浙江大學

硕 士 研 究 生 读 书 报 告

题目《segment ant 3d gaussians》 读书报告

作者姓名 林靖

作者学号 22351124

指导教师 李启雷

学科专业 软件工程

所在学院 软件学院

提交日期 2024 年 1月 8日

摘要

本文将会介绍《Segment Any 3D Gaussians》一文的内容，主要包括一下几个方面：1、文章所要解决的问题；2、解决问题的方法；3、论文总结。

1. 引言

在各个领域，如场景操作、自动标注和虚拟现实，亮度场中的交互式三维分割引起研究人员的广泛关注。以往的方法主要包括训练特征场以模仿自我监督视觉模型提取的多视角二维特征，将二维视觉特征提升到三维空间中，然后使用三维特征相似性来测量两个点是否属于同一对象。这些方法由于其简单的分割流程而快速，但代价是分割的粒度可能较粗，因为它们缺乏解析特征中包含的信息的机制（例如，分割解码器）。相反，另一种方法是将二维分割基础模型提升到三维，通过直接将多视角细粒度的二维分割结果投射到三维掩膜网格上。尽管这种方法可以产生精确的分割结果，但由于需要多次执行基础模型和体渲染，它会带来大量的时间开销，限制了交互性。特别是对于需要进行分割的复杂场景中的多个对象，这种计算成本是无法承受的。

上述讨论揭示了当前现有的范例在实现效率和准确性方面的困境，指出了限制现有范式性能的两个因素。首先，先前方法中使用的隐式辐射场阻碍了高效分割：必须遍历 3D 空间以获取 3D 对象。其次，使用 2D 分割解码器带来了较高的分割质量但低效率。因此，我们从辐射场的最新突破开始重新审视这个任务：3D 高斯喷洒（3DGS）由于其高质量和实时渲染能力而成为游戏规则改变者。它采用一组 3D 彩色高斯函数来表示 3D 场景。这些高斯函数的均值表示它们在 3D 空间中的位置，因此 3DGS 可以被看作是一种点云，有助于绕过广泛的、通常是空的 3D 空间的大量处理，并提供丰富的显式 3D 先验知识。有了这样的点云结构，3DGS 不仅实现了高效渲染，还成为了分割任务的理想候选者。

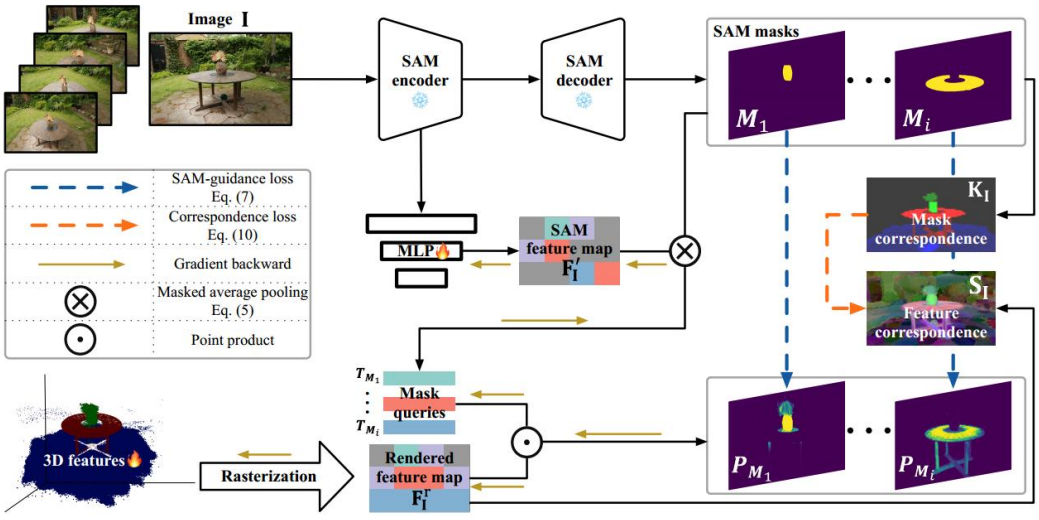
基于 3DGS，我们提出将 2D 分割基础模型（即 Segment Anything Model）的细粒度分割能力融入到 3D 高斯函数中。这种策略与之前的方法不同，之前的方法侧重于将 2D 视觉特征提升到 3D，并实现细粒度的 3D 分割。此外，它避免了推理过程中耗时的多次 2D 分割模型前向传递。通过基于 Segment Anything Model (SAM) 自动提取的掩模，在已提取的掩模基础上，训练 3D 高斯函数的特征。在推理过程中，通过输入提示生成一组查询，然后通过高效特征匹配来检索预期的高斯函数。我们称之为 Segment Any 3D Gaussians（SAGA），我们的方法可以在毫秒级实现细粒度的 3D 分

割，并支持包括点、涂鸦和掩模在内的各种提示类型。对现有基准的评估表明，SAGA 的分割质量与之前的最新技术相当。作为对 3D 高斯函数交互式分割的首次尝试，SAGA 具有很大的灵活性，可以适应多种提示类型，包括掩模、点和涂鸦。我们对现有基准进行的评估表明，SAGA 的性能与最新技术相当。值得注意的是，高斯函数特征的训练通常仅需要 5-10 分钟。随后，大多数目标对象的分割可以在毫秒级完成，实现近 1000 倍的加速。

2. 实现方法

如图所示，给定预先训练的 3DGS 模型 G 及其训练集 I，我们首先使用 SAM 编码器来提取 2D 特征图 F SAM I∈RC SAM×H×W 和 I 中每个图像 I∈R H×W 的一组多粒度掩码 MSAM I。然后，我们基于提取的掩码为 g 中的每个高斯 g 训练一个低维特征 fg∈RC，以聚合跨视图一致的多粒度分割信息（C 表示特征维度，为默认设置为 32）。这是通过精心设计的 SAM 制导损失实现的。为了进一步增强特征的紧凑性，我们从提取的掩码中导出逐点对应关系，并将它们提取到特征中（即对应关系损失）。

在推理阶段，对于具有相机姿态 v2 的特定视图，基于输入提示 P 生成一组查询 Q。然后，通过与学习到的特征的有效特征匹配，使用这些查询来检索相应目标的 3D 高斯。此外，我们还引入了一种高效的后处理操作，该操作利用 3DGS 的点云状结构提供的强 3D 先验来细化检索到的 3D 高斯。



SAGA 的整体 pipeline。给定预先训练的 3DGS 模型及其训练集，我们将低维 3D 特征附加到模型中的每个高斯。对于训练集中的每个图像，我们使用 SAM 来提取 2D 特征和一组掩码。然后，我们通过可微分光栅化绘制 2D 特征图，并训练具有两个损失的附加特征：即 SAM 制导

损失和对应损失。前者采用 SAM 特征来引导 3D 特征从模糊的 2D 掩模中学习 3D 分割。后者提取从掩模导出的逐点对应关系，以增强特征的紧凑性。

3D 高斯散射（3DGS）作为辐射场的最新进展，3DGS[21]使用可训练的 3D 高斯来表示 3D 场景，并提出了一种用于渲染和训练的高效可微分光栅化算法。

给定具有相机姿态的多视图 2D 图像的训练数据集 I，3DGS 学习一组 3D 彩色高斯 G＝{g1，g2，…，gN}，其中 N 表示场景中 3D 高斯的数量。每个高斯的平均值表示其在 3D 空间中的位置，协方差表示尺度。因此，3DGS 可视为一种点云。给定特定的相机姿势，3DGS 将 3D 高斯投影到 2D，然后通过混合与像素重叠的一组有序高斯 N 来计算像素的颜色 C：

$$\mathbf{C} = \sum_{i \in \mathcal{N}} \mathbf{c}_i \alpha_i \prod_{j=1}^{i-1} (1 - \alpha_j).$$

其中 ci 是每个高斯的颜色，αi 是通过评估协方差Σ乘以学习的每高斯不透明度的 2D 高斯来给出的。从等式（1）我们可以了解光栅化过程的线性：渲染像素的颜色是所涉及的高斯系数的加权和。

在我们的框架中，这种特性确保了 3D 特征与 2D 渲染特征的对齐。

分段任意模型（SAM）SAM 将图像 I 和一组提示 P 作为输入，并输出相应的 2D 分段掩码 M，即

$$\mathbf{M} = \mathrm{SAM}(\mathbf{I}, \mathcal{P}).$$

3D 高斯的初始分割 Gt 表现出两个主要问题：（i）存在多余的噪声高斯和（ii）忽略了目标对象的某些高斯积分。为了解决这个问题，我们使用了传统的点云分割技术[36，37，42]，包括统计滤波和区域增长。对于基于点和涂鸦提示的分割，采用统计滤波来滤除噪声高斯。对于掩码提示和基于 SAM 的提示，将 2D 掩码投影到 G t 上以获得一组经过验证的高斯，并投影到 G 上以排除不需要的高斯。所得到的经验证的高斯作为区域生长算法的种子。最后，应用基于球查询的区域增长方法从原始模型 G 中检索目标所需的所有高斯系数。

给定具有其特定相机姿态 v 的训练图像 I，我们首先根据预先训练的 3DGS 模型 G 来渲染相应的特征图。类似于等式（1），像素 p 的渲染特征 F

r I，p 计算为

$$\mathbf{F}_{\mathbf{I},p}^{r}=\sum_{i\in\mathcal{N}}\mathbf{f}_{i}\alpha_{i}\prod_{j=1}^{i-1}(1-\alpha_{j})$$

其中 N 是与像素重叠的高斯有序集合。在训练阶段，我们冻结 3D 高斯 G 的所有其他属性（例如，均值、协方差和不透明度），除了新附加的特征。

SAM 制导损失通过 SAM 自动提取的 2D 掩模 MI 是复杂和令人困惑的（即，3D 空间中的一个点可能被分割为不同视图上的不同对象/部分）。这种模糊的监督信号对从头开始训练 3D 特征提出了巨大的挑战。为了解决这个问题，我们建议使用 SAM 生成的功能进行指导。如图 2 所示，我们首先采用 MLP Φ 将 SAM 特征投影到与 3D 特征相同的低维空间：

$$\mathbf{F}_{\mathbf{I}}^{\prime}=\varphi(\mathbf{F}_{\mathbf{I}}^{\mathrm{SAM}}).$$

然后，对于 MSAM I 中的每个提取的掩码 M，我们通过掩码平均池运算获得相应的查询 TM∈RC：

$$\mathbf{T}_{\mathbf{M}}=\frac{1}{||\mathbf{M}||_{1}}\sum_{p=1}^{HW}\mathbb{1}(\mathbf{M}_{p}=1)\mathbf{F}_{\mathbf{I},p}^{\prime},$$

对应损失在实践中，我们发现 SAM 制导损失的学习特征不够紧凑，这降低了各种提示的分割质量（更多细节请参阅第 4 节中的消融研究）。受先前对比对应蒸馏方法[9，17]的启发，我们引入了对应损失来解决这个问题。

如前所述，对于训练集 I 中每个高度为 H、宽度为 W 的图像 I，使用 SAM 提取一组掩码 MI。考虑到 I 中的两个像素 p1、p2，它们可能属于 MI 中的多个掩码。设 Mp1 I、Mp2 I 分别表示 p1、p2 所属的掩码。直观地说，如果两个集合的并集上的交集更大，那么这两个像素应该共享更相似的特征。因此，掩码对应关系 KI（p1，p2）被定义为：

$$\mathbf{K}_{\mathbf{I}}(p_{1},p_{2})=\frac{|\mathcal{M}_{\mathbf{I}}^{p_{1}}\cap\mathcal{M}_{\mathbf{I}}^{p_{2}}|}{|\mathcal{M}_{\mathbf{I}}^{p_{1}}\cup\mathcal{M}_{\mathbf{I}}^{p_{2}}|}.$$

两个像素 p1，p2 之间的特征对应关系 SI（p1，p2）被定义为它们的渲染特征之间的余弦相似性：

$$\mathbf{S_I}(p_1, p_2) =< \mathbf{F}^r_{\mathbf{I},p_1}, \mathbf{F}^r_{\mathbf{I},p_2} >,$$

则对应损失定义为：

$$\mathcal{L}_{\text{corr}} = -\sum_{\mathbf{I}\in\mathcal{I}} \sum_{p_1}^{HW} \sum_{p_2}^{HW} \mathbf{K_I}(p_1, p_2)\mathbf{S_I}(p_1, p_2).$$

如果两个像素从不属于同一段，我们通过将 KI 中的 0 值条目设置为−1 来降低它们的特征相似性。

对于 SAM 制导损失（方程（7））和对应损失（方程。（10））这两个分量，SAGA 的最终损失为：

$$\mathcal{L} = \mathcal{L}_{\text{SAM}} + \lambda\mathcal{L}_{\text{corr}},$$

基于区域增长的过滤来自掩码提示或基于 SAM 提示的 2D 掩码可以作为准确定位目标的先验。最初，我们将掩码投影到分段的高斯 G t 上，产生一个经验证的高斯子集，表示为 G c。随后，对于 GC 内的每个高斯 g，我们计算其到同一子集中最近邻居的欧几里得距离 dg：

$$d^{\mathcal{G}^c}_{\mathbf{g}} = \min\{D(\mathbf{g}, \mathbf{g}')|\mathbf{g}' \in \mathcal{G}^c\}$$

尽管点提示和涂鸦提示也可以粗略定位目标，但基于它们的区域增长是耗时的。因此，我们只在有掩模的情况下应用基于区域生长的过滤。

3. 总结

本文介绍了一种新的交互式三维分割方法 SAGA。作为 3D 高斯交互式分割的第一次尝试，SAGA 使用两个精心设计的损失，有效地将来自分段任意模型（SAM）的知识提取到 3D 高斯中。训练后，SAGA 允许在各种输入类型（如点、涂鸦和掩码）之间进行快速、毫秒级的 3D 分割。进行了大量的实验来证明 SAGA 的效率和有效性。