

参赛学生姓名：陶誉文

中学：上海市民办平和学校（金鼎校区）

省份：上海 浦东新区

国家/地区：中国大陆

指导老师姓名：谭鑫

指导老师单位：华东师范大学

论文题目：CaRF: 在指代性3D 高斯泼溅分割中增强多视图一致性

摘要

指代性 3D 高斯泼溅分割 (R3DGS) 旨在根据语言表达在 3D 场景中定位相应区域, 但现有方法因依赖 2D 伪监督和视图特定特征学习而存在跨视图一致性不足的问题。本研究旨在解决 R3DGS 中的多视图不一致挑战, 提升 3D 语言锚定的几何一致性。本项目提出相机感知指代场 (CaRF), 直接在 3D 高斯空间操作的可微分框架。CaRF 包含两个核心创新: GFCE 将相机几何融入高斯-文本交互以建模视图相关变化; 训练中配对视图监督 (ITPVS) 在训练期间对齐校准视图间的每个高斯逻辑值, 缓解单视图过拟合。在三个基准测试上的实验表明, CaRF 在 Ref-LERF、LERF-OVS 和 3D-OVS 数据集上的 mIoU 相对最先进方法平均提高 16.8%、4.3% 和 2.0%。消融研究验证了 ITPVS 和 GFCE 的有效性, 定性结果显示 CaRF 在不同视图间产生几何一致的掩码。CaRF 成功解决了 R3DGS 中的多视图一致性问题, 为 3D 语言锚定建立了新基线, 对具身 AI、AR/VR 交互和自主感知具有应用价值。

关键词: 3D 高斯泼溅, 语义分割, 多视图一致性

Abstract

Referring 3D Gaussian Splatting Segmentation (R3DGS) aims to localize regions in 3D scenes based on language expressions, but existing methods suffer from insufficient cross-view consistency due to reliance on 2D pseudo supervision and view-specific feature learning. This study addresses the multi-view inconsistency challenge in R3DGS to enhance geometric consistency of 3D language grounding. This project proposes Camera-Aware Referring Field (CaRF), a differentiable framework operating directly in 3D Gaussian space. CaRF incorporates two core innovations: GFCE integrates camera geometry into Gaussian-text interactions to model view-dependent variations; In-Training Paired-View Supervision (ITPVS) aligns per-Gaussian logits across calibrated views during training to mitigate single-view overfitting. Experiments on three benchmarks show CaRF achieves average improvements of 16.8%, 4.3%, and 2.0% in mIoU over state-of-the-art methods on Ref-LERF, LERF-OVS, and 3D-OVS datasets. Ablation studies validate the effectiveness of ITPVS and GFCE,

while qualitative results show CaRF produces geometrically consistent masks across views. CaRF successfully addresses the multi-view consistency problem in R3DGS, establishing a new baseline for 3D language grounding, with application value for embodied AI, AR/VR interaction, and autonomous perception.

Keywords: 3D Gaussian Splatting, semantic segmentation, multi-view consistency

目录

图目录	iii
表目录	v
1 引言	1
2 相关工作	3
2.1 3D 神经表示	3
2.2 高斯泼溅中的 3D 分割	4
2.3 指代性 3D 高斯泼溅分割	4
3 预备知识：符号与任务定义	5
4 CaRF：相机感知指代场	7
4.1 动机与框架概述	7
4.2 高斯场相机编码	9
4.3 训练中配对视图监督	11
4.4 训练过程	12
5 实验	13
5.1 实验设置	13
5.2 在 Ref-Lerf 数据集上的结果	14
5.3 在 3D 开放词汇分割数据集上相对于最先进技术的结果	15
5.4 消融研究	17
5.5 可视化	20
5.6 计算与精度之间的平衡	21
6 结论与讨论	22

2026 S.-T. Yau High School Science Award
仅用于2026丘成桐中学科学奖论文公示

图目录

1	本项目方法的动机。(a) 先前的方法在每次迭代中执行单视图栅格化和使用单个 GT 掩码进行监督, 这可能会引入视图间的不一致性。(b) CaRF 在训练期间添加多视图联合监督, 以显式强制执行跨视图一致性, 从而在推理时产生跨视图一致、抗伪影的掩码。	2
2	ReferSplat [9] 本项目 CaRF 在 <i>Ramen</i> 和 <i>Waldo Kitchen</i> 场景上的定性比较。ReferSplat 未能保持 多视图一致性 : 在 <i>Ramen</i> 中, 水杯在视图 1 中完全丢失, 在视图 2 中显示出严重的渲染伪影; 在 <i>Waldo Kitchen</i> 中, ”冰箱”掩码泄漏到相邻区域并显示碎片化边界。相比之下, CaRF 在不同视图间产生连贯且几何一致的掩码, 有效保留了细粒度对象的细节。	6
3	提出的 CaRF 框架概述。给定校准的多视图 RGB 图像和一个文本查询, CaRF 首先通过置信度加权选择生成鲁棒的伪掩码。每个高斯通过一个与语言嵌入通过跨交互模块交互的指代特征进行增强。然后, 相机感知编码集成了来自相机内参和外参的几何信息以增强视图一致性。最后, 指代特征被栅格化为 2D 响应图, 并在训练中监督下进行优化, 以实现几何感知、视图一致的指代分割。	8
4	在 Ref-LERF 数据集上跨两个代表性场景和两个校准视图的定性比较。每行显示来自 GS-Grouping [36]、 ReferSplat [9]、本项目 CaRF 和 GT 掩码的结果。GS-Grouping 由于缺乏细粒度的空间推理, 经常混淆外观相似的附近对象, 而 ReferSplat 由于监督源自单视图伪掩码, 产生视图不一致或不完整的分割。相比之下, CaRF 通过利用相机感知几何和配对视图一致性约束, 跨多个视图生成空间连贯和语义精确的掩码。	20
5	在 Ref-LERF 数据集上跨两个代表性场景和两个校准视图的失败案例。每行显示来自 ReferSplat [9]、本项目的 CaRF 和 GT 掩码的结果。大多数错误源于不准确或未对齐的伪 GT 掩码 (见最后一行)。	21

-
- 6 CaRF 与 ReferSplat [9] 的定量比较。(a) 参数增加 (ΔParams), 其中 CaRF 仅引入了 25.7K 参数的小额开销。(b) FLOPs 差异 (ΔFLOPs), 显示两种方法共享相同的渲染管道, 计算成本几乎相同。(c) 在三个基准测试上的性能改进 (ΔmIoU)。 22

表目录

1	在 Ref-Lerf 数据集上与最先进方法在 mIoU ($\uparrow$) 方面的比较。更高的值表示更好的性能。粗体数字表示最佳结果。最后一行报告了 CaRF 相对于 ReferSplat 的相对改进 (%).	15
2	在 LERF-OVS 数据集上与最先进方法在 mIoU ($\uparrow$) 方面的定量比较。更高的值表示更好的性能。粗体数字表示最佳结果。最后一行报告了 CaRF 相对于 ReferSplat 的相对改进 (%).	16
3	在 3D-OVS 数据集上与最先进方法在 mIoU ($\uparrow$) 方面的定量比较。更高的值表示更好的性能。粗体数字表示最佳结果。最后一行报告了 CaRF 相对于 ReferSplat 的相对改进 (%).	16
4	方法上的消融研究。ITPVS 和 GFCE 分别表示训练中配对视图监督和高斯场相机编码组件。	17
5	在 Ramen 和 Kitchen 场景上评估的高斯选择算法的消融研究。	18
6	相机编码融合设计的消融研究。	18
7	视图数量的消融研究。	19

1 引言

指代性 3D 高斯泼溅分割 (R3DGS) [9], 建立在 3D 高斯泼溅 (3DGS) [16] 之上, 学习一个每个场景的语言场, 将每个高斯映射到一个语义特征, 使得能够从校准图像和自由形式的文本查询 (例如, “窗边的红色物体”) 中对描述的 3D 区域进行空间感知的、开放词汇的检索和分割。这种根据语言描述在 3D 中解释和定位对象的能力对于现实世界的应用至关重要, 例如具身 AI [30]、自动驾驶 [11, 7] 和 VR/AR 系统 [14, 21]。与传统的 3DGS 分割 [17, 24, 41, 3, 37, 20, 4] 专注于类别级或实例级分离不同, R3DGS 必须解释通常编码复杂空间关系和细粒度对象属性的自由形式语言表达。此类描述可能描述在某些视角下部分遮挡或不可见的目标, 使得一致的多视图空间推理至关重要, 并使 R3DGS 成为 3D 场景理解中一个更具挑战性且未被充分探索的问题。

最近的进展在将语言理解与 3D 场景表示集成方面取得了显著进展, 但在实现几何感知和视图一致的推理方面仍然存在重大挑战。ReferSplat [9] 迈出了重要一步, 引入了位置感知的跨模态交互和高斯-文本对比学习, 共同对齐语言和几何特征以推理空间关系和细粒度对象属性。其他代表性方法, 如 LangSplat [24]、Feature3DGS [41]、GaussianGrouping [37] 和 OpenGaussian [35], 通过提取 2D 视觉-语言特征或将 2D 掩码提升到 3D, 将 3DGS 分割扩展到开放词汇和语言引导的理解。尽管它们取得了成功, 但这些方法仍然严重依赖于单视图图像特征监督, 这会过拟合视图特定的伪影, 并且没有约束相同的高斯在校准视图中应如何表现, 如 1(a) 所示。此外, 它们对不可微分的 2D 特征投影或掩码提升的依赖限制了对原生 3D 几何的利用, 导致多视图一致性下降和空间推理能力减弱。

为了进一步研究 R3DGS 中多视图不一致的根本原因, 本项目分析了现有流程, 并观察到单视图伪监督倾向于过拟合视图特定的伪影, 而没有约束相同的高斯在校准视图间应如何表现 (见图 1(a)), 导致泛化能力下降和掩码不一致。尽管一些先前的研究在相关的 3D 任务中探索了多视图一致性, 但直接将此类策略应用于 R3DGS 被证明是无效的。预处理/后处理方法, 如 FastLGS [13]、econSG [40]、FMLGS [32] 和 OmniSeg3D [38], 试图强制跨视图一致性。然而, 它们的不可微分组件使其对阈值高度敏感且容易产生误差累积, 限制了它们在指代场景所需的细粒度推理中的适用性 [42]。相比之下, 可微分聚类方法, 包

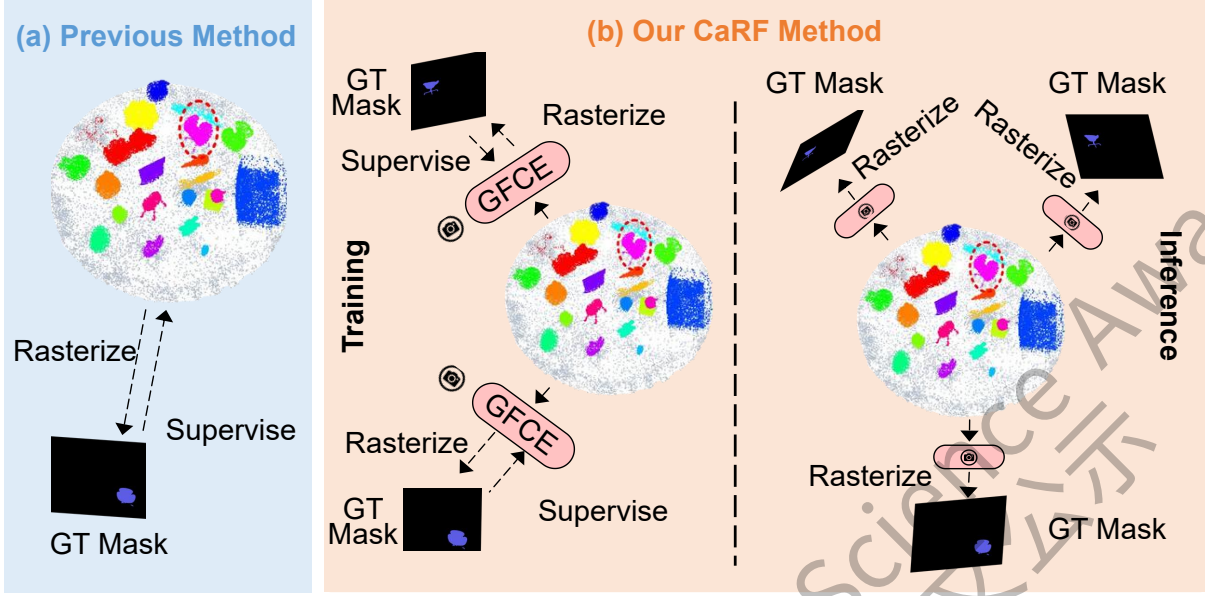


图 1: 本项目方法的动机。(a) 先前的方法在每次迭代中执行单视图栅格化和使用单个 GT 掩码进行监督, 这可能会引入视图间的不一致性。(b) CaRF 在训练期间添加多视图联合监督, 以显式强制执行跨视图一致性, 从而在推理时产生跨视图一致、抗伪影的掩码。

括 Gaussian Grouping [37]、Contrastive Gaussian Clustering [31]、OpenGaussian [35] 和 Click-Gaussian [4], 通过特征正则化促进跨视图稳定性。然而, 这些方法仍然主要依赖于从基础模型中提取的 2D 图像级特征, 未能充分利用对鲁棒空间推理至关重要的内在 3D 几何 [39]。这些限制表明, 在 R3DGS 中实现鲁棒的多视图一致性是 **非平凡的**, 需要一个新的、完全可微分的、基于语言的框架, 在原生 3D 空间内显式地强制执行跨视图对齐。

为此, 提出了 **相机感知指代场 (CaRF)**, 一个新颖的 R3DGS 框架 (见图 1(b)), 旨在实现鲁棒的多视图一致性。本项目首先引入了高斯场相机编码 (GFCE), 它将相机参数嵌入到高斯特征空间中, 以捕获视图相关的线索, 并在高斯-查询匹配期间增强几何推理。通过注入显式的几何先验, GFCE 使模型能够解释不同视角下的空间变化, 并更准确地将视觉特征与语言线索对齐。在此基础上, 进一步提出了训练中配对视图监督 (ITPVS), 它在单个迭代中将每个选定的高斯投影到两个校准视图中, 以暴露跨视图差异并惩罚视图特定的过拟合。GFCE 和 ITPVS 相互补充, 其中 GFCE 为可靠的视图锚定提供几何上下文, 而 ITPVS 通过配对视图优化强制执行一致性, 使 CaRF 能够在 3D 高斯场中实现视图一致的指代分割。

在三个基准数据集上的广泛实验表明, CaRF 在 Ref-LERF、LERF-OVS 和 3D-OVS

数据集上, 相对于最先进的方法, mIoU 平均分别提高了 16.8%、4.3% 和 2.0%。此工作首次对 R3DGS 进行了系统分析, 并为高斯场中的 3D 语言锚定建立了视图一致的基础。

本项目的贡献可以总结如下:

- 对 R3DGS 进行了全面的实证分析, 揭示了现有流程的基本局限性, 最显著的是它们对视图特定伪监督的依赖, 这导致了跨视图的预测不一致。
- 提出了高斯场相机编码 (GFCE), 一种几何条件化机制, 将相机姿态纳入高斯-文本交互, 建模视图相关的变化并增强指代场的几何保真度。
- 提出了训练中配对视图监督 (ITPVS), 一种可微分的 3D 原生策略, 在校准视图间对齐每个高斯的逻辑值, 减轻单视图过拟合并实现一致的多视图优化, 而不依赖于 2D 掩码。

2 相关工作

在本节中, 介绍先前基于 3DGS 的分割方法以及最近的 R3DGS 方法。

2.1 3D 神经表示

3D 表示的最新发展取得了显著进展, 其中神经辐射场 (NeRF) [22] 因其生成高质量新颖视图合成的能力而脱颖而出。尽管其有效, 但 NeRF 对隐式神经网络的依赖可能导致训练和渲染时间延长。为了解决这个问题, 基于显式表示的方法, 如体素和点云, 已经出现。最近, 3D 高斯泼溅 (3DGS) [16] 提出了一种使用一组 3D 高斯分布来表示 3D 场景的显式方法。通过采用 3D 高斯的快速可微分栅格化而不是体积渲染, 该方法实现了实时、高质量的渲染。自从 3DGS 引入以来, 其卓越的性能吸引了越来越多的关注, 导致了許多研究专注于其众多下游任务, 如高斯编辑 [2] 和 3D 语义分割 [24, 20]。由于其显式性质以及 2D 图像和 3D 场景之间的有效连接, 在这项工作中, 基于此范式, 特别关注 3D 点级的开放词汇理解。

2.2 高斯泼溅中的 3D 分割

w 与神经 3D 场景表示技术同时, 大型视觉语言模型 (VLM), 如 CLIP [26] 模型, 在各种下游任务的零样本迁移到广泛设置中表现出极强的能力。这一进展导致了开放词汇 3D 场景理解的研究, 其中神经场的神经渲染能力被用来将 VLM 与 3D 场景对齐。

因此, 几乎所有现有的 3DGS 相关语义分割工作都一致遵循基本范式: 首先利用强大的 VLM 从给定的多视图图像中提取语义特征, 然后通过训练附加到 3DGS 的语义或身份场来提升这些特征。一个主要分支是基于 2D 特征提取的方法 [24, 13], 旨在提取 2D 基础模型 [18, 19, 27] 中嵌入的语义知识, 并使用渲染的语义特征进行查询。LangSplat [24] 首先通过使用 SAM [18] 分割图像来实现这一点, 然后将生成的分层语义掩码输入 CLIP 以提取区域级语义嵌入。同时开发的 Feature3DGS [41] 使用 LSeg 和 SAM 提取 2D 特征, 并进一步使用 SAM 的解码器来解释 2D 渲染视图。另一个分支是基于 2D 掩码提升的方法 [37, 1, 10, 3, 35], 旨在从基础模型 (如 SAM) 提升 2D 掩码, 通常在跨视图一致性方面有所改进。GaussianGrouping 引入了一种对象关联技术作为预处理步骤, 以跨视图对齐 2D 分割图, 增强多视图一致性。此外, OpenGaussian [35] 通过利用 SAM 预测的掩码, 通过两阶段码本的粗到细特征离散化策略, 监督 3D 实例特征, 专注于 3D 开放词汇分割。虽然这些方法提供了一定程度的语言感知, 但它们主要关注类别级分割, 并且在理解任意和复杂的自然语言查询方面存在困难, 尤其是当涉及空间描述时, 这对于各种现实世界的应用 (如具身 AI、自动驾驶和 VR/AR 系统) 至关重要。

2.3 指代性 3D 高斯泼溅分割

R3DGS 将 2D 指代表达式分割 [6, 5] 和基于 3D 点的指代分割 [8, 34, 33] 的概念扩展到高斯泼溅领域, 使得能够在 3D 场景中对自由形式的自然语言查询进行查询。由 ReferSplat [9] 引入, R3DGS 旨在定位由语言表达描述的 3D 区域, 这些表达通常编码复杂的空间关系和细粒度的对象属性。ReferSplat 建立了一个空间感知的框架, 通过结合置信度加权的 IoU 掩码监督与高斯-文本对比学习, 在 Ref-LERF 数据集上实现了最先进的性能, 有效地将高斯特征与来自预训练大型语言模型的语言嵌入对齐, 以实现丰富的语义和空间理解。然而,

尽管有这些进步，其对单视图伪掩码监督的依赖限制了几何推理，并导致对特定视角的过拟合，从而导致空间一致性下降和跨视图泛化有限。

3 预备知识：符号与任务定义

3DGS [16] 将一个场景表示为一组 N 个各向异性高斯基元 $\mathcal{G} = \{G_i = (\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, \mathbf{c}_i, \alpha_i)\}_{i=1}^N$ ，其中 $\boldsymbol{\mu}_i \in \mathbb{R}^3$ 表示高斯中心， $\boldsymbol{\Sigma}_i \in \mathbb{R}^{3 \times 3}$ 其协方差， $\mathbf{c}_i \in \mathbb{R}^3$ 颜色，且 $\alpha_i \in [0, 1]$ 不透明度。每个高斯定义一个连续密度 $G_i(\mathbf{x}) = \alpha_i \exp(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)^\top \boldsymbol{\Sigma}_i^{-1}(\mathbf{x} - \boldsymbol{\mu}_i))$ 。给定校准的相机参数 $(\mathbf{K}, [\mathbf{R}|\mathbf{t}])$ ，每个高斯被投影到图像平面为 $\boldsymbol{\mu}'_i = \pi(\mathbf{K}(\mathbf{R}\boldsymbol{\mu}_i + \mathbf{t}))$ ，其中 $\pi(\cdot)$ 表示透视投影。像素 $\mathbf{p}$ 的渲染颜色通过 alpha 合成所有高斯获得：

$$\mathbf{C}(\mathbf{p}) = \sum_{i=1}^N T_i(\mathbf{p}) \alpha'_i(\mathbf{p}) \mathbf{c}_i, \quad (1)$$

其中 $T_i(\mathbf{p}) = \prod_{j < i} (1 - \alpha'_j(\mathbf{p}))$ 是累积透射率。高斯参数使用光度损失进行优化：

$$\mathcal{L}_{\text{photo}} = \sum_{\mathbf{p}} \|\mathbf{C}(\mathbf{p}) - \mathbf{I}_{\text{gt}}(\mathbf{p})\|_2^2, \quad (2)$$

这使得能够进行可微分的、实时的优化，以实现高保真度的场景重建。

R3DGS [9] 使用可学习的语义特征向量 $\mathbf{f}_i \in \mathbb{R}^d$ 扩展每个高斯 $\mathbf{G}_i$ ，形成一个语言场 $\mathcal{F} = \{\mathbf{f}_i\}_{i=1}^N$ 。给定一个自由形式的文本查询 q ，一个预训练的语言编码器产生一个上下文嵌入矩阵 $\mathbf{E} \in \mathbb{R}^{L \times d}$ ，其中 L 是词元数量。为了桥接语言语义和几何表示，引入了一个跨交互模块 $\phi(\cdot, \cdot)$ 来通过交叉交互细化高斯特征：

$$\mathbf{g}_i = \phi(\mathbf{f}_i, \mathbf{E}), \quad (3)$$

其中 $\mathbf{g}_i$ 表示增强的高斯特征嵌入。该模块使每个高斯能够选择性地关注描述其空间和语义属性的语言线索，从而加强语义锚定并提高视觉相似区域之间的可区分性。每个高斯的指

代分数通过聚合其与所有词嵌入的相似度获得：

$$m_i = \sum_j g_i^\top e_j, \quad (4)$$

其中 e_j 表示 $\mathbf{E}$ 中的第 j 个词嵌入。这些每个高斯的分数随后被渲染成 2D 指代掩码 $\mathbf{M}_{\text{pred}}$ ，通过二元交叉熵损失由伪真实掩码 $\mathbf{M}_{\text{gt}}$ 监督：

$$\mathcal{L}_{\text{BCE}} = \text{BCE}(\mathbf{M}_{\text{pred}}, \mathbf{M}_{\text{gt}}). \quad (5)$$

此外，ReferSplat [9] 引入了对象级对比学习目标以加强跨模态对齐。对于每个查询，选择具有前 τ 百分位相似度分数的高斯作为正样本，并且它们的平均特征 $\mathbf{f}_g$ 形成用于对比优化的高斯嵌入，可以表示为：

$$\mathcal{L}_{\text{con}} = \text{Con}(\mathbf{f}_g, \mathbf{e}_t), \quad (6)$$

其中 $\mathbf{e}_t$ 是相应的句子级嵌入。总训练损失结合了 Eqs. (5) and (6)。

尽管这种单视图监督有效地增强了跨模态对齐，但它缺乏跨校准视角的显式约束。因此，模型可能过拟合视图特定的伪掩码伪影（见图 2），导致不一致的多视图行为和降低的空间连贯性。本项目旨在解决这个问题。

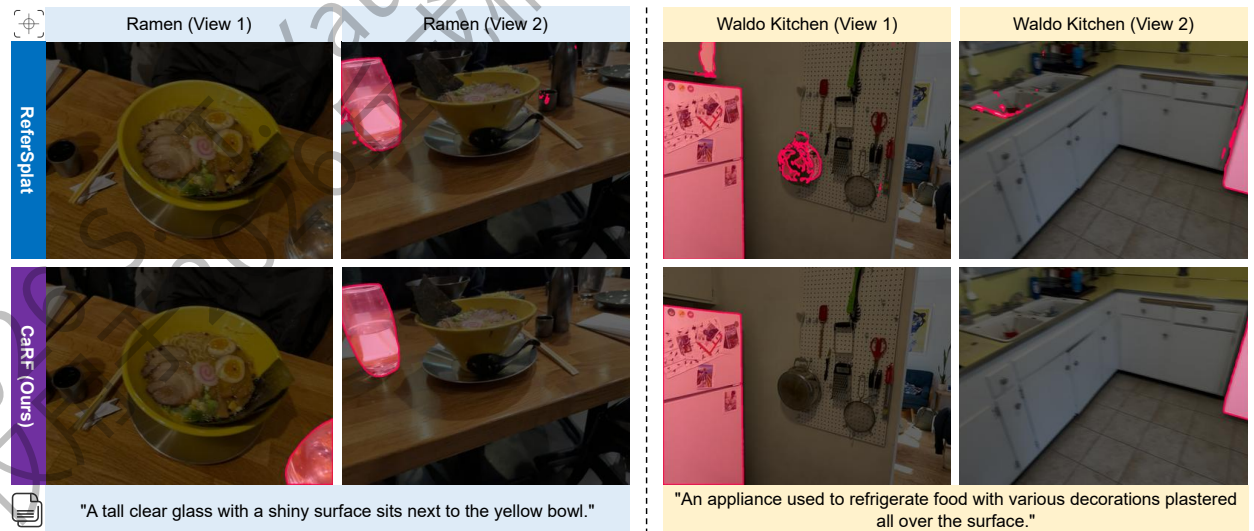


图 2: **ReferSplat** [9] 本项目 **CaRF** 在 *Ramen* 和 *Waldo Kitchen* 场景上的定性比较。ReferSplat 未能保持 **多视图一致性**: 在 *Ramen* 中, 水杯在视图 1 中完全丢失, 在视图 2 中显示出严重的渲染伪影; 在 *Waldo Kitchen* 中, ”冰箱”掩码泄漏到相邻区域并显示碎片化边界。相比之下, CaRF 在不同视图间产生连贯且几何一致的掩码, 有效保留了细粒度对象的细节。

4 CaRF: 相机感知指代场

本节介绍了提出的 CaRF 框架, 用于在 R3DGS 中实现视图一致的推理。首先介绍动机和整体架构, 然后描述其两个关键组件: 训练中配对视图监督 (ITPVS) 和高斯场相机编码 (GFCE)。最后, 概述了训练过程。

4.1 动机与框架概述

动机。 R3DGS 旨在定位与自由形式文本查询相对应的 3D 区域。尽管最近的方法如 ReferSplat [9] 通过跨模态对比学习取得了有希望的结果, 但它们在保持视图间一致性方面存在困难 (见 图 2 左侧)。因为它们的伪监督源自 2D 渲染, 模型倾向于过拟合视图特定的伪影, 并且缺乏对相同高斯在不同校准视图下应如何表现的约束。这导致不一致的掩码和降低的 3D 空间推理能力。为了克服这些限制, 本项目引入了 **CaRF**, 一个几何感知的 3D 原生框架, 直接在高斯场内强制执行语义和几何对齐。

核心思想。 核心思想是使指代过程在设计上视图一致: CaRF 不是依赖 2D 伪掩码, 而是联合学习多个校准视图, 并将每个高斯表示条件化于显式的相机几何。这使得能够端到端地优化一个统一的指代场, 该场保持跨视图连贯性和空间可解释性。

框架概述。 图 3 展示了 CaRF 框架, 这是一个用于 R3DGS 的视图一致范式。每个场景由一组校准的多视图 RGB 图像和一个文本查询表示。遵循先前工作 [9], 首先在 RGB 输入上训练一个标准的 3DGS 模型以恢复准确的场景几何。对于每个指代表达式, 使用 Grounded-SAM [28] 生成 K 个候选掩码, 每个掩码关联一个置信度分数。不直接选择置信度最高的掩码 (这通常不可靠), 而是计算所有候选掩码之间的置信度加权 IoU, 并选择具

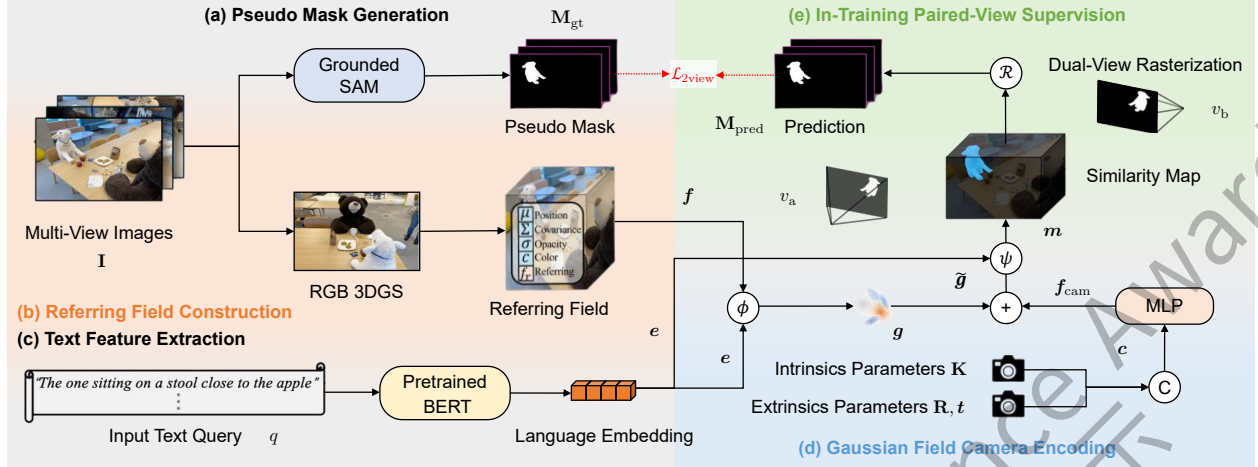


图 3: 提出的 CaRF 框架概述。给定校准的多视图 RGB 图像和一个文本查询，CaRF 首先通过置信度加权选择生成鲁棒的伪掩码。每个高斯通过一个与语言嵌入通过跨交互模块交互的指代特征进行增强。然后，相机感知编码集成了来自相机内参和外参的几何信息以增强视图一致性。最后，指代特征被栅格化为 2D 响应图，并在训练中监督下进行优化，以实现几何感知、视图一致的指代分割。

有最高聚合分数的掩码作为伪真实掩码 $\mathbf{M}_{gt}$ (见 图 3(a))。这个过程为后续训练产生了更空间连贯和鲁棒的监督。

如图 3(b) 所示，每个高斯 G_i 使用一个可学习的指代特征 $\mathbf{f}_i \in \mathbb{R}^D$ 进行扩展，形成一个集成语言和空间信息的语义场。同时，输入文本查询通过预训练的 BERT [12] 编码为每个词的嵌入 $\mathbf{e}$ (见 图 3(c))。一个跨交互模块 $\phi(\mathbf{f}_i, \mathbf{e})$ 融合了高斯和语言特征，加强了语义对齐并实现了细粒度的空间推理。随后，响应 m_i 通过一个相机感知编码模块（在 节 4.2 中详述）进行细化，该模块注入相机内参 $\mathbf{K}$ 和外参 $[\mathbf{R}|\mathbf{t}]$ 以编码视图相关的几何，从而减轻校准视图间不一致的监督（见 图 3(d)）。

最后，如图 3(e) 所示，CaRF 将指代特征栅格化为 2D 响应图 $\mathbf{M}_{pred}$ 而不是 RGB 颜色，并使用由 $\mathbf{M}_{gt}$ 引导的训练中监督。

在 节 4.2 and 4.3 中介绍的高斯场相机编码 (GFCE) 和训练中配对视图监督 (ITPVS) 模块共同使 CaRF 能够以端到端的方式在高斯表示中学习一个几何感知且视图一致的指代场。

4.2 高斯场相机编码

动机。 R3DGS 的一个基本挑战源于 3D 感知内在的视图依赖性。方法如 ReferSplat [9] 和 LangSplat [24] 主要依赖于隐式特征收敛或基于 2D 的一致性约束，而没有显式建模每个高斯在不同相机姿态下是如何投影的。因此，它们的指代特征通常在不同视图间编码不一致的证据，导致语义漂移和不稳定的优化。其他策略，包括 FastLGS [13]、econSG [40]、FMLGS [32] 和 OmniSeg3D [38]，通过多视图特征融合或图像空间正则化来缓解这个问题，但它们仍然对底层 3D 几何不可知，并且未能建模真正的空间对应关系。

为此，本项目提出了 **GFCE**，一个显式地将特征计算条件化于相机几何的模块。关键见解是使每个高斯 **视图感知**：其表示应适应当前相机的几何配置，同时在不同视角下保持一致的语义含义。这种显式的几何条件化实现了连贯的多视图推理，并增强了语言引导分割的鲁棒性。

相机参数编码。 每个校准相机由其内参矩阵 $\mathbf{K} \in \mathbb{R}^{3 \times 3}$ 和描述其相对于世界坐标系姿态的外参定义。后者由世界到相机的变换矩阵 $\mathbf{W} \in \mathbb{R}^{4 \times 4}$ 表示，该矩阵被分解为旋转矩阵 $\mathbf{R} = \mathbf{W}[:, 3:] \in \mathbb{R}^{3 \times 3}$ 和平移向量 $\mathbf{t} = \mathbf{W}[:, 3] \in \mathbb{R}^3$ 。首先通过连接向量化的旋转和平移来构建一个紧凑的 12 维姿态描述符：

$$\mathbf{c}_{\text{ext}} = [\text{vec}(\mathbf{R}); \mathbf{t}] \in \mathbb{R}^{12}, \quad (7)$$

其中 $\text{vec}(\cdot)$ 表示将 $\mathbf{R}$ 展平为 9 维向量的向量化操作。

虽然外参已经确定了相机在空间中的位置和方向，但它们没有捕捉到同一个 3D 点在不同成像条件下的投影方式。因此，本项目进一步将内参参数，包括焦距和主点坐标，纳入描述符。这些内参控制投影尺度、视野和光学中心，这对于建模视图相关的畸变和确保相机间的几何可比性至关重要。我将归一化的内参与外参描述符连接起来，形成统一的相机编码：

$$\mathbf{c} = [\mathbf{c}_{\text{ext}}; \text{norm}(\mathbf{K})], \quad (8)$$

随后通过一个 MLP 进行投影以获得相机感知嵌入。这种丰富的表示提供了相机校准及其

空间姿态的整体视图，允许每个高斯特征以精确的观察几何为条件，并提高视图间的一致性。

相机特征生成。 姿态描述符 $\mathbf{c}$ 通过一个轻量级多层感知机 (MLP) 映射到相机特征嵌入：

$$\mathbf{f}_{\text{cam}} = \text{MLP}_{\text{cam}}(\mathbf{c}) \in \mathbb{R}^d. \quad (9)$$

这个过程将视图相关的几何嵌入到一个潜在表示中，该表示可以以可微分的方式与语义特征交互。

视图感知的高斯-查询交互。 在高斯特征和文本嵌入之间进行跨交互之后，获得每个高斯的增强特征 $\mathbf{g}_i$ (见 Eq. (3))。然后，通过加性调制将相机编码注入到指代场中：

$$\tilde{\mathbf{g}}_i = \mathbf{g}_i + \mathbf{f}_{\text{cam}}, \quad (10)$$

这产生了一个相机条件化的特征 $\tilde{\mathbf{g}}_i$ ，它共同反映了语言、空间和几何信息。最终的高斯-语言相似度分数计算为

$$m_i = \psi(\tilde{\mathbf{g}}_i, \mathbf{E}) = \sum_j \tilde{\mathbf{g}}_i^\top \mathbf{e}_j, \quad (11)$$

其中 $\mathbf{e}_j$ 表示文本查询中第 j 个词的特征。通过将视图几何直接纳入特征计算，GFCE 使模型能够协调视角变化，保持语义一致性，并保持完全可微分以进行端到端优化。

优势。 在特征空间中编码相机参数避免了对不可微分重投影或掩码提升操作的需要，确保了平滑的梯度传播。加性集成策略平衡了稳定性和表达能力，引入了视图特定的偏置而不扭曲底层的语义表示。此外，将高斯特征条件化于姿态允许网络区分几何引起的外观变化与真实的语义差异，从而提高了 3D 语言锚定的可解释性和鲁棒性。

4.3 训练中配对视图监督

动机。 单视图监督没有对相同高斯在不同视角下应如何行为提供任何约束。因此，模型可能过拟合视图特定的伪影或伪掩码噪声，产生跨视图的不一致预测。现有方法如 Open-Gaussian [35] 和以实例为中心的聚合损失 [24, 37] 通过将掩码内像素或高斯拉向共享原型，同时推远不同掩码的高斯来解决特征对齐问题。虽然对于稳定的分割或聚类有效，但这种基于原型的训练对于 3D 高斯泼溅 (3DGS) 中的指代分割是次优的，原因有二：(i) 它压缩了查询相关的区别，迫使语义不同的实例（例如，“窗边的杯子”与“门附近的杯子”）尽管在查询上下文中不同，却共享一个单一原型；(ii) 它在掩码的 2D 特征空间中强制执行一致性，而不是直接在校准视图间监督相同的 3D 高斯，忽略了几何对齐。因此，当监督有噪声（例如，SAM 掩码、遮挡）或语言查询是关系型时，这些方法倾向于过度平滑特征并对查询特定、几何依赖的区别欠拟合。

为此，本项目引入了 **ITPVS**，它在单个高斯的级别上显式地强制执行多视图一致性。对于给定的文本查询 q ，相同的选定高斯被投影到两个共享重叠内容的校准视图中。通过联合损失耦合这两者，模型学习在尊重 3D 几何的同时，跨视角产生语义一致的预测。

双视图栅格化。 在通过位置感知的跨模态交互 [9] 获得高斯-文本相似度之后，选择与查询嵌入具有较高相似度的高斯进行渲染。与标准监督中在单个视图 v_a 上训练不同，选择另一个观察同一目标对象的校准视图 v_b ，并同时渲染两个视图。对于每个视图 $v \in \{v_a, v_b\}$ ，预测的指代掩码 $\mathbf{M}_{\text{pred}}^{(v)}$ 通过体积 alpha 合成生成：

$$\mathbf{M}_{\text{pred}}^{(v)} = \sum_{i=1}^{N_v} m_i \alpha_i(v) \prod_{k=1}^{i-1} (1 - \alpha_k(v)), \quad (12)$$

其中 m_i 表示来自 Eq. (11) 的高斯-语言相似度分数， $\alpha_i(v)$ 是第 i 个高斯在视图 v 中的不透明度， N_v 是该视图中可见高斯的数量。这种双视图栅格化显式地在多个校准视角下监督相同的 3D 高斯，在保持几何对应的同时实现跨视图的一致监督。

加权双视图 BCE 损失。 给定两个视图的渲染掩码，为每个计算二元交叉熵损失：

$$\mathcal{L}_{\text{bce}}^{(v)} = - \sum_p \left[\hat{y}_v \log y_v + (1 - \hat{y}_v) \log(1 - y_v) \right]. \quad (13)$$

其中 $\hat{y}_v$ 和 y_v 分别表示视图 v 在像素 p 处的真实掩码值和预测掩码值。为了联合优化两个视图，本项目使用固定权重 α 和 $(1 - \alpha)$ 组合它们，可以表示为：

$$\mathcal{L}_{2\text{view}} = \alpha \mathcal{L}_{\text{bce}}^{(v_a)} + (1 - \alpha) \mathcal{L}_{\text{bce}}^{(v_b)}. \quad (14)$$

这个复合目标耦合了两个视图的监督，允许来自一个视角的梯度信号正则化另一个，从而加强几何一致性。

优势。 ITPVS 利用高斯表示的显式 3D 性质来对齐多视图语义，而不依赖于昂贵的重投影或外部几何约束。通过在配对视图上监督相同的高斯，它有效地将真实的语义对应与视图相关的外观变化区分开来。这种方法不仅在有噪声监督下增强了跨视图稳定性和鲁棒性，而且加强了模型的几何可解释性，在整个训练过程中促进一致的 3D 语言锚定。

4.4 训练过程

CaRF 的整体训练联合优化了语义精度、几何一致性和跨模态对齐。如 ?? 1 所述，每次迭代都涉及配对视图监督和对比优化。具体来说，模型由配对视图损失 $\mathcal{L}_{2\text{view}}$ （见 Eq. (14)，在节 4.3 中介绍）引导以强制执行跨视图一致性，和一个对象级对比损失 $\mathcal{L}_{\text{con}}$ （见 Eq. (6)）以在共享语义空间中对齐高斯和文本嵌入。

总训练目标公式化为：

$$\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_{2\text{view}} + \lambda_2 \mathcal{L}_{\text{con}}, \quad (15)$$

其中 λ_1 和 λ_2 是平衡几何和语义目标的加权系数。这种统一的优化方案使 CaRF 能够实现具有一致 3D 锚定和跨多个视图的鲁棒语言引导推理的端到端学习。

Algorithm 1: CaRF 的训练过程

输入: 校准的多视图图像 $\{\mathbf{I}^{(v)}\}$ 及其 $(\mathbf{K}^{(v)}, \mathbf{W}^{(v)})$, 文本查询 q

输出: 高斯场景中的指代场参数

// 阶段 0: 预处理

- 1 在 $\{\mathbf{I}^{(v)}\}$ 上训练 RGB 3DGS 以恢复几何。
 - 2 对于每个查询 q , 通过 Grounded-SAM 获取 K 个候选掩码, 并使用置信度加权 IoU 选择 $\mathbf{M}_{\text{gt}}$ 。
 - // 阶段 1: 一次训练迭代
 - 3 采样一对校准视图 (v_a, v_b) 。
 - 4 将文本 q 编码为词嵌入 $\mathbf{E} = [\mathbf{e}_1, \dots, \mathbf{e}_L]$ 。
 - 5 对于每个高斯 i , 执行跨模态交互以获得 $\mathbf{g}_i = \phi(\mathbf{f}_i, \mathbf{E})$ 。
 - 6 为 v_a 和 v_b 生成相机描述符, 并通过 MLP 投影得到 $\mathbf{f}_{\text{cam}}^{(v_a)}$ 和 $\mathbf{f}_{\text{cam}}^{(v_b)}$ 。
 - 7 计算视图感知特征 $\tilde{\mathbf{g}}_i^{(v)} = \mathbf{g}_i + \mathbf{f}_{\text{cam}}^{(v)}$, 以及相应的指代分数 $m_i^{(v)}$ 。
 - 8 通过 alpha 合成渲染指代掩码 $\mathbf{M}_{\text{pred}}^{(v_a)}$ 和 $\mathbf{M}_{\text{pred}}^{(v_b)}$ 。 // 见 Eq. (12)
 - 9 计算双视图损失 $\mathcal{L}_{2\text{view}}$ 作为两个视图上 BCE 的加权和。 // 见 Eq. (14)
 - 10 通过 $m_i^{(v)}$ 选择前 τ 个高斯, 求平均形成 $\mathbf{f}_g$, 并计算对比损失 $\mathcal{L}_{\text{con}} = \text{Con}(\mathbf{f}_g, \mathbf{e}_t)$ 。
 - 11 形成总损失 $\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_{2\text{view}} + \lambda_2 \mathcal{L}_{\text{con}}$ 。
 - 12 通过反向传播更新网络参数。
-

5 实验

本项目进行了大量的实验来验证 CaRF 在 3DGS 中实现视图一致和语义精确的指代分割方面的有效性。本项目还提供了消融研究以分析每个提出组件的贡献, 以及定性可视化来说明几何和语言对齐。

5.1 实验设置

数据集。 在三个代表性基准上评估本项目提出的 CaRF 方法: **Ref-LERF**、**LERF-OVS** 和 **3D-OVS**, 它们共同涵盖了 3D 语言锚定中的多样化挑战。**Ref-LERF** 侧重于具有复杂空间关系和遮挡的场景特定指代表达式; 采用其官方数据分割, 并遵循 ReferSplat [9] 引入的置信度加权 IoU 伪掩码协议。**LERF-OVS** 将 3D 高斯泼溅扩展到跨多个场景的开放词汇分割; 通过与 LangSplat [24] 和 LangSplat-V2 对齐, 使用类别级文本短语进行查询来评估。**3D-OVS** 针对大规模、类别级和房间级的 3D 开放词汇分割, 如最近的工作 GAGS 和 OpenGaussian [35] 中所使用。这些数据集共同全面评估了 CaRF 在 (i) 细粒度空间推理、(ii) 开放词汇泛化和 (iii) 跨不同尺度和场景类型的鲁棒 3D 分割方面的能力。

评估指标 (mIoU)。 采用平均交并比 (mIoU) 作为主要评估指标, 遵循先前 3DGS-语言锚定的工作。对于每个查询-视图对, 本项目计算预测掩码 $\hat{Y}$ 和真实掩码 Y 之间的交并比为

$$\text{IoU}(Y, \hat{Y}) = \frac{|Y \cap \hat{Y}|}{|Y \cup \hat{Y}|}. \quad (16)$$

然后将得到的 IoU 值在 Ref-LERF 的所有查询和场景上平均, 或在 LERF-OVS 和 3D-OVS 的类别和房间子集上平均, 产生一个单一的 mIoU 分数, 反映分割精度和跨场景一致性。

实现细节。 首先预训练一个仅 RGB 的 3DGS 模型直至收敛, 然后在学习指代场之前冻结其几何参数 (μ, Σ , 不透明度和球谐函数), 遵循标准的 3DGS-语言实践以确保准确的可见性和不透明度监督。文本嵌入使用 BERT 提取, 跨交互 [9] 聚合每个高斯的词级响应以获得增强特征 g_i 。在训练期间, 本项目采样至少具有 30% 可见像素重叠的配对视图 (v_a, v_b); Eq. (14) 中的加权因子 α 默认为 0.5, 但可以选择性地根据可见比例进行缩放, 而不影响整体性能趋势。伪掩码的生成遵循 ReferSplat, 通过生成 K 个 SAM 衍生的候选掩码及其置信度分数, 并通过置信度加权 IoU 聚合选择最终掩码, 这比单分数选择提供了更大的稳定性。模型使用 Adam 优化器进行 30k 次迭代优化, 指代场和对比头的学习率为 2.5×10^{-3} , 相机 MLP 和门控参数的学习率为 1×10^{-4} 。特征维度设置为 $d = 128$, 使用混合精度训练和梯度裁剪为 1.0。所有实验均在 RTX A6000 GPU 上进行, 其他超参数 (例如, $\lambda_1 = 1$ 和 $\lambda_2 = 1$) 遵循 ReferSplat [9]。所有报告的结果是五次独立运行的平均值, 并且基线结果在相同设置下复现以进行公平比较。

5.2 在 Ref-Lerf 数据集上的结果

如表 1 所示, CaRF 在 Ref-Lerf 上设立了新的最先进水平, 将平均 mIoU 从 25.0 (ReferSplat) 提升到 29.2 (+16.8%), 在所有场景上取得一致增益: Ramen (+18.4%)、Figurines (+18.1%)、Teatime (+9.2%) 和 Kitchen (+22.9%)。按场景划分, 最大的相对改进出现在 *Kitchen* 和 *Ramen*, 其中强烈的遮挡、杂乱和精细结构使得单视图伪监督特别脆弱; 将高斯-文本相似度条件化于校准的相机几何 (GFCE) 减少了视图引起的外观混淆, 而配对视图监督 (ITPVS) 约束相同的高斯在重叠视角间保持一致, 遏制了溢出和边界伪影。*Figurines* 同样受益于存

表 1: 在 Ref-Lerf 数据集上与最先进方法在 **mIoU** ($\uparrow$) 方面的比较。更高的值表示更好的性能。粗体数字表示最佳结果。最后一行报告了 CaRF 相对于 ReferSplat 的相对改进 (%).

方法	发表	Ramen	Figurines	Teatime	Kitchen	平均
SPIIn-NeRF [23]	ICCV 2023	7.3	9.7	11.7	10.3	9.8
Grounded SAM [28]	arXiv 2024	14.1	16.0	16.9	16.2	15.8
LangSplat [24]	CVPR 2024	12.0	17.9	7.6	17.9	13.9
GS-Grouping [37]	ECCV 2024	27.9	8.6	14.8	6.3	14.4
GOI [25]	ACM MM 2024	27.1	16.5	22.9	15.7	20.6
ReferSplat [9]	ICML 2025	28.3	24.3	27.2	20.1	25.0
CaRF (本项目)	–	33.5	28.7	29.7	24.7	29.2
改进 (Δ , %)	–	+18.4	+18.1	+9.2	+22.9	+16.8

在相互遮挡和深度不连续情况下的几何感知匹配, 而 *Teatime* 显示出较小但稳定的增益, 可能是因为更简单的几何和更大的对象已经产生更强的伪掩码, 留下了更少的改进空间。总的来说, 表 1 中的一致改进支持了这样的主张: 将单视图监督转换为相机感知、配对视图训练对于稳定指代 3DGS 中的多视图推理是有效的。

5.3 在 3D 开放词汇分割数据集上相对于最先进技术的结果

如表 2 对于 LERF-OVS 数据集和表 3 对于 3D-OVS 数据集所示, CaRF 在所有子集上持续超越先前的 3DGS 语言方法, 证明了其在开放词汇理解和几何感知推理方面的有效性。

在 LERF-OVS 基准测试中, CaRF 实现了 54.9 的平均 mIoU, 比 ReferSplat (52.6) 高出 4.3%。按场景划分, 最大的相对改进出现在 *Kitchen* (6.9%) 和 *Figurines* (4.7%), 其中视觉杂乱和遮挡使得跨视图一致性特别具有挑战性。这些增益验证了提出的相机感知条件化 (GFCE) 和配对视图监督 (ITPVS) 的有效性, 它们共同减少了伪掩码的模糊性并稳定了高斯-文本对应关系。在所有场景上的稳定改进, 包括简单情况如 *Teatime* (1.8%), 表明 CaRF 不仅在复杂空间配置下增强了鲁棒性, 而且在干净条件下保持了语义精度。

在 3D-OVS 数据集上, CaRF 达到了 94.7 mIoU, 比 ReferSplat 的 92.9 提高了 2.0%。尽管总体幅度似乎小于 LERF-OVS 上的幅度, 但它反映了所有五个类别的一致增强, 包括具

表 2: 在 LERF-OVS 数据集上与最先进方法在 **mIoU** ($\uparrow$) 方面的定量比较。更高的值表示更好的性能。粗体数字表示最佳结果。最后一行报告了 CaRF 相对于 ReferSplat 的相对改进 (%)。

方法	发表	Ramen	Figurines	Teatime	Kitchen	平均
Feature-3DGS [41]	CVPR 2024	43.7	58.8	40.5	39.6	45.6
LEGaussians [29]	CVPR 2024	46.0	60.3	40.8	39.4	46.6
LangSplat [24]	CVPR 2024	51.2	65.1	44.7	44.5	51.4
GS-Grouping [37]	ECCV 2024	45.5	60.9	40.0	38.7	46.3
GOI [25]	ACM MM 2024	52.6	63.7	44.5	41.4	50.6
ReferSplat [9]	ICML 2025	53.1	64.1	50.1	43.3	52.6
CaRF (本项目)	–	55.2	67.1	51.0	46.3	54.9
改进 (Δ , %)	–	+4.0	+4.7	+1.8	+6.9	+4.3

表 3: 在 3D-OVS 数据集上与最先进方法在 **mIoU** ($\uparrow$) 方面的定量比较。更高的值表示更好的性能。粗体数字表示最佳结果。最后一行报告了 CaRF 相对于 ReferSplat 的相对改进 (%)。

方法	发表	Bed	Bench	Room	Sofa	Lawn	平均
Feature-3DGS [41]	CVPR 2024	83.5	90.7	84.7	86.9	93.4	87.8
LEGaussians [29]	CVPR 2024	84.9	91.1	86.0	87.8	92.5	88.5
LangSplat [24]	CVPR 2024	92.5	94.2	94.1	90.0	96.1	93.4
GS-Grouping [37]	ECCV 2024	83.0	91.5	85.9	87.3	90.6	87.7
GOI [25]	ACM MM 2024	89.4	92.8	91.3	85.6	94.1	90.6
ReferSplat [9]	ICML 2025	90.2	93.8	94.1	90.8	95.5	92.9
CaRF (本项目)	–	92.1	94.2	96.8	93.2	97.3	94.7
改进 (Δ , %)	–	+2.1	+0.4	+2.9	+2.6	+1.9	+2.0

有挑战性的大规模室内场景 (*Room*, *Sofa*) 和开放环境 (*Lawn*)。最显著的增益出现在 *Room* (2.9%) 和 *Sofa* (2.6%)，其中准确的多视图推理对于在广泛空间范围内保持一致的分割至关重要。这些结果证实了 CaRF 的显式几何条件化和多视图训练有助于在开放词汇、场景尺度设置下实现更好的泛化。

总的来说，在 LERF-OVS 和 3D-OVS 上的结果表明，CaRF 有效地桥接了 3D 高斯泼溅中语言理解和几何锚定之间的差距。其在多样化场景和数据集上的一致改进突显了其对于现实世界开放词汇 3D 分割的鲁棒性和可扩展性。

表 4: 方法上的消融研究。ITPVS 和 GFCE 分别表示训练中配对视图监督和高斯场相机编码组件。

索引	组件		结果	
	ITPVS	GFCE	Ramen	Kitchen
基线	✗	✗	28.3	20.1
1	✓	✗	31.6	22.4
2	✗	✓	24.3	13.5
本项目	✓	✓	33.5	24.7

5.4 消融研究

本项目进行了消融研究以评估 CaRF 中的关键设计选择，包括 ITPVS 和 GFCE 的效果、高斯选择策略、相机编码融合设计以及训练视图数量。结果表明，ITPVS 和 GFCE 共同增强了多视图一致性和空间推理，余弦相似度比重排序提供了更稳定的选择，基于 MLP 的融合最好地集成了相机几何，并且 2 视图设置实现了最佳的精度-效率权衡。

模块有效性。 进行了消融实验以验证两个提出组件 ITPVS 和 GFCE 的贡献。如表 4 所示，加入 **ITPVS** (索引 1) 相比基线 (ReferSplat) 产生了一致的增益，在 *Ramen* 上提高了 3.3% mIoU，在 *Kitchen* 上提高了 2.3%。这证实了配对视图监督有效地减轻了对单视图伪影的过拟合并增强了训练期间的多视图一致性。

引入 **GFCE** (索引 2) 通过显式地将高斯特征条件化于相机参数，进一步改善了几何推理。该模块使模型能够在相似度聚合之前捕获视图相关的线索，当与 ITPVS 结合时，导致更好的空间对齐和语义连贯性。然而，当 GFCE 单独在单视图训练下使用时，性能显著下降到基线以下（在 *Ramen* 上为 24.3 对比 28.3，在 *Kitchen* 上为 13.5 对比 20.1）。这种下降是因为，没有配对视图监督，相机条件化的特征缺乏稳定的多视图对应关系，在特征空间中引入了模糊性和噪声。

当两个模块都集成时（**本项目的**），CaRF 实现了最佳整体性能，在 *Ramen* 和 *Kitchen* 上分别达到 33.5 和 24.7 mIoU。这些结果表明 ITPVS 和 GFCE 相互补充，其中 ITPVS 提供了跨视图的一致几何约束，而 GFCE 丰富了模型的视图感知表示，共同形成了一个用于跨视图连贯语言锚定的鲁棒框架。

表 5: 在 Ramen 和 Kitchen 场景上评估的高斯选择算法的消融研究。

方法	Ramen	Kitchen
LERF 相关性评分	31.2	23.2
余弦相似度	33.5	24.7

表 6: 相机编码融合设计的消融研究。

方法	Ramen	Kitchen
MLP	33.5	24.7
后融合	25.6	18.3
语言编码	28.3	22.4

高斯选择算法评估。 进一步评估了此项目框架中使用的高斯选择策略。如表 5 所示，我比较了默认的基于余弦相似度的选择与 LERF 风格的相关性评分机制。余弦相似度方法实现了最高性能，在 *Ramen* 上 mIoU 增益为 2.3%，在 *Kitchen* 上为 1.5%。

尽管简单，但纯粹依赖余弦相似度可能会降低具有接近相似度分数的高斯之间的可区分性。为了缓解这个问题，先前的工作 [15] 引入了一个重排序过程，通过将每个高斯的激活与规范参考特征进行比较来归一化相似度。具体来说，对于每个渲染的语言嵌入 $\mathbf{f}$ 和文本查询嵌入 $\mathbf{e}$ ，相关性分数计算为

$$r_i = \frac{\exp(\mathbf{f} \cdot \mathbf{e})}{\exp(\mathbf{f} \cdot \mathbf{e}) + \exp(\mathbf{f} \cdot \mathbf{f}_i^{\text{canon}})}, \quad (17)$$

其中 $(\cdot)$ 表示点积， $\mathbf{f}_i^{\text{canon}}$ 表示从集合 {“object”, “thing”, “stuff”, “texture”} 中选择的规范词元的 CLIP 嵌入。

然而，如定量结果所示，这种重排序过程在 *Ramen* 上降低了模型性能 5.6%，在 *Kitchen* 上降低了 4.5%。下降的原因是 Ref-LERF 数据集中的指代表达式通常包含宽泛或通用的术语，这些术语在语义上已经与这些规范词元重叠。因此，归一化抑制了相关的相似度激活并减弱了正确高斯的响应。这表明，虽然重排序在细粒度场景中可能提高选择性，但直接余弦相似度对于 3D 高斯场中的自由形式指代表达式仍然是更有效和鲁棒的度量。

表 7: 视图数量的消融研究。

方法	结果		效率
	Ramen	Kitchen	训练
2 视图	33.5	24.7	420.3/epoch
3 视图	33.7	23.1	630.5/epoch
4 视图	32.4	24.1	840.7/epoch

高斯场相机编码设计。 我研究了将相机几何融入高斯-文本交互的不同策略。如表 6 所示，比较了三种变体：(a) 基于 MLP 的融合，它将姿态参数连接起来并通过 MLP 投影到高斯特征维度；(b) 后融合，它在相似度计算之后添加相机特征；和 (c) 语言编码融合，它将相机特征附加到 BERT 编码的查询中以形成视图相关的文本嵌入。

其中，基于 MLP 的融合实现了最佳性能（在 *Ramen* 和 *Kitchen* 上分别为 33.5 和 24.7 mIoU），表明直接将高斯特征条件化于几何嵌入提供了编码视图相关先验的最有效方式。后融合导致性能明显下降，因为在相似度计算之后添加相机特征在特征交互期间移除了有意义的几何对应关系。语言编码方法表现中等但仍然较差，因为将几何线索合并到文本编码器中扰乱了语言语义并破坏了优化的稳定性。总之，本项目采用基于 MLP 的设计作为最终配置，它为鲁棒的多视图推理提供了几何意识和语言一致性的平衡集成。

视图数量的影响。 我研究了每次迭代训练视图数量 $V \in \{2, 3, 4\}$ 的影响，同时保持一致的视图间重叠和目标可见性。如表 7 所示，将 V 增加到超过两个提供了可忽略的性能改进，但显著增加了计算成本。

两个主要因素解释了这种行为。首先，来自额外视图的边际增益迅速减少，因为大多数新添加的视图主要贡献被遮挡或边界区域，其中跨相机共同可见的像素很少。由于跨视图监督主要在这些共同可见区域上有效，整体效益趋于稳定。其次，训练时间几乎随 V 线性增长，导致精度和效率之间不利的权衡。因此，采用 2 视图配置作为 ITPVS 的默认设置，因为它实现了性能和计算效率之间的最佳平衡。更高视图的配置仅保留用于高度遮挡的场景，其中额外的视角有意义地扩展了共同可见区域。

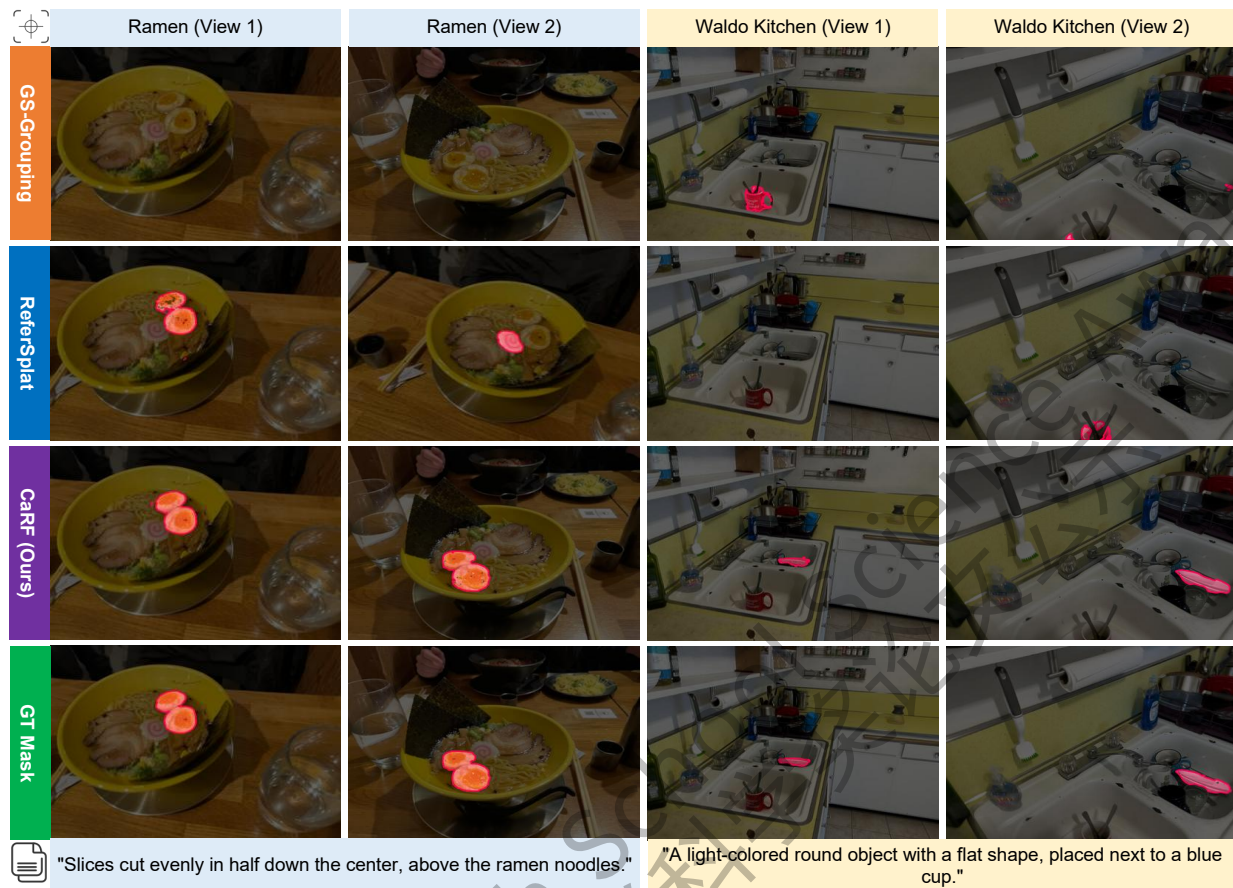


图 4: 在 Ref-LERF 数据集上跨两个代表性场景和两个校准视图的定性比较。每行显示来自 **GS-Grouping** [36]、**ReferSplat** [9]、本项目 **CaRF** 和 **GT** 掩码的结果。GS-Grouping 由于缺乏细粒度的空间推理，经常混淆外观相似的附近对象，而 ReferSplat 由于监督源自单视图伪掩码，产生视图不一致或不完整的分割。相比之下，CaRF 通过利用相机感知几何和配对视图一致性约束，跨多个视图生成空间连贯和语义精确的掩码。

5.5 可视化

除了图 2，本项目在图 4 and 5 中展示了来自 Ref-LERF 数据集的更多定性示例，将 CaRF 与最近的基线（包括 Gaussian Grouping 和 ReferSplat）进行比较。如图 4 所示，CaRF 提供了更准确和完整的分割，与文本描述紧密对齐。Gaussian Grouping 由于其依赖于类别级聚类先验，经常错误识别目标区域，而 ReferSplat 在不同场景间表现出不稳定的性能，在某些情况下产生清晰的结果，但在其他情况下产生不一致或碎片化的掩码。相比之下，CaRF 在多个视角下保持连贯的空间锚定，并捕获细粒度的对象边界，反映了其强大的跨视图推理以及语言和几何表示之间的对齐。在失败案例中（见图 5），例如 Ramen 和 Waldo Kitchen 场景，大多数错误源于不准确或模糊的伪真实掩码，而不是模型本身。在 Ramen 场景中，

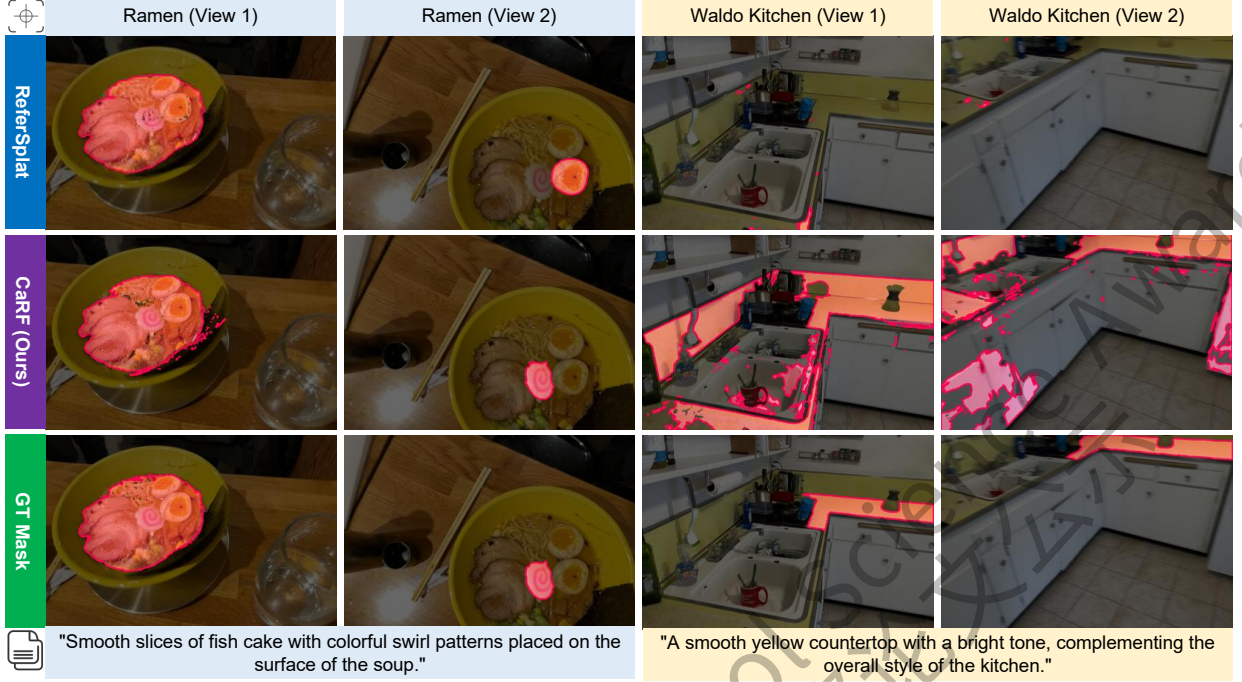


图 5: 在 Ref-LERF 数据集上跨两个代表性场景和两个校准视图的失败案例。每行显示来自 **ReferSplat** [9]、本项目的 **CaRF** 和 **GT** 掩码的结果。大多数错误源于不准确或未对齐的伪 **GT** 掩码 (见最后一行)。

鱼糕区域的伪掩码与周围的汤纹理重叠，而在 Waldo Kitchen 场景中，台面的掩码未能与真实对象边界对齐，并偶尔包含背景区域。尽管有这些不完美的标注，CaRF 仍然比基线产生更平滑和语义一致的分割 (见 图 5 右侧)，展示了对标签噪声和不完美监督的鲁棒性。

5.6 计算与精度之间的平衡

如图 6 所示，CaRF 在计算和精度之间实现了有利的平衡。额外的参数主要来自 GFCE 模块中的轻量级 MLP (见 Eq. (9))，它将相机姿态映射到 128 维潜在表示。这种紧凑的设计仅引入了 25.7K 额外参数，同时保持了整体流程不变。Eq. (14) 中提出的损失仅在训练期间应用，确保推理保持与 ReferSplat 中一样高效。因此，CaRF 不会产生额外的 FLOPs，因为它重用了相同的高斯渲染和合成过程。尽管有这种可忽略的计算开销，CaRF 在 Ref-LERF、LERF-OVS 和 3D-OVS 上持续超越基线，mIoU 分别提高了 16.8%、4.3% 和 2.0%。这些结果表明，通过相机感知编码融入几何先验显著增强了跨视图推理和语义锚定，同时保持了轻量级的推理效率。

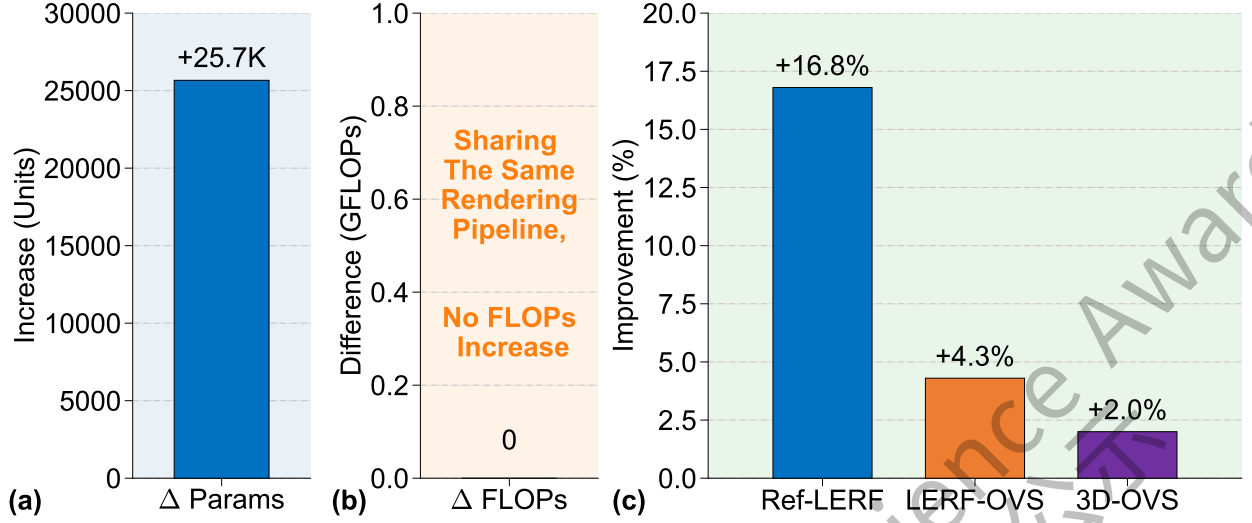


图 6: CaRF 与 ReferSplat [9] 的定量比较。(a) 参数增加 (Δ Params), 其中 CaRF 仅引入了 25.7K 参数的小额开销。(b) FLOPs 差异 (Δ FLOPs), 显示两种方法共享相同的渲染管道, 计算成本几乎相同。(c) 在三个基准测试上的性能改进 (Δ mIoU)。

6 结论与讨论

在本项目中, 提出了 CaRF, 一个用于 3D 高斯泼溅的相机感知指代场, 解决了 R3DGS 中多视图不一致的核心挑战。通过引入 ITPVS 和 GFCE, CaRF 在保持细粒度、语言引导的分割性能的同时, 强制执行视图间的几何一致性。在三个代表性数据集上的广泛实验表明, CaRF 在 mIoU 方面持续优于先前的方法, 为高斯场中的 3D 语言锚定建立了一个强大的最先进基线。除了其技术贡献, 这项工作还提高了 3D 语言接地的鲁棒性, 可能有益于下游应用, 如具身 AI、自主感知以及 AR/VR 环境中的人机交互。

尽管有效, CaRF 仍然与先前的 R3DGS 框架共享一些局限性。首先, 它依赖于伪掩码, 尚未纳入显式的 3D 视觉接地或关系约束, 这可能会增强空间推理和掩码质量。集成从大规模 3D 基础模型中提取的知识可能进一步提高其细粒度分割能力。其次, 与现有方法一样, CaRF 需要每个场景的优化。尽管 ITPVS 提高了训练稳定性, 但配对视图机制引入了额外的计算开销, 限制了在非常大或资源受限环境中的可扩展性。未来的工作将探索场景可泛化的训练策略和轻量级公式, 以扩展 CaRF 在动态和实时 3D 感知任务中的适用性。

参考文献

- [1] J. Cen, J. Fang, Z. Zhou, C. Yang, L. Xie, X. Zhang, W. Shen, and Q. Tian. Segment any 3d gaussians. In *AAAI*, 2025. [4](#)
- [2] Y. Chen, Z. Chen, C. Zhang, F. Wang, X. Yang, Y. Wang, Z. Cai, L. Yang, H. Liu, and G. Lin. Gaussianeditor: Swift and controllable 3d editing with gaussian splatting, 2023. [3](#)
- [3] S. Choi, H. Song, J. Kim, T. Kim, and H. Do. Click-gaussian: Interactive segmentation to any 3d gaussians, 2024. [1](#), [4](#)
- [4] S. Choi, H. Song, J. Kim, T. Kim, and H. Do. Click-gaussian: Interactive segmentation to any 3d gaussians. In *European Conference on Computer Vision*, pages 289–305. Springer, 2024. [1](#), [2](#)
- [5] H. Ding, C. Liu, S. He, X. Jiang, and C. C. Loy. Mevis: A large-scale benchmark for video segmentation with motion expressions, 2023. [4](#)
- [6] H. Ding, C. Liu, S. Wang, and X. Jiang. Vision-language transformer and query generation for referring segmentation, 2021. [4](#)
- [7] Q. Gu, A. Kuwajerwala, S. Morin, K. M. Jatavallabhula, B. Sen, A. Agarwal, C. Rivera, W. Paul, K. Ellis, R. Chellappa, C. Gan, C. M. de Melo, J. B. Tenenbaum, A. Torralba, F. Shkurti, and L. Paull. Conceptgraphs: Open-vocabulary 3d scene graphs for perception and planning, 2023. [1](#)
- [8] S. He, H. Ding, X. Jiang, and B. Wen. Segpoint: Segment any point cloud via large language model, 2024. [4](#)

-
- [9] S. He, G. Jie, C. Wang, Y. Zhou, S. Hu, G. Li, and H. Ding. ReferSplat: Referring segmentation in 3d gaussian splatting. In *International Conference on Machine Learning (ICML)*, 2025. [iii](#), [iv](#), [1](#), [4](#), [5](#), [6](#), [7](#), [9](#), [11](#), [13](#), [14](#), [15](#), [16](#), [20](#), [21](#), [22](#)
 - [10] S. Jain, S. Tulsiani, and A. Gupta. Gaussiancut: Open-vocabulary 3d scene segmentation via 3d gaussians. In *NeurIPS*, 2024. [4](#)
 - [11] K. M. Jatavallabhula, A. Kuwajerwala, Q. Gu, M. Omama, T. Chen, A. Maalouf, S. Li, G. Iyer, S. Saryazdi, N. Keetha, A. Tewari, J. B. Tenenbaum, C. M. de Melo, M. Krishna, L. Paull, F. Shkurti, and A. Torralba. Conceptfusion: Open-set multimodal 3d mapping, 2023. [1](#)
 - [12] G. Jawahar, B. Sagot, and D. Seddah. What does bert learn about the structure of language? In *ACL 2019-57th Annual Meeting of the Association for Computational Linguistics*, 2019. [8](#)
 - [13] Y. Ji, H. Zhu, J. Tang, W. Liu, Z. Zhang, X. Tan, and Y. Xie. Fastlgs: Speeding up language embedded gaussians with feature grid mapping. In *AAAI*, 2025. [1](#), [4](#), [9](#)
 - [14] Y. Jiang, C. Yu, T. Xie, X. Li, Y. Feng, H. Wang, M. Li, H. Lau, F. Gao, Y. Yang, and C. Jiang. Vr-gs: A physical dynamics-aware interactive gaussian splatting system in virtual reality, 2024. [1](#)
 - [15] K. Jun-Seong, G. Kim, K. Yu-Ji, Y.-C. F. Wang, J. Choe, and T.-H. Oh. Dr. splat: Directly referring 3d gaussian splatting via direct language embedding registration. In *CVPR*, 2025. [18](#)
 - [16] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis. 3d gaussian splatting for real-time radiance field rendering, 2023. [1](#), [3](#), [5](#)
 - [17] J. Kerr, C. M. Kim, K. Goldberg, A. Kanazawa, and M. Tancik. LERF: Language embedded radiance fields. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 19729–19739, 2023. [1](#)
 - [18] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick. Segment anything, 2023. [4](#)
 - [19] B. Li, K. Q. Weinberger, S. Belongie, V. Koltun, and R. Ranftl. Language-driven semantic segmentation, 2022. [4](#)

-
- [20] H. Li, R. Qin, Z. Zou, D. He, B. Li, B. Dai, D. Zhang, and J. Han. Langsurf: Language-embedded surface gaussians for 3d scene understanding. *arXiv*, 2024. [1](#), [3](#)
 - [21] K. Liu, F. Zhan, J. Zhang, M. Xu, Y. Yu, A. E. Saddik, C. Theobalt, E. Xing, and S. Lu. Weakly supervised 3d open-vocabulary segmentation, 2024. [1](#)
 - [22] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng. Nerf: Representing scenes as neural radiance fields for view synthesis, 2020. [3](#)
 - [23] A. Mirzaei, T. Aumentado-Armstrong, K. G. Derpanis, J. Kelly, M. A. Brubaker, I. Gilitschenski, and A. Levinshtein. Spin-nerf: Multiview segmentation and perceptual inpainting with neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20669–20679, 2023. [15](#)
 - [24] Z. Qin, Y. Huang, B. Chen, W. Ji, L. Li, C. Zhang, and et al. Langsplat: Language-driven open-vocabulary 3d gaussian splatting. In *CVPR*, 2024. [1](#), [3](#), [4](#), [9](#), [11](#), [13](#), [15](#), [16](#)
 - [25] Y. Qu, S. Dai, X. Li, J. Lin, L. Cao, S. Zhang, and R. Ji. Goi: Find 3d gaussians of interest with an optimizable open-vocabulary semantic-space hyperplane. In *ACM MM*, 2024. [15](#), [16](#)
 - [26] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision, 2021. [4](#)
 - [27] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson, E. Mintun, J. Pan, K. V. Alwala, N. Carion, C.-Y. Wu, R. Girshick, P. Dollár, and C. Feichtenhofer. Sam 2: Segment anything in images and videos, 2024. [4](#)
 - [28] T. Ren, S. Liu, A. Zeng, J. Lin, K. Li, H. Cao, J. Chen, X. Huang, Y. Chen, F. Yan, et al. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024. [7](#), [15](#)
 - [29] J.-C. Shi, M. Wang, H.-B. Duan, and S.-H. Guan. Language embedded 3d gaussians for open-vocabulary scene understanding. In *CVPR*, 2024. [16](#)

-
- [30] O. Shorinwa, J. Tucker, A. Smith, A. Swann, T. Chen, R. Firoozi, M. K. III, and M. Schwaiger. Splat-mover: Multi-stage, open-vocabulary robotic manipulation via editable gaussian splatting, 2024. [1](#)
 - [31] M. C. Silva, M. Dahaghin, M. Toso, and A. D. Bue. Contrastive gaussian clustering: Weakly supervised 3d scene segmentation, 2024. [2](#)
 - [32] X. Tan, Y. Ji, H. Zhu, and Y. Xie. Fmlgs: Fast multilevel language embedded gaussians for part-level interactive agents. *arXiv*, 2025. [1](#), [9](#)
 - [33] C. Wang, S. He, X. Fang, M. Wu, S.-K. Lam, and P. Tiwari. Taylor series-inspired local structure fitting network for few-shot point cloud semantic segmentation, 2025. [4](#)
 - [34] C. Wang, M. Wu, S.-K. Lam, X. Ning, S. Yu, R. Wang, W. Li, and T. Srikanthan. Gpsformer: A global perception and local structure fitting-based transformer for point cloud understanding, 2024. [4](#)
 - [35] J. Wu, C. Wu, Z. Fan, C. Sun, K. Zhang, Y. Wang, and et al. Opengaussian: Towards open-vocabulary 3d scene understanding with 3d gaussian splatting. In *NeurIPS*, 2024. [1](#), [2](#), [4](#), [11](#), [13](#)
 - [36] M. Ye, M. Danelljan, F. Yu, and L. Ke. Gaussian grouping: Segment and edit anything in 3d scenes. In *European conference on computer vision*, pages 162–179. Springer, 2024. [iii](#), [20](#)
 - [37] Q. Ye, Z. Li, S. Chen, S. Zhou, Y. Zheng, M. Geng, and et al. Gaussiangrouping: 3d gaussian splatting with instance-aware segmentation via multi-view association. In *ECCV*, 2024. [1](#), [2](#), [4](#), [11](#), [15](#), [16](#)
 - [38] H. Ying, Y. Yin, J. Zhang, F. Wang, T. Yu, R. Huang, and L. Fang. Omniseg3d: Omniversal 3d segmentation via hierarchical contrastive learning, 2023. [1](#), [9](#)
 - [39] H. Ying, Y. Yin, J. Zhang, F. Wang, T. Yu, R. Huang, and L. Fang. Omniseg3d: Omniversal 3d segmentation via hierarchical contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20612–20622, 2024. [2](#)
 - [40] C. Zhang and G. H. Lee. econsg: Efficient and multi-view consistent open-vocabulary 3d semantic gaussians. In *ICLR*, 2025. [1](#), [9](#)

-
- [41] X. Zhou, B. Shi, H. Xu, K. Zhang, S. Wang, F. Xiao, and et al. Feature3dgs: 3dgs with 2d feature distillation for 3d scene understanding. In *CVPR*, 2024. [1](#), [4](#), [16](#)
- [42] R. Zhu, S. Qiu, Z. Liu, K.-H. Hui, Q. Wu, P.-A. Heng, and C.-W. Fu. Rethinking end-to-end 2d to 3d scene segmentation in gaussian splatting. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3656–3665, 2025. [1](#)

致谢

本项目旨在解决指代表达三维高斯分割中的多视角不一致问题。我负责执行研究框架，包括构思研究思路、整理公开数据、设计方法、编写与修改代码、实施全部训练与评测，并撰写参赛报告。本项目不涉及人类受试者，实验使用公开的 Ref-LERF、LERF-OVS 和 3D-OVS 数据集以及公开基线代码。

本研究的选题来源于我对三维高斯泼溅（3DGS）及 ReferSplat 等工作的阅读、复现和比较。在复现实验中，我发现同一自然语言目标从不同相机视角观察时，分割结果容易发生漂移，遂将“如何在三维高斯空间中提升跨视角一致性”确定为研究问题。围绕这一问题，我完成了文献调研，下载并整理三套公开数据，核对相机参数、伪掩码和训练划分，设计了训练中成对视角监督（ITPVS）与高斯场相机编码（GFCE），配置环境并修改训练代码，制定基线与消融实验方案，汇总 mIoU 及多视角可视化结果并制作图表，独立撰写了参赛报告的摘要、引言、方法、实验、讨论和结论。

研究过程中遇到的第一个困难是基线环境依赖复杂、代码版本和张量维度不一致。我通过隔离环境、固定依赖版本、逐模块检查输入输出，并先在小场景上验证流程，使基线稳定复现。随后，成对视角训练出现遮挡差异、显存占用和视角对应不稳定等问题，我依据相机标定建立视角配对，在同一组高斯上对齐预测，并冻结 RGB 分支、调整损失权重和训练配置。面对伪掩码噪声及不同场景提升不均的情况，我通过三套数据集的基线比较、模块消融和定性结果核查定位问题，同时如实保留失败场景与局限。写作阶段则反复核对实验日志、表格和图像，确保结论与证据一致。

在研究过程中，谭鑫老师、谢源老师提供了学术建议并对论文进行审阅，帮助我理清技术逻辑、改进实验呈现与文字表述。本参赛报告的核心方法、主要实验及全部文稿均由我独立完成，两位老师未代替我完成核心方法、代码、实验或参赛报告的任何部分。此外，在参赛报告之外，本项目的研究成果另整理为一篇学术论文，由我与两位老师共同撰写完成，并已提交预印本平台及学术会议；该论文与本次参赛报告为两份不同稿件，特此说明。

感谢 ReferSplat 作者公开代码与数据，并感谢 3DGS、Grounded-SAM、BERT 等开源项目为本研究提供基础。除上述学术建议、论文审阅及公开资源外，无其他个人或机构代为完成数据获取、分析计算、程序开发、实验实施或参赛报告撰写；本项目亦未使用有偿科研代做或论文写作服务。

导师指导过程

· 2025 年 5 月：基础理论学习与范式梳理

- 系统学习 3D Gaussian Splatting 的核心流程、渲染方程、可微分体积渲染与 splat rasterizer 原理，以及可优化参数的构成。
- 梳理 3DGS 语义分割的三类范式（特征蒸馏、二维掩码提升、前馈推理），比较各自的优点与局限，为后续选题奠定基础。

· 2025 年 6 月：三维指代表达分割任务理解

- 系统理解三维指代表达分割（3D-RES）任务的定义、方法脉络与主要技术路线。
- 谭鑫老师、谢源老师就研究方向与问题定义提供学术建议。

· 2025 年 7 月：文献精读与方法设计

- 精读 ReferSplat 论文，完整理解 R3DGS 任务设定与 Referring Field 设计，并总结其依赖单视角伪掩码监督、缺乏显式多视角约束的局限。
- 独立提出并设计高斯场相机编码（GFCE），将相机内外参编码后显式注入 per-Gaussian 指代特征。
- 独立提出并设计训练中成对视角监督（ITPVS），通过双视图联合渲染与加权 BCE 损失在三维原生空间对齐 per-Gaussian logits。

· 2025 年 8 月：代码实现与实验验证

- 复现 ReferSplat 基线代码，修改训练流程，分别验证 GFCE 与 ITPVS 的有效性，并完成初步场景测试。

- 在 Ref-LERF、LERF-OVS 和 3D-OVS 三套数据集上完成全部对比实验。
- 开展模块消融、高斯选择算法、相机编码融合方式及视图数量等多组消融与超参数实验，并如实记录失败场景与方法局限。

· 2025 年 9 月 - 11 月：参赛报告撰写与论文审阅

- 独立撰写参赛报告的摘要、引言、方法、实验、讨论和结论，反复核对实验日志、表格与可视化结果，确保结论与证据一致。
- 谭鑫老师、谢源老师对论文进行审阅，在技术逻辑、实验呈现及文字表述方面提出修改意见。

尽管他们提供了重要帮助，谭鑫老师和谢源老师的全部指导均为无偿，未向本人收取任何费用。感谢两位老师的帮助！

导师简历

谭鑫老师 谭鑫，华东师范大学计算机科学与技术学院青年研究员（紫江青年学者）、院长助理，博士生导师，同时为上海人工智能实验室双聘副研究员。他于 2017 年至 2022 年就读于上海交通大学计算机科学与工程系，并参加上海交大—香港城市大学联合培养项目，先后取得两校博士学位；本科毕业于重庆大学自动化学院。谭老师的主要研究兴趣是通用场景理解与构建，已在 TPAMI、TIP、CVPR、ICCV 等国际知名期刊和会议发表论文 40 余篇。他主持了国家自然科学基金青年项目和重点项目课题、重庆市自然科学基金面上项目、CCF-腾讯犀牛鸟基金等，入选中国科协青年人才托举工程、中国图象图形学学会博士学位论文激励计划、上海市扬帆计划与晨光计划。他还担任中国图学学会可视化与认知计算专委会秘书、CCF 计算机辅助设计与图形学专委会执行委员，以及 Pattern Recognition、The Visual Computer 等期刊编委。

谢源老师 谢源，华东师范大学计算机科学与技术学院教授、博士生导师，国家优秀青年科学基金获得者，国家人工智能学院（上海）全时导师，教育部军民融合创新团队核心成员。他的研究方向涵盖计算机视觉与机器学习，包括基于深度学习的低层视觉、深度网络压缩、多视图学习与迁移学习等，近年聚焦跨模态大模型持续学习理论与方法以及持续策略驱动的具身智能。谢老师先后主持科技部科技创新 2030“新一代人工智能”重大项目课题、军委科技委国防创新特区项目及国家自然科学基金，在国内外权威期刊和顶级会议上发表学术论文百余篇，相关成果被引用超过 1.5 万次。他提出的图像复原算法获得 CVPR 2020 图像视频复原与增强挑战赛去雾和摩尔纹两个赛道冠军，曾任 IEEE TIP 编委及 IJCAI、CIKM、ACM MM 等顶会领域主席，相关成果获上海市科技进步特等奖、CSIG 科技进步一等奖、吴文俊人工智能科学技术奖自然科学二等奖等。