

Pipeline: 3DGs初始化 → 多模态稠密语义特征提取 → 高维特征量化压缩 → 索引语义嵌入高斯 → 不确定性引导语义平滑

3.2 稠密语言特征提取 (Dense Language Feature Extraction)

从多视角2D图像中提取像素级稠密语义特征。

问题: CLIP仅能输出“图像级”全局特征,无法提供像素级语义标注。仅用单一特征存在边界模糊。

CLIP + DINO 混合特征:

1. CLIP特征提取(全局语义)

借鉴3DViS分层随机裁剪技术,对每张多视角图像进行多尺度裁剪(不同大小、位置的图像块)。输入

CLIP视觉编码器,提取各层特征并聚合,最后归一化得到像素级稠密CLIP特征

提供与语言对齐的全局语义信息,支持开放词汇查询

$$F_{I,x,y}^{CLIP} \in \mathbb{R}^{d_{CLIP}}$$
 (d_{CLIP} 为CLIP编码器输出维度)

$$I = RCP, \Sigma, C, \alpha, pcam$$

2. DINO特征提取(局部边界)

DINO为自监督视觉Transformer模型,直接提取像素级稠密特征。

增强语义区域边界区分度,解决CLIP特征边界模糊问题

$$F_{I,x,y}^{DINO} \in \mathbb{R}^{d_{DINO}}$$
 (d_{DINO} 为DINO编码器输出维度)

3. 混合特征融合

将CLIP与DINO的像素级特征在通道维度拼接,形成最终混合稠密语言特征

$$F_{I,x,y} = F_{I,x,y}^{CLIP} \oplus F_{I,x,y}^{DINO}$$

$$\text{混合特征维度 } d = d_{CLIP} + d_{DINO}$$

3.3 语言特征量化 (核心创新1: 解决内存爆炸)

局部语义冗余: 同一物体的像素级特征高度相似,无需存储完整高维特征

场景语义有限: 单一场景仅覆盖CLIP语义空间的极小部分

量化目标:

将高维混合特征 $F \in \mathbb{R}^d$ 映射为离散特征空间中的索引,仅存储索引而非原始特征。优化离散特征空间,确保量化后特征能精确还原原始语义。

Step:

1. 构建离散特征空间 S

$$S = \{f_i \in \mathbb{R}^d \mid i=1, 2, \dots, N\}$$

N —特征空间大小(控制压缩率) f_i —语义基础向量(可学习)

初始化: 随机初始化 N 个基础向量,后续训练优化,能覆盖当前场景所有语义

2. 特征量化与索引生成

对每个像素的混合特征 $F_{I,x,y}$, 在离散空间 S 中检索“最相似的基础向量”,用其索引 m 表示该特征。

相似度度量: 采用余弦相似度

$$D(F, f_i) = \cos(F^{CLIP} \cdot f_i^{CLIP}) + \lambda_{DINO} \cos(F^{DINO} \cdot f_i^{DINO})$$

λ_{DINO} 超参数,调节DINO特征在相似度计算中的贡献, f_i^{CLIP} —基础向量 f_i 中的 CLIP 部分

索引选择: 选择相似度最高的基础向量索引 $m = \arg \max_i D(F, f_i)$

量化特征重构: 通过索引 m 对应的基础向量重构量化特征 \hat{F}

$$\hat{F} = \sum_i f_i \cdot \text{onehot}(m)_i$$

$\text{onehot}(m)$ —索引 m 的独热编码,第 m 位为 1,其余为 0

3. 量化结果：语义索引图 M ：

每张多视角图经过量化后，得到一张语义索引图 $M \in R^{H \times W \times 1}$ ，其中每个像素值为对应的基础向量索引 m ($int, 1 \sim N$)。该索引图将作为后续 3D 高斯语义嵌入的监督目标。

4. 离散特征空间的优化（量化损失）

设计双重损失优化离散空间 S 和索引图 M

(1) 余弦相似度损失 \mathcal{L}_{cos} ：

最小化原始特征 F 与量化重构特征 \hat{F} 的语义差异：

$$\begin{aligned}\mathcal{L}_{cos}(F_i) = & (1 - \cos(F_i^{\text{CLIP}} \cdot \hat{F}_i^{\text{CLIP}})) \\ & + \lambda_{\text{DINO}}(1 - \cos(F_i^{\text{DINO}} \cdot \hat{F}_i^{\text{DINO}})).\end{aligned}$$

确保量化后的特征在 CLIP 和 DINO 两个维度上都与原始特征语义一致。

(2) 负载均衡损失 \mathcal{L}_{lb} ：

借鉴 Switch Transformer 设计，避免基础向量利用率不均（防止部分基础向量被过度使用，部分闲置）

$$\mathcal{L}_{lb} = \sum_{i=1}^N (\mathbf{r} \circ \mathbf{p}), \quad \begin{aligned}r & \in R^N && \text{—每个基础向量的利用率 (实际被选择次数 / 总选择次数)} \\ p & \in P^N && \text{—每个基础向量的平均选择概率} \\ \circ & \quad \text{逐元素乘积}\end{aligned}$$

通过最小化该损失，迫使所有基础向量被均匀使用

5. 量化总损失：

$$\mathcal{L}_q = \lambda_{cos} \mathcal{L}_{cos} + \lambda_{lb} \mathcal{L}_{lb}. \quad \lambda_{cos} = 1 \quad \lambda_{lb} = 0.5$$

3.4 语言嵌入 3D Gaussians (核心创新 2：语义嵌入与平滑)

不为高斯点直接存储索引，而是学习“连续紧凑的语义向量”，并通过不确定性引导的平滑机制解决多视角矛盾

1. 3D 高斯的紧凑语义特征：

1) 扩展原有的 3D 高斯属性

在原有的几何/外观属性基础上，为每个 3D 高斯点新增“紧凑语义向量” $SG \in R^{d_s}$ ， d_s 为超参 ($D=8$)

2) 语义向量的渲染与解码

通过可微分光栅化，将 3D 高斯的紧凑语义向量 SG 投影为 2D 语义特征图，通过小型 MLP 解码器映射回“离散语义索引分布”，与量化得到的索引图 M 对齐：

语义渲染：给定相机姿态 P_{cam} ，通过语义专用光栅化函数 R_s ，将所有高斯点的 SG 渲染为 2D 语义特征图 $R_s(G; P_{cam}) \in R^{H \times W \times d_s}$

MLP 解码：用小型 MLP 解码器 D 将 2D 语义特征图映射为“索引分布” $\hat{M} \in R^{H \times W \times N}$ (每个像素对应 N 个基础向量的概率)，通过 softmax 归一化。 $\hat{M} = \text{softmax}(D(R_s(G; P_{cam})))$

监督优化：通过交叉熵损失 (CE) 最小化预测索引分布 \hat{M} 与真实图 M 的差异

$$\mathcal{L}_{CE} = CE(\hat{M}, M), \quad M \text{ — 量化得到的离散索引图}$$

2. 语义特征平滑：

1) 语义不确定性学习：

为每个高斯点新增“语义不确定性” $\mu \in [0, 1]$ 用于衡量该点语义特征的稳定性

定义： $\mu=0$ ，语义完全稳定 (多视角特征一致) $\mu=1$ 相反

初始化： $\mu=0$ 后续通过损失迭代更新

优化损失：

a. 带权重的交叉熵损失：用不确定性加权 CE 损失，降低高不确定性点的监督权重。

$$\mathcal{L}_{CE} = \frac{\sum CE(\hat{M}, M) \circ (1 - R_u(G; P_{cam}))}{H \times W},$$

$R_u(G; P_{cam}) \in R^{H \times W \times 1}$ — 不确定性值的 2D 渲染图

$1 - R_u$ — 权重因子，高不确定性点 (R_u 大) 的 CE 损失权重降低

b. 不确定性正则化损失：防止所有高斯点的从收敛到 1

$$\mathcal{L}_u = \frac{\sum R_u(\mathcal{G}; p_{\text{cam}})}{H \times W}.$$

2) $\mathcal{L}_s = \lambda_{\text{CE}} \mathcal{L}_{\text{CE}} + \lambda_u \mathcal{L}_u$, $\lambda_{\text{CE}}=1$ $\lambda_u=1$

3) 自适应空间平滑损失：

利用空间相邻点语义相似的先验，通过低频率特征约束，降低语义向量的空间频率

(1) 生成低频率平滑特征：用小型 MLP 输入 3D 高斯点的位置编码 (PE)，生成低频率平滑语义特征 S_{MLP} 和 G_{MLP}
 $S_{\text{MLP}} = \text{MLP}(\text{PE}(p))$,

$\text{PE}(p)$ — 低频率 MLP 的归纳偏置是学习低频信号，因此 S_{MLP} 天然具备空间平滑性

(2) 自适应平滑损失： $\mathcal{L}_{\text{smo}} = \|S_{\text{MLP}} - S_G^*\|_2 + \max(u_G^*, w_s) \|S_{\text{MLP}} - S_G\|_2$ 紧凑语义向量

——梯度停止 $\|S_{\text{MLP}} - S_G^\|$ ——让 S_{MLP} 靠近当前 S_G (固定 S_G ，优化 MLP 生成与 S_G

相近的平滑特征) $\max(u_G^*, w_s) \|S_{\text{MLP}} - S_G\|_2$ ——让 S_G 靠近 S_{MLP} (固定 S_{MLP} ，优化 S_G 向平滑特征靠拢)
且权重由不确定性 u_G 控制：

u_G 越高 (语义越不稳定)，权重越大， S_G 越向平滑特征靠拢

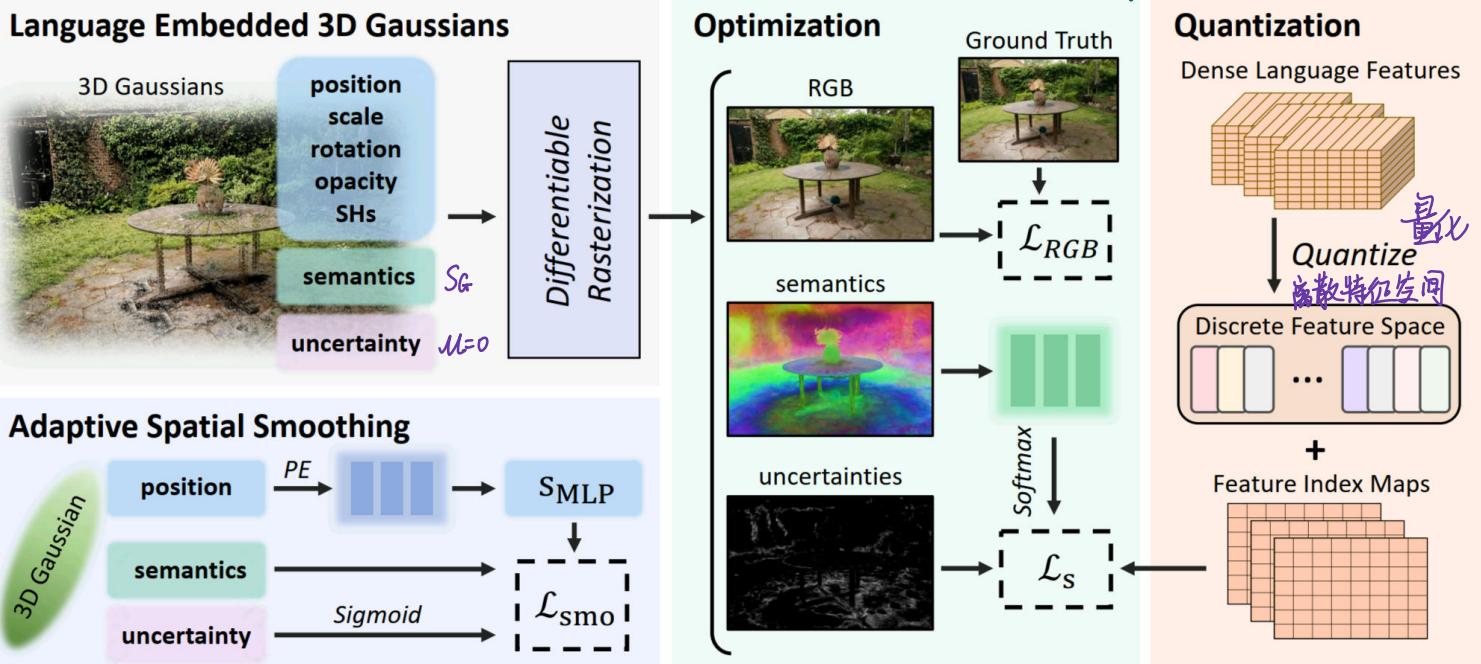
$w_s=0.1$ ，最小权重，确保即使 $u_G=0$ ，也会施加轻微平滑

不稳定区域强平滑，稳定区域弱平滑

3. 总优化损失

$$L = \lambda_{\text{RGB}} \cdot \mathcal{L}_{\text{RGB}} + \lambda_s \cdot \mathcal{L}_s + \lambda_{\text{smo}} \cdot \mathcal{L}_{\text{smo}}$$

\mathcal{L}_{RGB} — 3DGs 原始损失 \mathcal{L}_s — 语义嵌入损失 \mathcal{L}_{smo} — 语义平滑损失



1. 初始化：3D Gaussian 几何、外观、紧凑语义向量 S_G ，不确定性 $u=0$

2. 特征提取：对多视角图像提取 CLIP+DIIM 混合稠密特征

3. 量化优化：通过 L_s 优化离散特征空间 S 和语义索引图 M

4. 语义嵌入优化：通过 L_s 优化 S_G 和不确定性 u (渲染 $S_G \rightarrow$ MLP 解码 \rightarrow 与 M 计算 CE 损失)

5. 语义平滑优化：通过 L_{smo} 优化 S_G 和 MLP

6. 迭代更新：同步优化几何外观和语义属性

Limitation: 细粒度几何语义识别不足 (高分辨率场景)

向量量化导致的细粒度语义损失

高反光/半透明物体的语义检测挑战