文章编号:1671-6833(2025)05-0035-08

融合 CLIP 和 3D 高斯的多模态场景编辑算法

曹仰杰, 王伟平, 李振强, 谢 俊, 吕润峰

(郑州大学 网络空间安全学院,河南 郑州 450002)

摘 要:针对 3D 场景编辑算法对标注数据过度依赖和计算复杂度高的问题,提出了一种融合 CLIP 与 3D 高斯的多模态场景编辑算法(CLIP2Gaussian)。首先,利用 SAM 从多视角图像中提取目标掩码,并引入双向传播策略实现不同视角之间的掩码一致性;其次,将提取的掩码通过 CLIP 进行语义标签分配,并映射到 3D 高斯点,实现 3D 场景的语义嵌入;最后,采用可微分渲染机制对 3D 高斯参数进行优化,同时引入空间一致性正则化策略,通过聚类增强语义标签在 3D 空间中的一致性与稳定性。实验结果表明:CLIP2Gaussian 在 LERF 数据集上 IoU 达到 61.23%,语义分割任务中单次文本查询响应时间为 0.57 s,准确率和效率均优于 LERF。消融实验进一步验证了所提算法在最小扰动原始场景的前提下对目标区域的精准编辑。

关键词:3D 重建; 零样本学习; 场景理解; 场景编辑; 3D 高斯

中图分类号:TP391;TP751.1 **文献标志码**:A **doi**:10.13705/j.issn.1671-6833.2025.05.016

3D 场景编辑技术是机器与现实世界交互的核心技术,被广泛应用于虚拟现实、智能机器人和自动驾驶等领域。然而,由于现实场景中环境复杂,使得场景编辑算法面临诸多挑战,包括计算复杂度高和场景解耦困难等问题,尤其是在处理大规模场景时往往需要大量的计算资源。此外,3D 标注数据的稀缺性也限制了基于监督学习的场景编辑算法的准确性与适用性。因此,如何实现高效的交互式编辑并提升模型在开放世界下的泛化能力是当前研究的核心问题。

目前,基于点云的 3D 场景编辑方法虽在某些特定任务中具有良好表现[1-2],但过度依赖人工标注数据,难以满足零样本识别任务的需求,且训练成本较高。此外,点云与图像之间在语义层面缺乏有效融合,导致语义信息难以反馈至几何表示,限制了此类方法在复杂编辑场景中的应用能力。为克服上述缺点,NeRF[3] 采用隐式 MLP 对 3D 场景进行建模,实现高质量的重建效果。LERF[4] 在 NeRF 的基础上引入了语义嵌入机制,以增强对场景的语义理解能力,但仍存在训练时间长、渲染效率低等问题。

高斯溅射(Gaussian splatting, GS)[5] 作为一种高效的 3D 场景表示方式,通过各向异性高斯分布对辐射场进行建模,在保障渲染质量的同时显著提升了渲染效率,为实现实时、高效的场景编辑提供了新思路。已有研究尝试将 GS 应用于场景编辑任务,例如 Gaussian Grouping[6] 结合分割模型(segment anything model, SAM)[7] 对高斯点进行实例级分类与编辑,然而,这类方法仍局限于封闭词汇空间,缺乏对开放世界语义的理解能力,难以满足多样化语义编辑的实际需求。

随着多模态大模型的快速发展,3D 场景编辑技术迎来了新的发展机遇,尤其是对比语言-图像预训练(contrastive language-image pre-training, CLIP)[8] 模型的提出,实现了视觉与语言之间的多模态语义对齐,显著增强了模型对开放世界的理解能力,为缓解现有方法在标注依赖和泛化能力方面的限制提供了新的解决思路。

综上所述,本文提出了 CLIP2Gaussian 算法,该算法将 3D 高斯的高效渲染特性与 CLIP 模型的多模态语义对齐能力相结合,构建了一个支持开放词汇的语义理解框架并有效缓解了对人工标注数据的依赖。CLIP2Gaussian 算法通过语义增强的 3D 高斯

收稿日期:2025-03-07;**修订日期**:2025-04-13

基金项目:国家自然科学基金资助项目(62302458);郑州市协同创新重大专项(20XTZX06013)

作者简介:曹仰杰(1976—),男,河南濮阳人,郑州大学教授,博士,博士生导师,主要从事机器智能与人机交互、大数据智能处理、云计算与高性能计算研究,E-mail:caoyj@zzu.edu.cn。

引用本文:曹仰杰,王伟平,李振强,等. 融合 CLIP 和 3D 高斯的多模态场景编辑算法[J]. 郑州大学学报(工学版),2025,46(5):39-42. (CAO Y J, WANG W P, LI Z Q, et al. Multimodal Scene Editing Algorithm Integrating CLIP and 3D Gaussian[J]. Journal of Zhengzhou University(Engineering Science), 2025,46(5):39-42.)

表示对 3D 场景进行建模,并为其中的实例对象分配语义标签。该方法不仅实现了对场景的精准建模,还能通过语义查询识别特定目标,大幅提升了模型在 3D 开放环境下的理解能力,同时为后续的场景编辑任务提供了有力支持。

1 相关工作

1.1 零样本学习

零样本学习(zero-shot learning,ZSL)旨在识别训练集中未出现的实例对象,而传统 3D 场景编辑方法依赖大量 3D 标注数据,因而在识别“未见过”对象时表现有限。目前,ZSL 研究主要集中在 2D 图像领域,针对 3D 场景理解与编辑的探索仍较少[9]。近年来,部分研究尝试将 CLIP 模型的零样本学习能力应用到至 3D 视觉任务中。例如,PointCLIP[10]通过将点云投影至 2D 图像空间并与文本特征对齐,实现了零样本识别;MaskCLIP[11]则利用 CLIP 辅助 2D 密集预测,表现出良好的泛化能力;LERF[4]通过语言引导的特征嵌入,提升了 NeRF 在开放词汇场景中的语义理解能力,推进了 3D 开放世界的研究进展。

1.2 高斯溅射 GS

GS 是一种新兴的 3D 场景重建方法,通过结合 3D 高斯分布与 Splatting 技术,实现高效的实时渲染。后续工作如 Gaussian Grouping[6]和 FlashSplat[12]通过引入 SAM 实现语义分割,从而支持 3D 高斯的语义追踪与精准编辑。尽管 SAM 在空间分割方面表现出色,其语义理解能力仍然有限。相比之下,CLIP 通过视觉-语言对齐展现出更强的语义感知能力[13]。

1.3 基于辐射场的场景编辑

3D 场景编辑是 3D 视觉技术走向实际应用的关键,尽管基于辐射场的编辑方式具有效果,但仍面临诸多挑战。现有 NeRF 方法如 Blended-NeRF[14]依赖用户手动选区,适用性受限;Instruct-NeRF2NeRF[15]则通过文本控制编辑,依赖 2D 结果,易产生全局误差。相比之下,3D 高斯表示具备更强的空间可控性与表达能力,为实现高精度、高灵活性的场景编辑提供了更优方案。

2 本文方法

本文提出的 CLIP2Gaussian 算法充分利用了 CLIP[8]的文本编码能力,并通过可微分渲染将语义信息嵌入 3D 场景,实现了对实例对象的精准识别,为后续场景编辑提供支持。CLIP2Gaussian 模型架构如图 1 所示,包含以下 3 个关键步骤:

步骤 1 利用 SAM 从多视角图像中提取目标掩码,并引入双向传播策略以确保不同视角下的掩码一致性。

步骤 2 使用 CLIP 对掩码进行语义标签分配,并将语义信息映射到 3D 高斯点,实现 3D 语义嵌入。

步骤 3 借助可微分渲染优化高斯属性,并通过聚类执行空间一致性正则化实现语义追踪。

本节将详细介绍 CLIP2Gaussian 的设计关键。为实现细粒度语义理解,模型在保留高斯原有属性的基础上,向每个高斯点嵌入一个可学习的语义特征,从而提升其在开放场景中的理解与编辑能力。

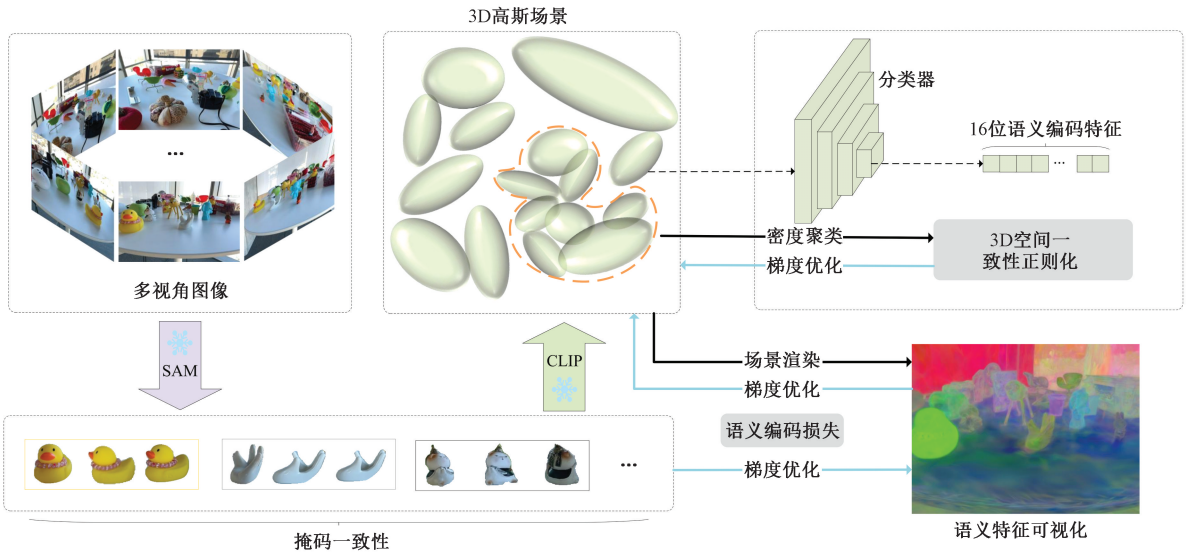

图 1 CLIP2Gaussian 模型架构

Figure 1 CLIP2Gaussian model architecture

2.1　3D 场景表示

本文基于 GS[5] 使用各向异性的 3D 高斯分布表示 3D 场景,其采用基于点的渲染技术(α-混合)将 3D 场景渲染到 2D 平面上,3D 高斯场景表示如式(1)所示:

$$G(\boldsymbol{x}) = \exp\left(-\frac{1}{2}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{A}^{-1}\boldsymbol{x}\right)。\tag{1}$$

式中:$\boldsymbol{x}$ 表示高斯点中心的空间信息($\boldsymbol{x} \in \mathbf{R}^3$),同时也是高斯平均值;$\boldsymbol{A}$ 为协方差矩阵,用来确定高斯点的形状,可以分解为旋转矩阵 $\boldsymbol{R}$ 和缩放矩阵 $\boldsymbol{S}$,用于可微优化,在优化过程中可以通过改变两个矩阵进而调整高斯椭球的形状、大小、方向。$\boldsymbol{A}$ 如式(2)所示:

$$\boldsymbol{A} = \boldsymbol{RSS}^{\mathrm{T}}\boldsymbol{R}^{\mathrm{T}}。\tag{2}$$

每一个高斯点都由一组属性表征:$g_i = \{\boldsymbol{x}, \boldsymbol{A}, c, \alpha\}$,其中,$c$ 为 RGB 颜色($c \in \mathbf{R}^3$);α 表示不透明度。在渲染时,通过将高斯点投影到平面上,其扩散痕迹由不透明度叠加形成。3D 高斯在具有场景表达能力的同时,还具备可微分和显式支持快速渲染的特点。N 个有序点重叠成一个像素点的颜色值 C 如式(3)所示:

$$C = \sum_{i \in N} c_i \alpha_i \prod_{j=1}^{i-1}(1 - \alpha_j)。\tag{3}$$

式中:c_i 和 α_i 分别表示给定高斯点的颜色值和密度,由高斯函数确定。

2.2　多视角图像和掩码关联

为提升模型在开放世界的理解能力并减少对 3D 标注的依赖,CLIP2Gaussian 采用 SAM 从多视角图像中提取 2D 掩码,从而实现实例对象的初步分割。然而,SAM 自动生成的掩码存在跨视角一致性差、分割杂乱等问题,难以准确统计实例数量。为此,本文引入具备零样本分割能力的 DEVA[16],对 SAM 提取的掩码进行跨视角关联与优化,生成更清晰、更稳定的实例分割结果,有效提升了训练数据质量和模型的鲁棒性。图 2 为 SAM 提取掩码和本文方法对比。

(a) 真实图像　(b) SAM提取掩码　(c) 本文方法

图 2　SAM 提取掩码和本文方法对比

Figure 2　Comparison between SAM extract mask and the proposed method

2.3　语义特征分配与嵌入

SAM 在分割图像方面展示了强大的泛化能力,但其分割结果缺乏语义识别能力。相比之下,CLIP 作为一种基于对比文本-图像学习的多模态模型,通过训练数十亿个文本-图像对,获得了强大的视觉理解能力,提供了广泛的文本与图像对齐的潜在空间。因此,本文通过引入 CLIP,可以更好地解决 SAM 在语义识别方面的不足。

2.3.1　语义特征分配

为了在场景视图中生成一致的语义编码特征,本文引入基于 CLIP 文本编码器的语义一致性正则化机制。在 3D 场景中,除原有高斯特征外,为每个高斯点嵌入一个长度为 16 的可学习语义向量,用于区分不同实例对象。通过 CLIP 模型提取已关联掩码图像中实例的视觉特征,并将其与文本标签共同嵌入统一语义空间。通过计算文本-图像对的相似度,选取匹配度最高的标签作为实例的语义分类结果,从而增强语义一致性并提高识别精度。

$$\hat{\boldsymbol{T}}(i) = \max(\boldsymbol{v}_i \cdot \boldsymbol{t}_i)。\tag{4}$$

式中:$\boldsymbol{v}_i$ 表示实例对象图像特征;$\boldsymbol{t}_i$ 表示文本特征;$\hat{\boldsymbol{T}}(i)$ 表示一组文本标签中相似度最高的文本标签,为后续语义嵌入工作提供支持。

2.3.2　基于可微渲染的语义嵌入

本文采用可微分方式将嵌入语义编码的 3D 高斯渲染为 2D 图像,借助 GS 中的可微渲染器,实现与球谐函数颜色优化相似的端到端可训练渲染过程。每个高斯点都可以在 2D 图像中评估其影响权重 α',通过混合与像素重叠的 N 个有序高斯分布,计算其对单个像素的累积贡献。

$$\boldsymbol{E}_{\mathrm{t}} = \sum_{i \in N} \boldsymbol{e}_i \alpha'_i \prod_{j=1}^{i-1}(1 - \alpha'_j)。\tag{5}$$

其中,每个像素最终渲染的 2D 掩码语义特征 $\boldsymbol{E}_{\mathrm{t}}$ 是每个高斯的长度为 16 的语义编码 $\boldsymbol{e}_i$ 的加权和,这一过程依赖于每个高斯点对该像素的影响权重 α'_i,通过计算协方差矩阵 $\boldsymbol{A}_{2\mathrm{D}}$ 的 2D 高斯函数与每个点的不透明度 α_i 的乘积来计算 α'_i。

$$\boldsymbol{A}_{2\mathrm{D}} = \boldsymbol{JWA}_{3\mathrm{D}}\boldsymbol{W}^{\mathrm{T}}\boldsymbol{J}^{\mathrm{T}}。\tag{6}$$

式中:$\boldsymbol{A}_{3\mathrm{D}}$ 为三维协方差矩阵;$\boldsymbol{A}_{2\mathrm{D}}$ 为经过渲染后的 2D 协方差矩阵;$\boldsymbol{J}$ 为 3D 到 2D 投影仿射近似的雅可比矩阵;$\boldsymbol{W}$ 为世界坐标到相机坐标的变换矩阵。

2.3.3　可微渲染实现语义嵌入

对图像掩码进行关联后,使用 CLIP 获取实例标签,假设场景中有 K 个实例对象。为了区分不同实例对象的语义编码,本文设计了语义编码损失 L_{t}

用于更新语义编码,包含以下两个部分:

(1)2D 语义编码损失。实例标签从 2D 图像中提取,不能直接用于监督 3D 高斯的语义编码 $\boldsymbol{e}_i$,因此,将式(5)中渲染的 2D 语义编码 $\boldsymbol{E}_t$ 作为输入,定义一个具有多层卷积的分类器,将其特征维度提升到 $K+1$ 维,然后对 $\mathrm{softmax}(f(\boldsymbol{E}_t))$ 进行标签分类,其中 K 为 3D 场景中实例对象的数量。采用标准交叉熵损失 L_{2D} 进行分类。

(2)3D 空间一致性损失。为了提高实例对象分类的精度,除了使用 2D 监督的标准交叉熵损失外,本文还引入了 3D 空间一致性正则化损失来约束语义编码。通过密度对高斯点进行聚类分析,识别出密度较高的区域,并将这些区域中的点划分为同一类实例对象。利用 3D 空间一致性,可以更有效地在基于点的 α-混合期间监督 3D 对象内部的 3D 高斯。在式(7)中,F 表示为线性层之后组合的 softmax 运算,在计算语义编码损失时使用,将 m 个采样点的 KL 散度损失表示为

$$L_{3D}=\frac{1}{m}\sum_{j=1}^{m}D_{KL}(P\parallel Q)=\frac{1}{m}\sum_{j=1}^{m}\frac{1}{|Q_j|}\cdot\sum_{e_i\in Q_j}F(\boldsymbol{e}_j)\log\frac{F(\boldsymbol{e}_j)}{F(\boldsymbol{e}_i)}。\tag{7}$$

式中:P 包含一个 3D 高斯的采样语义编码 $\boldsymbol{e}_i$;集合 $Q=\{\boldsymbol{e}'_1,\boldsymbol{e}'_2,\cdots,\boldsymbol{e}'_k\}$ 为根据密度进行聚类分析的聚类集合。鉴于开放场景中实例对象形态的复杂性,本文采用基于密度的 DBSCAN 聚类算法[17],该算法可识别任意形状簇并有效处理噪声点,基于此,本文引入的空间一致性损失聚类结果如图 3 所示。

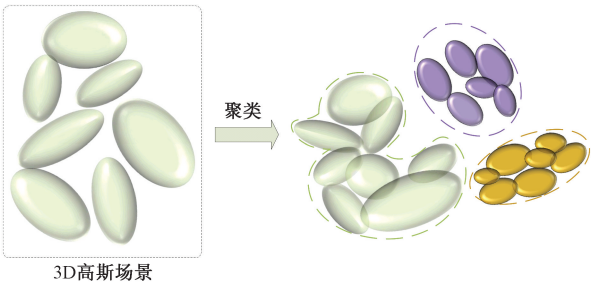

图 3 空间一致性损失聚类结果

Figure 3 Clustering results of spatial consistency loss

结合 3D 高斯原有的图像渲染损失 L_{ori},将端到端的整体损失表示为

$$L_{render}=L_{ori}+L_t=L_{ori}+\lambda_{2D}L_{2D}+\lambda_{3D}L_{3D}。\tag{8}$$

2.4 场景编辑

算法 1 CLIP2Gaussian 算法。

输入:数据集 *datasets*,场景中可分割的语义类别总数 *num_classes*,迭代次数 *iter*,*lambda* 正则化权重系数等;

输出:包含语义特征的 3D 点云,场景渲染的 2D 图像。

① 初始化 3D 点云和设备信息:$p\leftarrow$3D 点云和位姿信息;

② 提取多视角图像:$m\leftarrow(m_1,m_2,\cdots,m_K)$,通过 SAM 提取多视角图像的掩码;

③ 掩码关联:$(\hat{M}_1,\hat{M}_2,\cdots,\hat{M}_K)\leftarrow\mathrm{DEVA}(m)$,使用预训练的零样本分割追踪器 DEVA 在多视图间传播和关联掩码;

④ 提取文本信息:$t\leftarrow(t_1,t_2,\cdots,t_K)$,使用 CLIP 提取实例对象语义信息;

⑤ 初始化场景信息:s、a、c、$t\leftarrow$ 初始场景(协方差、不透明度、颜色、语义标签);

⑥ 迭代优化:

对不同相机视角进行采样训练:$V,I,\hat{M}\leftarrow\mathrm{SampleTrainingView}(i)$;

对图像和语义编码进行光栅化渲染:$I,E_t\leftarrow\mathrm{Rasterize}(p,s,a,c,t,V)$;

执行优化:$L_{render}\leftarrow L_{ori}+L_t$;

更新模型参数:$p,s,a,c,t\leftarrow\mathrm{Adam}(\nabla L)$;

⑦ 场景训练完成。

经过 3D 高斯训练和语义特征嵌入后,场景可以通过一组包含不同语义编码的 3D 高斯进行表示。针对编辑任务,仅调整与目标相关的 3D 高斯点即可完成编辑。对于 3D 物体的添加和删除,通过文本语义识别相关的高斯点进行直接添加或删除,无须参数优化。编辑后为确保场景平滑,删除相关的 3D 高斯点并通过 LAMA[18] 的 2D 绘制结果进行修复。本文提出的高斯编辑方法采用细粒度掩码建模,确保多个实例的编辑互不干扰,且无须重新训练整个 3D 场景,从而显著提高了编辑效率。

3 实验结果与分析

3.1 数据集

为了评估模型在开放世界中的理解能力和对象实例定位精度,本文使用 LERF[4] 数据集进行测试,该数据集包含 3 个场景,每个场景平均提供 7.7 个文本查询及对应掩码标签。通过将每个场景的文本查询结果与精确掩码进行比较,使用边界度量 mBIoU 评估分割质量。为评估文本语义嵌入对 3D 场景重建质量的影响,本文在 Mip-NeRF 360[19] 数据集的 7 组场景中进行了测试。

在 3D 场景对象删除任务中,本文根据用户提供的文本语义进行语义追踪,识别场景中与目标语

义编码相匹配的 3D 高斯分布,并通过凸包算法提取对应的高斯值。对于对象修复,在删除区域附近初始化新的高斯函数,并通过图像掩码进行微调以恢复场景。

3.2　评价指标与参数设置

本文使用 5 个广泛流行的评价指标综合评估模型的性能。

(1)mIoU:表示模型在各类别上的分割结果的平均交并比,是对所有类别的 IoU 求平均。

(2)mBIoU:表示平均边界交并比,用于评估模型在边界区域的分割质量。

计算公式如下:

$$\begin{cases} \mathrm{mIoU} = \dfrac{1}{C}\sum_{c=1}^{C}\dfrac{|\mathrm{Pred}(c)\cap \mathrm{GT}(c)|}{|\mathrm{Pred}(c)\cup \mathrm{GT}(c)|}; \\ \mathrm{mBIoU} = \dfrac{1}{C}\sum_{c=1}^{C}\dfrac{|\mathrm{Pred}_{\mathrm{boundary}}(c)\cap \mathrm{GT}_{\mathrm{boundary}}(c)|}{|\mathrm{Pred}_{\mathrm{boundary}}(c)\cup \mathrm{GT}_{\mathrm{boundary}}(c)|}。 \end{cases} \tag{9}$$

式中:$\mathrm{Pred}(c)$ 和 $\mathrm{GT}(c)$ 分别为类别 c 的预测区域和真实区域;$\mathrm{Pred}_{\mathrm{boundary}}(c)$ 和 $\mathrm{GT}_{\mathrm{boundary}}(c)$ 分别为类别 c 的预测边界区域和真实边界区域;C 为总类别数。

(3)*PSNR*[20]:表示像素级误差,通过计算语义嵌入后图像与原始图像之间的均方误差来衡量图像质量,其值越大表示图像质量越好。

$$PSNR = 10 \cdot \log \frac{MAX_{\mathrm{I}}^2}{MSE}。 \tag{10}$$

式中:MAX_{I} 为图像最大像素;MSE 表示原始图像和渲染图像的误差。

(4)*SSIM*[21]:表示结构相似性指数,将图像的亮度、对比度和结构的相似度组合成一个总的相似度指数来衡量图像质量,其值越大表示图像质量越好。

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}。 \tag{11}$$

式中:μ_x 为均值;σ_x^2 为方差;σ_{xy} 为协方差。

(5)*LPIPS*[22]:表示学习感知图像块相似性,用于比较两幅图像之间的感知相似度,通过计算图像特征空间中的差异来衡量图像质量。

$$LPIPS(x,y) = \sum_l \frac{1}{\boldsymbol{H}_l \boldsymbol{W}_l}\sum_{h,w} \| \hat{y}_l^{hw} - \hat{x}_l^{hw} \|_2^2。 \tag{12}$$

式中:$\hat{x}_l^{hw}$ 为第 l 层特征图的特征值;$\boldsymbol{H}_l$ 和 $\boldsymbol{W}_l$ 为 l 层特征图的高和宽。

本文在训练模型时,高斯层和线性层均使用 Adam 优化器进行迭代优化,语义编码的学习率设定为 0.000 5。对于 3D 正则化损失,选择样本总数为 1 000,邻域半径为 0.5,最小采样点数为 4 进行聚类。所有数据均在 NVIDIA A40 GPU 上进行了 30 000 次迭代训练。

3.3　多视图重建和语义分割

为探究跨视图掩码一致性对模型训练性能的影响,本文在 LERF[4] 数据集的 figurines 场景上进行了消融实验。初始实验采用 SAM 提取 2D 掩码,然而其输出掩码普遍存在结构混乱、语义缺乏一致性等问题,且实例识别主要依赖掩码面积,难以精确定位目标,导致训练与推理阶段性能显著下降。为验证模型在实际应用中的效率,本文进一步评估了模型对开放词汇的查询速度,实验结果如表 1 所示。

表 1　在 figurines 场景下掩码相关性消融实验

Table 1　Mask correlation ablation experiment under the scenarios of figures

SAM[7]	GS[5]	掩码一致性	IoU/%	查询时间/s
√			46.74	7.77
√	√		OOM	OOM
√	√	√	**61.23**	**0.57**

实验结果表明,若仅使用 SAM 进行掩码提取,单次文本查询平均耗时达 7.77 s。而采用本文提出的基于 SAM 的跨视图掩码关联策略后,IoU 提升至 61.23%,有效缓解了多视角一致性不足的问题,显著提升了 3D 语义特征的准确性。直接在 3D 高斯上对 CLIP 特征进行显式建模会造成显存溢出(out of memory,OOM),本文将语义特征压缩为 16 维(长度为 16 的一串编码)可学习向量并通过语义分类器进行预测,不仅解决了显存瓶颈问题,还进一步提升了分割精度。最终,CLIP2Gaussian 在保持更高分割准确率的同时,文本查询速度显著提高,验证了跨视图掩码一致性机制在复杂 3D 场景语义理解中的重要作用。

在定量评估中,基于 LERF 数据集在多个 3D 场景中将模型预测结果与真实掩码进行对比,结果如表 2 所示。CLIP2Gaussian 在多个场景中均取得最佳效果。上述结果充分验证了 CLIP2Gaussian 在开放词汇语义分割中的优势,特别是在复杂 3D 场景中的实例识别与边界定位方面表现更加出色。本文在 3D 场景中对比了现有开放词汇分割方法 LSeg[23] 与所提出的 CLIP2Gaussian,并在图 4 中展示了可视化结果。实验表明,CLIP2Gaussian 在掩码预测方面具有更高的精度,边界更清晰,分割效果更准确。为全面评估模型性能,图 4 还采用主成分分析对语义编码特征进行可视化,直观展示了不同语义之间的特征分布关系。

表2 在不同场景下分割精度对比结果

Table 2 Comparison of segmentation accuracy across different scenarios

模型	figurines		ramen		teatime	
	mIoU	mBIoU	mIoU	mBIoU	mIoU	mBIoU
LERF[4]	33.5	30.6	28.3	14.7	49.7	42.6
Gaussian Grouping[6]	69.7	67.9	77.0	68.7	71.7	66.1
CLIP2Gaussian	**86.2**	**84.1**	**77.9**	**70.1**	**72.6**	**67.3**

表3展示了语义编码对高斯场景重建质量的影响。实验结果表明,CLIP2Gaussian在将语义信息嵌入3D高斯场景的过程中,对局部高斯特征进行调整,导致几何结构和颜色分布发生微小变化,进而引起*PSNR*和*SSIM*的轻微下降。然而,在*LPIPS*感知指标上,CLIP2Gaussian与原始GS方法保持相当的视觉质量。尽管在部分重建指标上略有降低,但CLIP2Gaussian通过引入语义编码显著提升了模型的功能。借助CLIP与SAM的结合,该方法实现了从2D零样本分割向3D语义理解的扩展,可精确识别并操作任意实例对象。整体而言,CLIP2Gaussian在保持渲染质量的基础上,显著增强了3D场景的语义理解能力和下游编辑的可操作性,展示出更强的灵活性与拓展性。

(a) 场景渲染图像 (b) Lseg(掩码预测) (c) 本文方法(掩码预测) (d) 本文方法(特征可视化)

图4 开放词汇3D语义分割在LERF数据集上的定性比较

Figure 4 Qualitative comparison of open-vocabulary 3D semantic segmentation on the LERF dataset

表3 语义编码融合后的场景重建质量对比

Table 3 Comparative of Scene reconstruction quality following the integration of semantic encoding

算法	语义编码	*PSNR*	*SSIM*	*LPIPS*
GS[5]		28.69	0.870	0.182
CLIP2Gaussian	√	27.86	0.832	0.193

3.4 3D对象删除

图5为3D对象删除与场景修复。如图5所示,本文通过语义追踪实现对场景中指定实例对象的删除。具体而言,根据语义标签识别并定位目标对象所对应的高斯点,并对其进行删除或修改。删除操作可能导致局部区域的高斯分布变得稀疏或杂乱,需要对该区域进行引导式修复,从而提升场景的整体一致性与视觉质量。

(a) 示例1 (b) 示例2 (c) 示例3 (d) 示例4

图5 3D对象删除与场景修复

Figure 5 3D object removal and scene restoration

4　结论

针对场景编辑算法对标注数据的过度依赖和计算复杂度高的问题,本文提出了一种融合 CLIP 与 3D 高斯表示的多模态场景编辑方法 CLIP2Gaussian。该方法通过结合 CLIP 的语义理解能力与 GS 实时渲染优势,有效提升了 3D 高斯的语义感知能力。为实现跨视角语义一致性,本文引入基于语义编码的 3D 高斯追踪策略,结合 2D 掩码与 3D 空间一致性监督,支持基于文本的语义驱动编辑。在此基础上,设计了高效的场景编辑方案,可实现目标实例的精准删除与修改。实验结果表明,CLIP2Gaussian 不仅具备高效的场景渲染能力,还支持用户通过文本实现语义检索与编辑操作,显著提升了 3D 场景的交互能力。

参考文献:

[1] 纪勇, 刘丹丹, 罗勇, 等. 基于霍夫投票的变电站设备三维点云识别算法[J]. 郑州大学学报(工学版), 2019, 40(3): 1-6, 12.
JI Y, LIU D D, LUO Y, et al. Recognition of three-dimensional substation equipment based on Hough transform [J]. Journal of Zhengzhou University (Engineering Science), 2019, 40(3): 1-6, 12.

[2] 陈义飞, 郭胜, 潘文安, 等. 基于多源传感器数据融合的三维场景重建[J]. 郑州大学学报(工学版), 2021, 42(2): 80-86.
CHEN Y F, GUO S, PAN W A, et al. 3D scene reconstruction based on multi-source sensor data fusion[J]. Journal of Zhengzhou University (Engineering Science), 2021, 42(2): 80-86.

[3] MILDENHALL B, SRINIVASAN P P, TANCIK M, et al. NeRF: representing scenes as neural radiance fields for view synthesis[EB/OL]. (2020-03-19)[2025-02-12]. https://doi.org/10.48550/arXiv.2003.08934.

[4] KERR J, KIM C M, GOLDBERG K, et al. LERF: language embedded radiance fields[C]//2023 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2023: 19672-19682.

[5] KERBL B, KOPANAS G, LEIMKUEHLER T, et al. 3D Gaussian splatting for real-time radiance field rendering [J]. ACM Transactions on Graphics, 2023, 42(4): 1-14.

[6] YE M Q, DANELLJAN M, YU F, et al. Gaussian Grouping: segment and edit anything in 3D scenes[EB/OL]. (2023-12-1)[2025-02-12]. https://doi.org/10.48550/arXiv.2312.00732.

[7] KIRILLOV A, MINTUN E, RAVI N, et al. Segment anything[C]//2023 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2023: 3992-4003.

[8] RADFORD A, KIM J W, HALLACY C, et al. Learning transferable visual models from natural language supervision[EB/OL]. (2021-02-26)[2025-02-12]. https://doi.org/10.48550/arXiv.2103.00020.

[9] CHERAGHIAN A, RAHMAN S, PETERSSON L. Zero-shot learning of 3D point cloud objects[C]//The 16th International Conference on Machine Vision Applications. Piscataway: IEEE, 2019: 1-6.

[10] ZHANG R R, GUO Z Y, ZHANG W, et al. PointCLIP: point cloud understanding by CLIP[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2022: 8542-8552.

[11] ZHOU C, LOY C C, DAI B. Extract free dense labels from CLIP[EB/OL]. (2021-12-2)[2025-02-12]. https://doi.org/10.48550/arXiv.2112.01071.

[12] SHEN Q H, YANG X Y, WANG X C. FlashSplat: 2D to 3D Gaussian splatting segmentation solved optimally[EB/OL]. (2024-09-12)[2025-02-12]. https://doi.org/10.48550/arXiv.2409.08270.

[13] WANG H X, VASU P K A, FAGHRI F, et al. SAM-CLIP: merging vision foundation models towards semantic and spatial understanding[C]//2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Piscataway: IEEE, 2024: 3635-3647.

[14] GORDON O, AVRAHAMI O, LISCHINSKI D. Blended-NeRF: zero-shot object generation and blending in existing neural radiance fields[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2023: 2941-2951.

[15] HAQUE A, TANCIK M, EFROS A A, et al. Instruct-NeRF2NeRF: editing 3D scenes with instructions[C]// 2023 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2023: 19683-19693.

[16] CHENG H K, OH S W, PRICE B, et al. Tracking anything with decoupled video segmentation[C]// 2023 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2023: 1316-1326.

[17] ESTER M, KRIEGEL H P, SANDER J, et al. A density-based algorithm for discovering clusters in large spatial databases with noise[C]// Proceedings of the Second International Conference on Knowledge Discovery and Data Mining. New York: ACM, 1996: 226-231.

[18] SUVOROV R, LOGACHEVA E, MASHIKHIN A, et al. Resolution-robust large mask inpainting with Fourier convolutions[C]//2022 IEEE/CVF Winter Conference on Applications of Computer Vision. Piscataway: IEEE, 2022: 3172-3182.

[19] BARRON J T, MILDENHALL B, VERBIN D, et al. Mip-NeRF 360: unbounded anti-aliased neural radiance fields[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2022: 5460-5469.

[20] WANG Y J, LI J H, LU Y, et al. Image quality evaluation based on image weighted separating block peak signal to noise ratio[C]//Proceedings of 2003 International Conference on Neural Networks and Signal. Piscataway: IEEE, 2003: 994-997.

[21] WANG Z, BOVIK A C, SHEIKH H R, et al. Image quality assessment: from error visibility to structural similarity[J]. IEEE Transactions on Image Processing, 2004, 13(4): 600-612.

[22] ZHANG R, ISOLA P, EFROS A A, et al. The unreasonable effectiveness of deep features as a perceptual metric[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 586-595.

[23] LI B Y, WEINBERGER K Q, BELONGIE S, et al. Language-driven semantic segmentation[EB/OL]. (2022-01-10)[2025-02-12]. https://doi.org/10.48550/arXiv.2201.03546.

Multimodal Scene Editing Algorithm Integrating CLIP and 3D Gaussian

CAO Yangjie, WANG Weiping, LI Zhenqiang, XIE Jun, LYU Runfeng

(School of Cyber Science and Engineering, Zhengzhou University, Zhengzhou 450002, China)

Abstract: To address the issues of excessive reliance on annotated data and high computational complexity in 3D scene editing algorithms, in this study a multimodal scene editing method named CLIP2Gaussian was proposed, which integrated CLIP with 3D Gaussian. Firstly, the algorithm employed SAM to extract target masks from multi-view images and introduced a bidirectional propagation strategy to ensure mask consistency across different views. Secondly, the extracted masks were assigned semantic labels using CLIP and mapped to 3D Gaussian points to enable semantic embedding in the 3D scene. Finally, a differentiable rendering mechanism was used to optimize the parameters of the 3D Gaussians, and a spatial consistency regularization strategy was introduced by applying clustering to enhance the consistency and stability of semantic labels in 3D space. Experimental results showed that CLIP2Gaussian achieved 61.23% IoU on the LERF dataset and a per-query response time of 0.57 seconds in semantic segmentation tasks, improving the speed by 54 times compared to LERF while achieving superior accuracy and efficiency. Ablation studies further verified that the proposed method enabled precisely editing of target regions with minimal disturbance to the original scene.

Keywords: 3D reconstruction; zero-shot learning; scene understanding; scene editing; 3D Gaussian