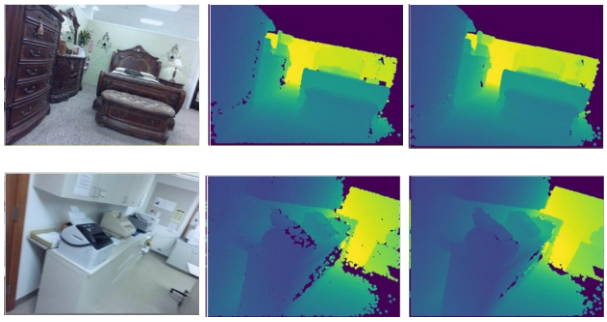


RGB 图像　　特征融合热力图

图5　多模态融合特征可视化结果

如图 5 所示, 左侧为原始 RGB 图像, 右侧为对应的融合特征热力图. 可视化结果表明, 经过密集深度图引导的特征融合后, 高响应区域 (图中亮黄色及绿色区域) 主要集中在椅子、桌子等前景目标物体上, 而地面、墙壁等背景区域的特征响应则受到明显抑制 (图中深蓝色区域). 这说明本文提出的融合模块有效地将图像中的丰富语义信息映射到了三维特征空间, 并实现了跨模态特征的精准对齐.

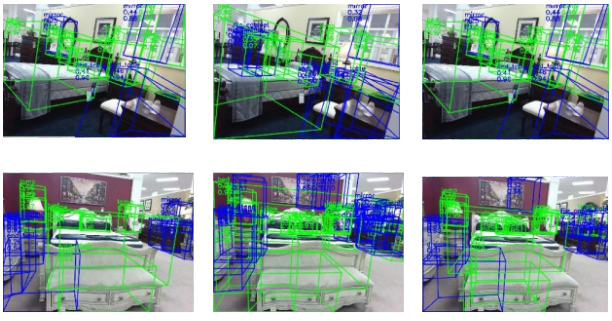
原始深度图中存在大量的信息空洞和不连续区域, 这会严重影响图像语义特征向三维空间的精准映射. 图 6 展示了经由本文提出的边缘感知深度补全算法处理后的结果. 通过对比可以明显观察到, 补全后的深度图有效填充了稀疏区域的深度值.



RGB 图像　　原始深度图　　补全后深度图

图6　深度补全可视化结果

为了直观验证各模块的有效性, 本文对消融实验结果进行了可视化对比, 如图 7 所示. 观察发现, 仅使用 FMGD 模块时, 模型能利用深度引导实现对场景主体的初精确定位; 引入 MQ-DETR 模块后, 得益于混合查询对全局信息的捕捉, 被遮挡物体和边缘小目标的检出率显著提高, 有效缓解了漏检问题; 而进一步加入 DD-3DIoU 策略后, 检测框的尺寸和方向与真实物体贴合度明显增强, 有效修正了定位偏差.



FMGD 可视化　MQ-DETR 可视化　DD-3DIOU 图可视化

图7　消融实验结果可视化

为了更直观地展示本文方法检测性能, 挑选了下面 6 个复杂场景下的检测结果进行可视化, 如图 8 所示, 并与基线网络进行对比. 实验表明, 基线网络在面对物体遮挡以及远距离目标表征稀疏问题时, 会出现不同程度的漏检误检; 而所提方法虽然在严重遮挡情况下也会出现漏检问题, 但总体表现更优, 证明了本文算法具有较好的检测性能. 然而, 该方法也会带来一定的误判, 这主要体现在模型对复杂背景的判别能力仍有不足.

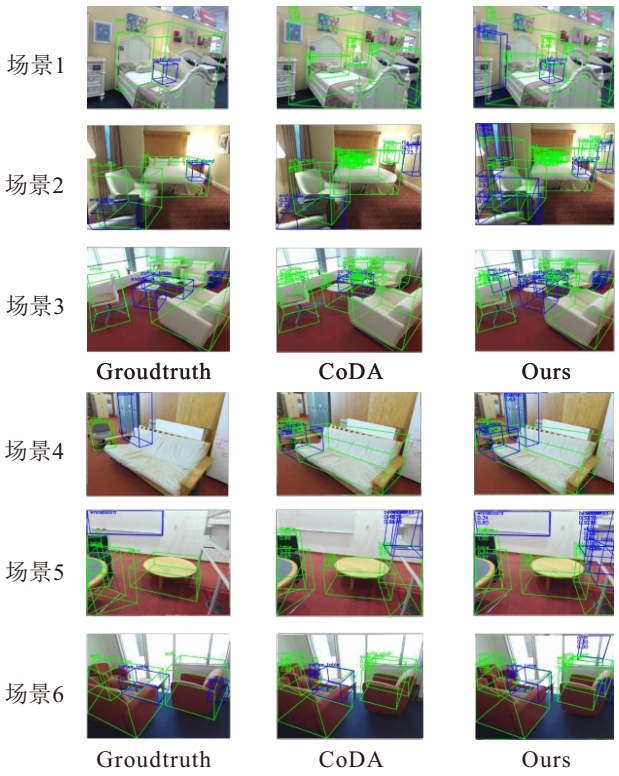


图8　实验结果可视化

5　结　论

本文提出了一种基于图像–点云–文本多模态融合的室内开放域三维目标检测方法. 通过引入密集深度图引导的早期融合策略, 实现了图像语义信息向三维空间的精确映射, 有效提升了点云特征的语义表达能力; 同时, 设计混合查询引导的室内 Transformer 检测器, 结合几何查询与可学习查询以及自适应注意力机制, 增强了对室内场景中目标几何结构与全局上下文信息的联合建模能力; 此外, 提出动态解耦 3D-IoU 损失增强策略, 通过解耦不同空间维度的梯度并引入尺度自适应权重机制, 显著提升了候选框定位精度与新物体发现能力. 实验结果表明, 所提出的方法在 SUN-RGBD 数据集上在新类别与基础类别检测性能方面均取得了稳定提升. 未来的工作将重点探索文本模态与视觉特征的深层融合机制. 本文将不再局限于仅利用文本进行后端的分类匹配, 而是计划在特征提取与编码阶段引入文本嵌入, 通过跨模态注意力机制实现语言先验信息对视觉特征定位与学习的早期引导. 旨在构建文本

引导视觉、视觉增强文本的互增强范式, 从而进一步提升模型在复杂、动态场景下的三维开放域检测性能与泛化能力.

参考文献 (References)

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

[2] 孙先涛, 闻勇, 陈文杰, 等. 基于语义分割与旋转目标检测的机器人抓取位姿估计[J]. 控制与决策, 2024, 39(9): 2913-2922.
(Sun X T, Wen Y, Chen W J, et al. Robot grasping pose estimation based on semantic segmentation and rotating target detection[J]. Control and Decision, 2024, 39(9): 2913-2922.)

[3] 周栋, 孙光辉, 吴立刚. 面向空间视觉目标检测的对抗攻击与防御算法[J]. 控制与决策, 2024, 39(7): 2161-2168.
(Zhou D, Sun G H, Wu L G. Adversarial attack and defense algorithms towards space visual object detection[J]. Control and Decision, 2024, 39(7): 2161-2168.)

[4] Cao Y, Zeng Y H, Xu H, et al. CoDA: Collaborative novel box discovery and cross-modal alignment for open-vocabulary 3D object detection[J/OL]. 2023, arXiv: 2310.02960.

[5] Jiao P K, Zhao N, Chen J J, et al. Unlocking textual and visual wisdom: Open-vocabulary 3D object detection enhanced by comprehensive guidance fromtext and image[C]. Computer Vision – ECCV 2024. Cham: Springer, 2025: 376-392.

[6] Huang R, Zheng H, Wang Y, et al. Training open-vocabulary monocular 3D detection model without 3D data[J]. Advances in Neural Information Processing Systems, 2024, 37: 72145-72169.

[7] Ju Y L, Yang T M, Yi L. ImOV3D: Learning open vocabulary point clouds 3D object detection from only 2D images[J]. Advances in Neural Information Processing Systems, 2024, 37: 141261-141291.

[8] Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision[J/OL]. 2021, arXiv: 2103.00020.

[9] Lu Y H, Xu C F, Wei X B, et al. Open-vocabulary point-cloud object detection without 3D annotation[C]. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, 2023: 1190-1199.

[10] Qi C R, Litany O, He K M, et al. Deep hough voting for 3D object detection in point clouds[C]. IEEE/CVF International Conference on Computer Vision. Seoul, 2019: 9276-9285.

[11] Cheng B W, Sheng L, Shi S S, et al. Back-tracing representative points for voting-based 3D object detection in point clouds[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, 2021: 8959-8968.

[12] Zhang Z W, Sun B, Yang H T, et al. H3DNet: 3D object detection using hybrid geometric primitives[C]. Computer Vision – ECCV 2020. Cham: Springer, 2020: 311-329.

[13] Wang H Y, Shi S S, Yang Z, et al. RBGNet: Ray-based grouping for 3D object detection[C]. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, 2022: 1100-1109.

[14] Zhou Y, Tuzel O. VoxelNet: End-to-end learning for point cloud based 3D object detection[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, 2018: 4490-4499.

[15] Rukhovich D, Vorontsova A, Konushin A. FCAF3D: Fully convolutional anchor-free 3D object detection[C]. Computer Vision – ECCV 2022. Cham: Springer, 2022: 477-493.

[16] Liu Z, Zhang Z, Cao Y, et al. Group-free 3D object detection via transformers[C]. IEEE/CVF International Conference on Computer Vision. Montreal, 2022: 2929-2938.

[17] Misra I, Girdhar R, Joulin A. An end-to-end transformer model for 3D object detection[C]. IEEE/CVF International Conference on Computer Vision. Montreal, 2022: 2886-2897.

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

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

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

[21] 于明, 邢章浩, 刘依. 基于非对称跨模态融合的 RGB-D 显著目标检测[J]. 控制与决策, 2023, 38(9): 2487-2495.
(Yu M, Xing Z H, Liu Y. RGB-D salient object detection with asymmetric cross-modal fusion[J]. Control and Decision, 2023, 38(9): 2487-2495.)

[22] Zhang Z H, Shen Y X, Li H, et al. MAFF-net: Filter false positive for 3D vehicle detection with multi-modal adaptive feature fusion[C]. IEEE 25th International Conference on Intelligent Transportation Systems. Macau, 2022: 369-376.

[23] Tan X, Chen X Y, Zhang G W, et al. MBDF-net: Multi-branch deep fusion network for 3D object detection[J/OL]. 2021, arXiv: 2108.12863.

[24] Wang Z X, Jia K. Frustum ConvNet: Sliding Frustums to aggregate local point-wise features for amodal 3D object detection[C]. IEEE/RSJ International Conference on Intelligent Robots and Systems. Macau, 2020: 1742-

1749.

[25] 佟国峰, 杨宇航, 彭浩, 等. 基于视觉语义与激光点云交融构建的 SLAM 算法[J]. 控制与决策, 2024, 39(1): 103-111.
(Tong G F, Yang Y H, Peng H, et al. SLAM algorithm based on fusion of visual semantics and laser point cloud[J]. Control and Decision, 2024, 39(1): 103-111.)

[26] Kuo W C, Cui Y, Gu X Y, et al. F-VLM: Open-vocabulary object detection upon frozen vision and language models[J/OL]. 2022, arXiv: 2209.15639.

[27] Wu X S, Zhu F, Zhao R, et al. CORA: Adapting CLIP for open-vocabulary detection with region prompting and anchor pre-matching[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, 2023: 7031-7040.

[28] Li R, Zhang D W, Wang Y C, et al. Open-vocabulary multi-object tracking with domain generalized and temporally adaptive features[J]. IEEE Transactions on Multimedia, 2025, 27: 3009-3022.

[29] Qorbani R, Villani G, Panagiotakopoulos T, et al. Semantic library adaptation: LoRA retrieval and fusion for open-vocabulary semantic segmentation[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, 2025: 9804-9815.

[30] Gupta A, Narayan S, Joseph K J, et al. OW-DETR: Open-world detection transformer[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, 2022: 9225-9234.

[31] Zhao Y J, Lin J Y, Lau R W H. Hierarchical cross-modal alignment for open-vocabulary 3D object detection[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2025, 39(10): 10501-10509.

[32] Zhong Y W, Yang J W, Zhang P C, et al. RegionCLIP: Region-based language-image pretraining[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, 2022: 16772-16782.

[33] Cheng T H, Song L, Ge Y X, et al. YOLO-world: Real-time open-vocabulary object detection[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, 2024: 16901-16911.

[34] Liu S L, Zeng Z Y, Ren T H, et al. Grounding DINO: Marrying DINO with grounded pre-training forOpen-set object detection[C]. Computer Vision – ECCV 2024. Cham: Springer, 2025: 38-55.

[35] Wang Z Y, Li Y L, Liu T C, et al. OV-Uni3DETR: Towards unified open-vocabulary 3D object detection via cycle-modality propagation[C]. Computer Vision – ECCV 2024. Cham: Springer, 2025: 73-89.

[36] Zhang R R, Guo Z Y, Zhang W, et al. PointCLIP: Point cloud understanding by CLIP[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, 2022: 8542-8552.

[37] Zhu X Y, Zhang R R, He B W, et al. PointCLIP V2: Prompting CLIP and GPT for powerful 3D open-world learning[C]. IEEE/CVF International Conference on Computer Vision. Paris, 2024: 2639-2650.

[38] Zeng Y H, Jiang C H, Mao J G, et al. CLIP2: Contrastive language-image-point pretraining from real-world point cloud data[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, 2023: 15244-15253.

[39] Wofk D, Ma F C, Yang T J, et al. Fastdepth: Fast monocular depth estimation on embedded systems[C]. International Conference on Robotics and Automation. Montreal, 2019: 6101-6108.

[40] Alhashim I, Wonka P. High quality monocular depth estimation via transfer learning[J/OL]. 2018, arXiv: 1812.11941.

作者简介

樊怡麟 (2002−), 男, 硕士研究生, 主要研究方向为计算机视觉、目标检测, E-mail: fylqust@163.com;

季雨昂 (2002−), 男, 硕士研究生, 主要研究方向为计算机视觉、多目标跟踪, E-mail: 18764138278@163.com;

秦修功 (1990−), 男, 高级工程师, 硕士, 主要研究方向为智慧空间机器人, E-mail: 13121990213@163.com;

杨方正 (2003−), 男, 硕士研究生, 主要研究方向为计算机视觉、多目标跟踪, E-mail: 2869195579@qq.com;

李辉 (1984−), 男, 副教授, 博士, 从事计算机视觉、行为 识别等研究, E-mail: lihui@qust.edu.cn.