中图法分类号:TP391.41　　**文献标识码**:A　　**文章编号**:1006-8961(XXXX)XX-0001-18

论文引用格式:Li Lifen, Zhang Sijia, Zhang Ronghua. Survey on semantic expansion methods of 3D Gaussian splatting[J/OL]. Journal of Image and Graphics, XXXX:1-18. DOI: 10.11834/jig.250582. (李丽芬, 张思佳, 张荣华. 三维高斯溅射语义扩展研究综述[J/OL]. 中国图象图形学报, XXXX:1-18. DOI: 10.11834/jig.250582.)[DOI:10.11834/jig.250582]

三维高斯溅射语义扩展研究综述

李丽芬[1,2],张思佳[1],张荣华[1,3*]

1. 华北电力大学　计算机系,河北　保定　071003; 2. 华北电力大学　河北省能源电力知识计算重点实验室,河北　保定　071003; 3. 华北电力大学　复杂能源系统智能计算教育部工程研究中心,河北　保定　071003

摘　要:三维高斯溅射(3D Gaussian splatting, 3DGS)作为新一代三维场景表示与重建的核心技术,在实时渲染与高质量重建方面取得了巨大成功,但其固有的语义缺失限制了其在智能场景理解与交互式图形学等任务中的应用。为弥补语义鸿沟,将视觉语言模型(vision-language models ,VLMs)多模态感知技术与3DGS框架深度融合成为研究前沿。尽管相关技术发展迅速,但现有研究缺乏系统性的技术分类与梳理。本文首次提出了一种基于语义信息流的三阶段分类框架,将3DGS语义扩展流程解构为语义感知、语义绑定与语义查询三个核心阶段,为系统性地解决开放词汇场景泛化难、动态语义一致性差及实时交互受限等核心问题提供了统一的分析视角。首先阐述3DGS基本原理与语义表征理论;然后,基于所提框架深入剖析了各阶段的代表性方法、核心挑战和技术路线;最后,总结了当前研究面临的技术挑战并展望了未来方向。

关键词:三维高斯溅射;视觉语言模型;语义感知;语义绑定;语义查询;开放词汇

Survey on semantic expansion methods of 3D Gaussian splatting

Li Lifen[1,2], Zhang Sijia[1], Zhang Ronghua[1,3*]

1. Department of Computer, North China Electric Power University, Baoding 071003 China; 2. Hebei Key Laboratory of Knowledge Computing for Energy & Power, Baoding 071003 China; 3. Engineering Research Center of Intelligent Computing for Complex Energy Systems, Ministry of Education, Baoding 071003 China

Abstract: 3D Gaussian splatting (3DGS) has recently emerged as an efficient scene representation paradigm, enabling real-time rendering of complex 3D environments. By modeling scenes using anisotropic Gaussian primitives and optimizing them through differentiable rasterization, 3DGS achieves high rendering quality and computational efficiency compared with neural radiance fields (NeRF). Despite these advantages, the standard 3DGS pipeline remains primarily appearance-driven and lacks explicit semantic understanding, which limits its utility in intelligent vision tasks. To address the issue of fragmented research in this field, this paper proposes a three-stage taxonomy of semantic information flow, categorizing relevant methods into semantic perception, semantic binding, and semantic querying. This study systematically reviews over 80 representative papers published between 2023 and 2026, comparing their data types, algorithmic structures, and performance metrics such as mIoU (mean intersection over union) and rendering speeds. The first stage, semantic perception, focuses on acquiring language-aligned semantic features. Most existing works rely on pretrained vision-language models (VLMs) such as contrastive language-image pre-training (CLIP), bootstrapping language-image pre-training (BLIP),

收稿日期:2025-11-18;**修回日期**:2026-03-18

*通信作者:张荣华 zronghua88@aliyun.com

基金项目:中央高校基本科研业务费专项资金资助(2020MS122)

Supported by: Fundamental Research Funds for the Central Universities (Grant No. 2020MS122)

OpenCLIP, and sigmoid language-image pre-training (SigLIP) to extract representations. These models provide powerful open-vocabulary capabilities and ensure generalization across diverse scenes. Beyond direct feature extraction, advanced methods incorporate geometric cues to enhance cross-view consistency. Depth maps, estimated surface normals, and epipolar constraints are utilized to guide semantic alignment across camera viewpoints. Meanwhile, cross-view self-distillation further reduces semantic noise by encouraging feature consistency in overlapping regions. Through contrastive learning, multi-view fusion, and latent space alignment, semantic perception modules construct a unified embedding manifold that ensures both visual coherence and linguistic flexibility, forming the foundation for subsequent semantic reasoning. The second stage, semantic binding, addresses how semantic information is encoded, organized, and embedded within the continuous Gaussian representation. In contemporary literature, three dominant binding procedures have been emphasized: feature-distillation-based binding, 2D-3D mask enhancement-based binding, and single-forward feedforward-network-based binding. 1) Feature-distillation-based binding leverages a teacher-student paradigm in which comprehensive semantic features extracted by pretrained VLMs (teacher) are distilled into the 3DGS representation (student). Typically, multi-view image features are projected or aggregated and used to supervise Gaussian-associated semantic descriptors via distillation losses that promote cross-view consistency. This approach benefits from the generalization capabilities of large-scale pre-trained models and reduces reliance on per-scene annotation, but it can inherit noisy or biased cues from the teacher and requires careful scheduling to avoid semantic drift. 2) 2D-3D mask enhancement-based binding emphasizes mask-level alignment between 2D renderings and 3D Gaussians. In this pipeline, rendered feature maps or preliminary 2D predictions are converted into semantic masks which are then lifted into 3D via visibility-aware back-projection or probabilistic fusion; conversely, 3D-consistent priors are used to refine 2D masks in an iterative enhancement loop. This bidirectional 2D-3D mask enhancement improves localization and reduces occlusion-induced artifacts, yielding well-defined instance boundaries and per-Gaussian semantics. The method is effective when high-precision 2D segmentations are available or can be refined jointly, though it may depend on accurate rendering and mask alignment strategies. 3) Single-forward feedforward-network-based binding uses a dedicated lightweight network that, in a single forward pass, predicts per-Gaussian semantic descriptors from rendered multi-view features or geometric cues. This one-shot approach prioritizes runtime efficiency and deterministic behavior: given current Gaussian parameters and rendered observations, the feedforward module outputs semantic vectors that are attached to Gaussians without iterative optimization. Such architectures are beneficial for real-time or interactive applications due to their low latency and fixed memory footprint, but they require carefully designed architectures and training procedures to generalize across scenes and maintain cross-view coherence. Collectively, these binding procedures present different trade-offs among semantic fidelity, computational cost, and implementation complexity. Feature distillation leverages extensive pretrained priors, mask-enhancement enforces explicit geometric-semantic consistency, and one-shot feedforward binding prioritizes efficiency—each addressing different practical constraints encountered in semantic 3DGS. The third stage, semantic querying, explores how the enriched semantic information within 3DGS is efficiently retrieved, manipulated, or interpreted. Current research mainly differentiates between 2D-space querying and 3D-space querying. In the 2D-space paradigm, methods such as LangSplat, LEGaussians and Feature Splatting compute similarity between text embeddings and rendered semantic feature maps. Relevance heatmaps are generated and subsequently converted into semantic masks through thresholding. Further advancements, such as those introduced in SLGaussian and econSG, incorporate cross-view memory banks or language feature caches, promoting querying speed and stability in open-vocabulary segmentation. GOI proposes an optimizable semantic hyperplane that automatically adapts similarity thresholds, making semantic querying differentiable and adaptive across scene content. In the 3D-space paradigm, methods operate directly within the Gaussian latent space. Techniques such as OpenGaussian, Dr. Splat, and seconGS compute text-Gaussian similarity volumetrically, enabling multi-scale retrieval of semantic entities in 3D space. Some methods introduce re-ranking mechanisms or transformer-based refinement to enhance discrimination among visually similar objects. By performing querying in 3D, these approaches avoid artifacts caused by partial occlusions and inconsistent view coverage, enabling more geometry-consistent semantic reasoning. Beyond static scenes, recent work has begun extending semantic expansion to dynamic 3D Gaussian Splatting, which incorporates temporal coherence and motion-aware semantic propagation. Dynamic semantic representations enable the modeling of evolving object states and temporal rela-

tionships, supporting applications such as semantic tracking, action-driven editing, and language-guided dynamic scene interaction. Although still in an early stage, this direction is expected to drive the development of unified 4D semantic scene understanding frameworks. In summary, this paper provides a structured review of semantic expansion methods in 3D Gaussian splatting based on the proposed semantic information flow taxonomy. By analyzing existing methods through the three stages of semantic perception, semantic binding, and semantic querying, the review clarifies their underlying design principles and interdependencies. Key challenges—such as multi-view semantic alignment, embedding efficiency, dataset bias, open-vocabulary scalability, and the balance between semantic richness and real-time performance—are examined. The paper further discusses potential future directions, including cross-modal pretraining for Gaussian fields, lightweight semantic binding strategies for mobile applications, and hierarchical semantic querying. These analytical perspectives address current limitations in scene understanding and support the development of next-generation semantic-aware 3D vision systems.

Key words: 3D Gaussian splatting; vision-language models; semantic perception; semantic binding; semantic querying; open vocabulary

0 引　言

近年来，以神经辐射场（neural radiance fields，NeRF）（Mildenhall等，2020）为代表的神经渲染方法在新视图合成和三维场景重建领域取得了显著进展，极大地推动了计算机视觉与图形学在相关领域的发展，但其在实时渲染效率方面存在明显瓶颈。Kerbl等人（2023）提出的三维高斯溅射（3D Gaussian splatting，3DGS）技术，能够利用快速可微的渲染器进行渲染和优化，极大地提高了训练和推理速度以及渲染结果的保真度。然而，传统3DGS方法主要聚焦于几何结构与物体外观属性的重建，缺乏对深层语义信息的表征与推理能力，这使其在开放词汇的语义查询和交互式编辑等图形学任务中存在固有语义鸿沟。

与此同时，以CLIP（Radford等，2021）、ALIGN（Jia等，2021）等为代表的视觉语言模型（vision-language models，VLMs）（Zhou等，2022；Kim等，2022；Cherti等，2023；Zhai，2023；Rao，2022；Zhai，2023；Li，2023；Alayrac，2022；Liu，2023）在大规模图文对对比学习领域取得了突破性进展。通过将视觉特征与文本语义映射到统一的嵌入空间，为开放词汇下的语义感知奠定了基础。将其有效且一致地集成到3DGS框架中，以构建支持自然语言交互的三维场景表征，不仅是增强机器三维世界认知能力的关键，也是图像处理与计算机图形学交叉领域的核心问题。

三维场景语义化建模的早期研究主要围绕神经辐射场框架展开。Semantic-NeRF（Zhi等，2021）通过为每个三维点附加语义标签，开创了三维语义神经场建模的新方向。DFF（Kobayashi等，2022）与Panoptic Lifting方法（Siddiqui等，2023）分别通过特征蒸馏框架与几何外观语义场的联合优化策略，进一步推动了语义神经场的发展。LeRF（Kerr等，2023）首次将视觉语言先验嵌入神经辐射场框架，支持开放词汇查询任务。尽管如此，受限于NeRF的渲染效率，研究逐渐转向3DGS语义扩展。LangSplat（Qin等，2024）通过自动编码器压缩CLIP特征维度，将语义嵌入3DGS并构建层次化语义结构；LEGaussians（Shi等，2023）引入量化机制以优化语义特征压缩，有效缓解了视觉不一致导致的语义偏差；OpenGaussian（Wu等，2024）则结合掩码平滑约束与码本聚类实现实例级语义关联。这些方法推动了3DGS语义扩展研究发展。

尽管3DGS语义扩展研究发展迅速，但现有综述尚未建立系统的技术分类体系。目前已有一些相关综述文献从宏观技术原理及通用应用任务方面进行了总结。如Wu等人（2024）侧重于展现3DGS“能做什么”，从功能和应用角度进行方法分类；Bao等人（2025）则全面概述了3DGS的基础技术（如压缩、加速、正则化等），属于大而全的通用技术综述。He等人（2025）聚焦于分割、编辑和生成等具体下游任务，展示了其在实际应用中的多样性。Chen和Wang（2024）对3DGS的研究现状和未来趋势进行了全面总结，为该领域研究提供了理论基础。与之不同，本文聚焦于“语义信息流”的微观演进，即语义信息如何从多模态输入中提取（感知）、如何注入三维

几何载体(绑定)、以及如何被高效检索(查询),填补了语义一致性与跨模态对齐机制的分析空白。本文首次提出了一种基于语义信息流的三阶段分类框架,将复杂的3DGS语义扩展流程解构为语义感知、语义绑定与语义查询三个阶段,如图1所示。通过此框架,本文旨在提供一个统一的分析视角,系统梳理各阶段的核心挑战、技术路线与代表性方法,为该领域提供清晰的技术发展脉络和理论体系。

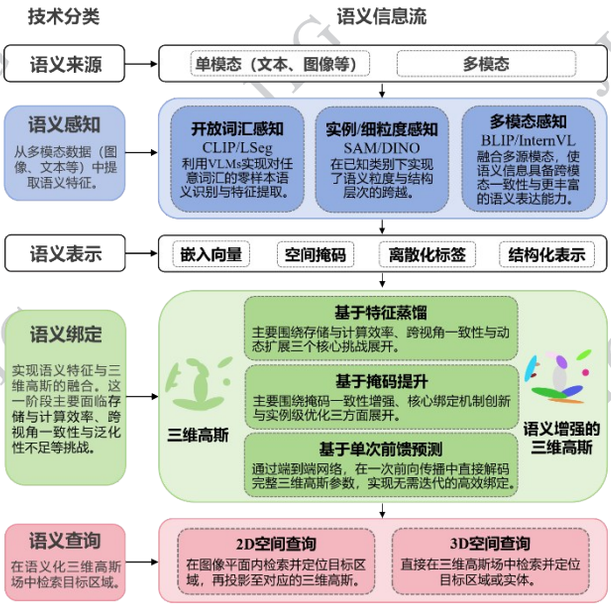


图1　基于语义信息流的语义扩展三阶段框架

Fig. 1　A three-stage framework for semantic augmentation via semantic information flow

1　理论基础

本节从语义扩展的目标和载体两个维度展开,为后续章节的分析奠定理论基础。

1.1　语义特征

语义扩展的核心目标是为三维场景赋予丰富、可被机器理解的多模态语义信息。这些信息以抽象特征的形式存在,称为语义特征。根据语义层次的不同,划分为基础语义与高层次语义两类。

1)基础语义主要描述物体或场景表层的、可直接感知的属性。

类别属性:指物体所属的基本类别,通常借助语义分割模型(如Mask R-CNN(He 等, 2017)、Swin-T(Liu 等, 2021)、SegFormer(Xie 等, 2021)、Mask-Former(Cheng 等, 2021)等)为像素或体素赋予类别标签。此类方法不仅能识别物体的类别,还进一步区分同一类别中的不同实例,实现实例级别的语义解析。

材质与光照属性:分别描述物体表面的物理材质特性和场景的光照条件,这些信息可结合三维高斯的颜色与透明度参数进行建模。

2)高层次语义建立在基础语义之上,描述物体和场景更为抽象的信息,如功能、空间关系和领域知识等。

对这些基础及高层次语义的有效感知,构成了三阶段框架中的第一阶段——语义感知。因此,如何从多模态数据中高效、准确地提取这些语义特征,是语义扩展流程的起点。

1.2　三维高斯溅射

3DGS(Kerbl 等,2023)将场景表达为一系列具有丰富参数的三维高斯,具体为

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

式中,均值$\boldsymbol{\mu}\in\boldsymbol{R}^{3}$表示三维高斯的中心位置,协方差矩阵$\boldsymbol{\Sigma}$控制三维高斯的大小以及旋转角度,通过旋转四元数$\boldsymbol{q}\in\boldsymbol{R}^{4}$和缩放向量$\boldsymbol{s}\in\boldsymbol{R}^{3}$进行重构表示和联合优化。

在渲染过程中,三维高斯首先被投影到二维平面,重叠的投影根据深度进行排序,从前往后依次进行可微α混合,最终得到的颜色表示为:

$$C^{j}=\sum_{i\in M}\boldsymbol{c}_{i}\alpha_{i}\prod_{j=1}^{i-1}(1-\alpha_{j}) \tag{2}$$

式中,$\boldsymbol{C}^{j}$表示第j个像素最终合成的颜色向量,$\boldsymbol{c}_{i}$是第i个三维高斯的颜色,α_{i}是第i个三维高斯投影的不透明度,它通过三维高斯的不透明度o与投影坐标得到,具体为:

$$\alpha=o\exp\left(-\frac{1}{2}(\boldsymbol{x}'-\boldsymbol{\mu}')^{T}\boldsymbol{\Sigma}'^{-1}(\boldsymbol{x}'-\boldsymbol{\mu}')\right) \tag{3}$$

式中,o是三维高斯的原始不透明度,$\boldsymbol{x}'$与$\boldsymbol{\mu}'$分别表示像素坐标及投影中心,$\boldsymbol{\Sigma}'$是二维协方差矩阵,具体为:

$$\boldsymbol{\Sigma}'=\boldsymbol{J}\boldsymbol{W}\boldsymbol{\Sigma}\boldsymbol{W}^{\mathrm{T}}\boldsymbol{J}^{\mathrm{T}} \tag{4}$$

式中,$\boldsymbol{W}$和$\boldsymbol{J}$分别表示视图变换矩阵和透视投影变换的仿射近似雅可比矩阵(Zwicker 等,2001)。

3DGS的显式表征特性使其在语义集成方面展现出显著优势,并为高效的语义查询与交互操作提供了几何驱动支撑。

在明确了语义特征与3DGS之后，后续章节将围绕如何实现语义特征与3DGS融合，即语义扩展展开。

2 三维高斯溅射语义扩展

本章将基于图1的语义信息流框架，分阶段系统阐述3DGS语义扩展的技术路线与研究进展。具体结构安排如下：

• 语义感知：重点论述从多模态数据（图像、文本等）中提取语义特征的技术演进脉络，并探讨相关方法的局限性以及未来发展趋势；

• 语义绑定：依据绑定的核心机制系统梳理主流技术范式，并分析各类范式为解决关键挑战采取的策略、内在贡献与固有局限；

• 语义查询：依据查询操作的空间维度，剖析从语义增强场景中响应开放词汇指令的实现机制。

通过对上述三阶段的系统分析，为理解不同方法的内在关联与演进逻辑提供统一视角。图2展示了各阶段代表性方法与模型之间的对应关系。

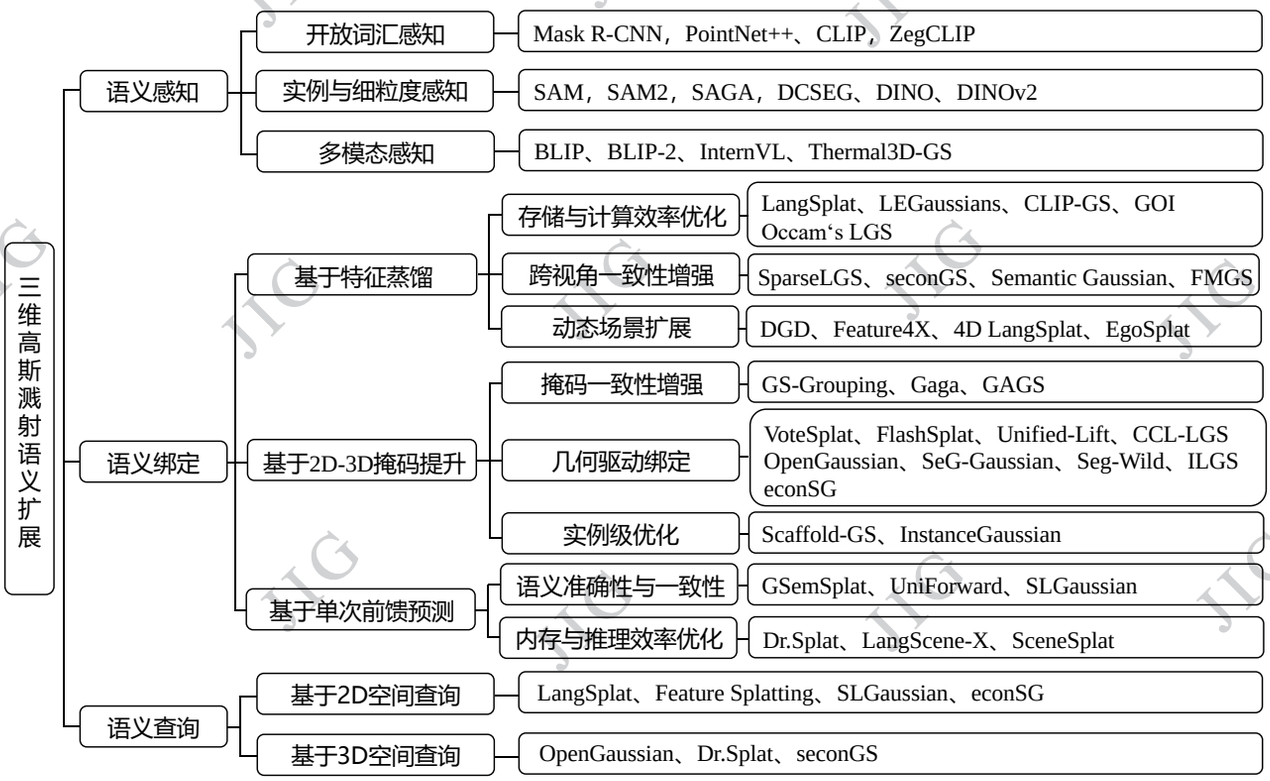


图2 3DGS语义扩展分类：语义感知、语义绑定与语义查询及其细化

Fig. 2 A taxonomy of semantic augmentation for 3DGS: perception, binding, and querying with subcategories

2.1 语义感知

语义感知是三阶段框架的起点，其核心目标是从多模态数据（如图像、文本等）中提取丰富且结构化的语义特征，为后续的3DGS语义绑定与查询奠定基础。如图3所示，语义感知技术的发展大致经历了开放词汇感知、实例与细粒度感知以及多模态感知三个阶段，体现了语义建模从封闭类别到开放词汇、从粗粒度到精细化、从单模态到多模态的范式转变。

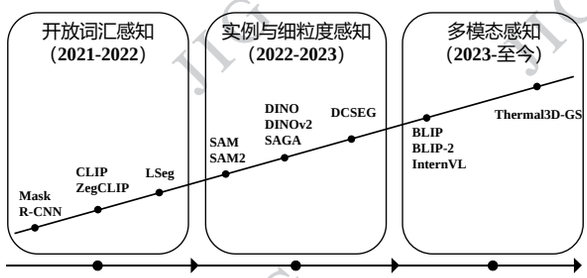


图3 语义感知方法技术演进脉络

Fig. 3 Evolution of semantic perception methods

2.1.1 开放词汇感知

早期研究主要依赖在封闭数据集上预训练的模型（如面向图像的Mask R-CNN（He 等，2017）与面向点云的PointNet++（Qi 等，2017））实现像素级与点级

的语义标签预测。尽管此类方法通过特征金字塔结构增强多尺度目标感知能力,但其语义泛化性受限,感知范围受预定义类别集约束,缺乏跨模态对齐机制,难以支持开放词汇的检索推理。

以CLIP(Radford 等,2021)为代表的VLMs通过大规模图文对对比学习,构建了统一的视觉语言嵌入空间,从根本上突破了封闭类别的限制,实现了零样本识别能力。在此基础上,ZegCLIP(Zhou 等,2022)与LSeg(Li 等,2022)进一步结合像素级解码器,实现了开放词汇下的语义分割,显著提高了视觉语义的细粒度感知能力。

2.1.2 实例与细粒度感知

为进一步提升语义粒度,该阶段技术主要分为提示驱动分割与自监督特征学习两类。

提示驱动分割以SAM(Kirillov 等,2023)及SAM2(Ravi 等,2024)为代表,通过点、框等交互式提示生成高质量的实例掩码。SAGA(Cen 等,2025)在此基础上引入层级实例分组,实现了"实例-部件"的层级语义建模;而DCSEG(Wiedmann 等,2025)则将开放词汇分割扩展至三维,通过重投影约束实现2D-3D语义迁移。

自监督特征学习方面,DINO(Caron 等,2021)与DINOv2(Oquab 等,2023)等模型通过无标注对比学习,提取出具备强几何结构感知能力的通用特征,为三维语义对齐与无标签学习提供了新的途径。王馨静等人(2026)提出了局部感知驱动的语义对齐策略,通过在特征提取阶段引入空间约束,缓解了CLIP全局特征在像素级任务中的语义偏移问题实现了从全局向局部的逻辑转变。

综上,实例与细粒度感知阶段实现了语义粒度与结构层次的跨越,为三维空间中的精细化语义建模奠定了基础。

2.1.3 多模态感知

随着场景复杂度的提升,单一视觉或文本模态难以满足语义理解需求。多模态感知通过整合图像、文本及其它模态信息,显著增强了系统的上下文理解能力。BLIP(Li 等,2022)、BLIP-2(Li 等,2023)与InternVL(Chen 等,2024)等多模态大模型采用统一编码架构,能捕捉更深层次的语义关联与跨模态指令理解能力。Thermal3D-GS(Chen 等,2024)进一步探索了红外与可见光的协同感知,为复杂环境(如夜间或烟雾场景)下的语义感知提供了新方向。针对精细边界捕捉与计算负担问题,贾迪等人(2025)提出频域引导的高效RGB-D分割网络,利用深度信息与高频特征校准边缘,为端侧环境下的高精度语义对齐提供了支撑。

尽管特定任务下的模型已初步实现轻量化,但多模态系统在向大规模3D场景扩展时仍普遍面临性能与开销的权衡难题。未来研究需进一步探索三维原生的轻量化融合机制,以构建全链路高效的语义感知体系。

2.2 语义绑定

本节系统梳理实现语义特征与3DGS融合的技术路线。根据其核心实现机制与演进范式,现有方法可划分为三类:1)基于特征蒸馏的方法;2)基于2D-3D掩码提升的方法;3)基于单次前馈预测的方法。下文将深入剖析每类方法面临的核心挑战、代表性工作及其内在贡献。

2.2.1 基于特征蒸馏的方法

基于特征蒸馏的方法为语义绑定奠定了基本框架,一般性流程如图4所示。该方法的核心是利用预训练的VLMs作为"教师模型",其提取的语义特征作为监督目标;通过可微分渲染,将3DGS的渲染输出与目标对齐,并利用蒸馏损失来优化三维高斯的语义属性。Feature 3DGS(Zhou 等,2023)首次通过特征场蒸馏将语义信息集成至3DGS。然而该范式在发展过程中面临存储效率、跨视角一致性与动态扩展三个核心挑战。围绕这些挑战,研究者们探索了不同的技术路径,形成了明晰的技术分支。

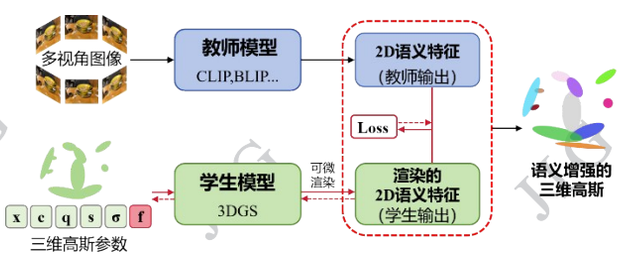


图4 基于特征蒸馏的3DGS语义绑定一般性流程图

Fig. 4 General pipeline of 3DGS semantic binding based on feature distillation

1) 存储与计算效率优化

VLMs产生的高维语义特征直接集成至3DGS会带来巨大存储与计算开销。针对此瓶颈,研究从数据优化与过程优化两个层面展开。

数据优化旨在直接降低语义信息的维度与冗

余。特征压缩是其主要路径，LangSplat（Qin 等，2023）和 LEGaussians（Shi 等，2023）分别通过场景自编码器和向量量化技术，将连续语义嵌入离散化或映射至低维潜空间，在保持语义表达能力的同时显著降低了内存占用。CLIP-GS（Liao 等，2024）引入语义属性紧凑性（semantic attribute compactness，SAC）机制，实现对象级特征的紧凑封装。GOI（Qu 等，2024）提出了基于特征码本的语义压缩策略，通过可学习的聚类码本，将高维语义特征映射为紧凑表示，在降低内存开销的同时保持了语义边界完整性。DF-3DGS（Dai 等，2025）通过颜色外观解耦的语义场蒸馏技术，在压缩语义特征的同时移除冗余外观参数，显著减少了三维高斯数量与存储需求。

过程优化则通过优化计算流程来提升效率。Occam's LGS（Cheng 等，2024）引入了一种免训练的全局优化方法，突破了传统依赖反向传播的思路，基于 3DGS 前向过程的概率公式规避了反向传播中的特征渲染开销，为实时应用提供了新思路。

上述效率优化策略从不同角度缓解了语义绑定的基础开销，但其优化过程也加剧了跨视角不一致性问题。

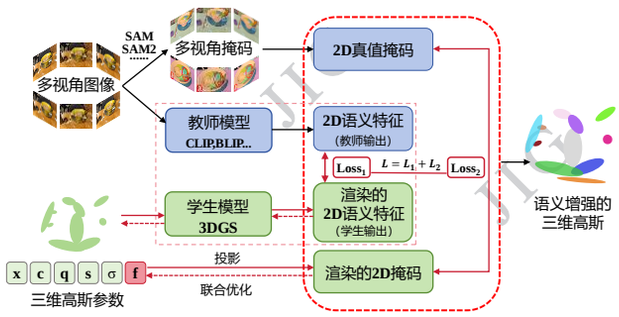
2） 跨视角语义一致性增强

语义一致性是指同一个三维实体在不同视角下应具有相同或语义连贯的特征表示。保证依赖 2D 监督的 3D 表示具备跨视角一致性是该范式的另一关键挑战。其技术进展分为两类。

显式约束策略通过在语义特征空间引入显式正则化来约束视角变化。SparseLGS（Hu 等，2024）通过建立 CLIP 特征与压缩特征间的双射关系，并结合三阶段多视图匹配策略保障稀疏视角下的一致性。seconGS（Yin 等，2025）利用 SAM2 生成跨视频帧一致掩码，并结合多帧平均策略降低动态场景的语义偏移。

映射机制改进策略通过优化 2D 到 3D 的映射过程约束一致性。Semantic-GS（Guo 等，2024）设计了一种通用的投影池化框架，支持开放词汇预测，在维持一致性的同时提高了泛化能力。FMGS（Zuo 等，2024）采用的多分辨率哈希编码（multi-resolution hash encoding，MHE）其编码结构具有空间连续性与平滑先验，能隐式促进跨视角一致性。

现有方法在多视角语义一致性方面取得了显著进展，但在极端稀疏视角下的鲁棒性仍是未来研究的重点。



3） 动态场景扩展

将语义绑定扩展至动态场景，核心挑战在于处理时序上的遮挡、位姿变化以及激增的计算开销。

DGD（Labe 等，2025）首次将时间维度引入 3DGS 框架，实现了外观、语义与几何属性的联合优化，提供了动态语义绑定的原型框架。Feature4X（Liu 等，2025）提出了一种高效的动态特征蒸馏方案，实现了图像与特征的高效同步渲染，显著降低了计算成本。同时，其权重插值机制有效保障了特征在时间维度上的平滑过渡，解决了时序一致性问题。4D LangSplat（Li 等，2025）将静态 CLIP 特征嵌入与多模态大语言模型（multimodal large language models，MLLMs）生成的动态文本描述相结合，通过状态可变形网络建模时间变化。EgoSplat（Li 等，2025）则针对自中心视角的频繁遮挡问题，提出了实例感知的时空瞬态预测机制以提升鲁棒性。然而，现有方法多依赖短程线性假设，且显式四维建模计算复杂。未来需探索更强大的时序架构和轻量化动态表征。

基于特征蒸馏的方法系统性地解决了语义绑定的基础问题，为整个领域奠定了基础。然而，其逐场景优化的范式在效率和泛化性上存在根本瓶颈。

2.2.2 基于 2D-3D 掩码提升的方法

基于特征蒸馏的绑定方法虽然在多视图间建立了稳定的语义，但受限于 2D 特征边界模糊与跨视角不一致等问题，其对象边界处精度不足。为此，研究开始利用 SAM 等 VLMs 生成的高质量实例掩码，将其作为强几何监督，通过多视图投影一致性将 2D 掩码提升至 3D 空间实现语义绑定，一般性流程如图 5 所示。其技术围绕掩码一致性增强、核心绑定机制创新与实例级优化三方面展开，以解决该范式面临的跨视角掩码不一致、绑定效率及泛化不足等挑战。

图 5 基于 2D-3D 掩码提升的 3DGS 语义绑定一

般性流程图

Fig. 5 General pipeline of 3DGS semantic binding with 2D - 3D mask enhancement

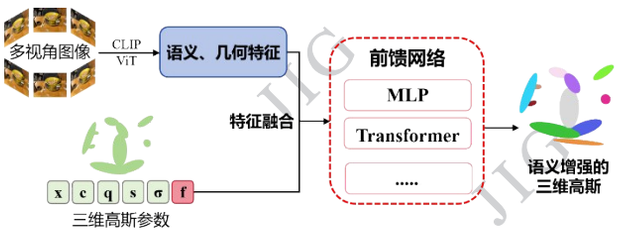
1) 掩码一致性增强

该类研究旨在提升初始掩码的跨视角一致性，为绑定提供更可靠的监督。GS-Grouping(Ye 等，2024)利用 3DGS 的可微渲染机制，通过交并比(intersection over union，IoU)筛选在多视角保持稳定投影关系的三维高斯，实现掩码的几何一致性过滤。Gaga(Lyu 等，2024)通过引入帧间光流信息与 3D 感知记忆库，将关键帧高质量掩码传播至相邻帧，有效缓解了动态场景下的遮挡与掩码断裂问题，提升了掩码完整性。GAGS(Wang 等，2024)则进一步提出多视图分割一致性损失，通过融合多粒度语义增强复杂场景中的掩码鲁棒性。

2) 几何驱动的绑定机制

该类研究通过创新绑定机制，降低逐场景优化方法的计算成本，并提升效率与泛化能力。VoteSplat(Jiang 等，2025)引入了基于 Hough 投票的快速绑定机制，通过学习三维高斯中心的偏移向量驱动语义信息向物体质心聚集，实现了无需迭代优化的实例绑定。FlashSplat(Shen 等，2024)则将掩码提升过程建模为线性规划问题，直接将 2D 掩码监督转化为 3D 掩码，无需迭代训练，显著降低了优化开销。

Unified-Lift(Zhu 等，2025)、OpenGaussian(Wu 等，2024)、CCL-LGS(Tian 等，2025)等通过特征码本机制实现结构化绑定，虽然仍需逐场景训练，但显著提升了训练和推理效率。此外，针对复杂场景的精细化语义建模，SeG-Gaussian (Zhang 等，2025) 引入了分割引导的优化策略，通过显式的物体分解将 2D 语义掩码转化为三维高斯的结构化约束，有效解决了物体边界模糊的问题；而 Seg-Wild(Bao 等，2025)则进一步聚焦于非受限场景(unconstrained collections)，利用 3DGS 建模外观与语义的不确定性，实现了在光照变化剧烈和遮挡严重场景下的鲁棒交互式分割。这些方法推动了掩码绑定从简单的几何投影向结构化、抗干扰的方向发展。econSG(Zhang 等，2025)提出基于置信区间引导的正则化方法(confidence-region regularization，CRR)，通过筛选高置信度区域增强跨视角一致性；ILGS(Jang 等，2025)则引入身份感知的语义一致性损失与掩码扩展机制，从损失函数层面强化绑定效果。



3) 实例级优化

针对分阶段绑定过程中可能存在的误差累积问题，该类研究侧重于对绑定结果进行全局优化，提升语义一致性与空间连贯性。Scaffold-GS(Lu 等，2023)是代表性工作，通过图优化平滑初始语义标签，有效纠正了离群点的错误标注。InstanceGaussian(Li 等，2025)构建了端到端框架，建立了一个更为通用、稳定的绑定范式，避免了分阶段优化的误差累积。

基于掩码提升的方法在语义边界精度、实例一致性与开放性方面取得了重要进展，但其性能依赖于 VLMs 的掩码质量，且逐场景优化的本质未变。

2.2.3 基于单次前馈预测的方法

前述基于特征蒸馏与掩码提升的方法虽然能获得高质量的语义绑定结果，但其逐场景优化模式计算密集且泛化能力受限。单次前馈预测范式为高效实现语义绑定提供了新思路，其一般性流程如图 6 所示。该方法通过构建端到端网络，在一次前向传播中直接解码完整的三维高斯参数，从而避免传统迭代优化的高成本。具体而言，多视图图像的视觉、语义及几何信息首先通过特征融合模块整合为统一的三维特征表示，常用的融合方式包括几何投影查询、多视图特征聚合或跨模态注意力等。随后，这些融合特征被输入前馈预测网络，该网络一般由多层感知机(multilayer perceptron，MLP)或轻量 Transformer 构成，可一次性预测三维高斯几何参数以及语义嵌入。尽管不同方法在具体结构上略有差异，但均遵循单次推理的统一范式，实现了快速且语义一致的 3DGS 绑定。其核心挑战在于如何保证生成结果在几何、外观与语义上的准确性及跨视角一致性。围绕上述挑战现有研究主要从以下技术路径展开探索。

图 6 基于单次前馈预测的 3DGS 语义绑定一般性流程图

Fig. 6 General pipeline of 3DGS semantic binding

via single forward feedforward network

1） 语义准确性与一致性增强

该类研究致力于在稀疏输入下提升语义标签识别的准确与跨视角一致性。GSemSplat（Wang 等，2024）与 UniForward（Tian 等，2025）分别通过并行解码器与双分支解耦解码器，显式分离几何、外观与语义属性的生成流程，从模型架构层面约束跨视角一致性。GaussianTR（Jiang 等，2024）通过特征对齐损失使网络输出与 CLIP 特征空间一致；SLGaussian（Chen 等，2025）创新性地引入“语言记忆库”模块，实现了从低维预测标签到高维语义特征的精准映射，增强了开放词汇能力。SemanticSplat（Li 等，2025）通过端到端训练直接预测各向异性三维高斯，实现了更高的语义一致性。

2） 内存与推理效率优化

作为面向应用的范式，高效率是其关键。Dr. Splat（Kim 等，2025）与 LangScene-X（Liu 等，2025）分别采用乘积量化（product quantization，PQ）技术和语言量化压缩器（language quantized compressor，LQC）技术，将高维 CLIP 特征压缩为紧凑的离散表示，大幅降低了内存占用与计算耗时。SceneSplat（Li 等，2025）提出了更高效的推理框架，在推理阶段仅依据三维高斯参数即可完成语义查询，无需实时图像输入，为实时交互应用提供了可能。

单次前馈预测范式代表了语义绑定走向通用化和实用化，虽然其在复杂场景下的精度与鲁棒性仍面临挑战，但展现出了巨大的发展潜力。

2.2.4 语义绑定方法的权衡与对比分析

基于上述三类绑定范式的技术原理，现有文献从计算效率、分割精度及泛化能力三个维度对其进行了广泛的实验验证与对比分析（Kerbl 等，2023；Li 等，2024；Zhang 等，2024；Wang 等，2024）。本节综合这些实验结果，总结各范式的核心权衡。

1）计算与存储开销

基于特征蒸馏的方法（如 Feature 3DGS）需要在每个场景的训练过程中为每个三维高斯存储高维语义特征（如 CLIP 的 768 维向量），导致显存占用相比纯几何 3DGS 增加 2 到 4 倍。虽然 LangSplat 通过场景自编码压缩了特征空间，但仍需额外的解码开销（Li 等，2024）。基于掩码提升的方法（如 GS-Grouping）利用离散的类别标识（category identification，ID）替代连续特征，显著降低了存储需求，但其依赖繁重的 2D 分割预处理（如 SAM 推理），增加了整体流程的时间成本（Zhang 等，2024）。基于单次前馈的方法（如 GSemSplat）彻底摒弃了逐场景优化，实现了毫秒级的实时推理，且无需存储高维特征，在端侧部署上具有显著优势（Wang 等，2024）。

2）语义分割精度与边界锐度

在标准数据集（如 LERF、Replica）上，特征蒸馏方法通常能获得较高的平均交并比（mean intersection over union，mIoU），但由于特征场的平滑插值特性，其预测结果在物体边缘处往往存在模糊现象，导致边界 F1 分数（boundary F1 score）相对较低（Li 等，2024）。掩码提升方法通过硬分配（hard assignment）实现了锐利的物体边界，SeG-Gaussian 的实验表明，其在复杂物体边界处的 mIoU 和 boundary F1 均优于特征蒸馏基线（Zhang 等，2025）。然而，该类方法的性能严重依赖于 2D 分割器的质量，容易受到分割噪声的影响，导致跨视角的语义不一致（Zhang 等，2024）。前馈预测方法通过学习通用的几何-语义先验，在处理遮挡和视角变化时表现出较好的鲁棒性，但在细粒度纹理的语义对齐上，其精度往往略低于针对特定场景优化的蒸馏方法（Wang 等，2024）。

3）泛化能力与鲁棒性

特征蒸馏和掩码提升方法均属于逐场景优化范式，难以直接泛化到未见过的场景。相比之下，单次前馈方法（如 GSemSplat、SceneSplat）展现了强大的零样本（zero-shot）泛化能力，能够在未见场景上直接推理出语义信息（Wang 等，2024；Li 等，2024）。然而，Seg-Wild（Bao 等，2025）的研究指出，在非受限场景（unconstrained collections）中，光照变化和动态遮挡会显著影响掩码提升方法的稳定性，而基于特征蒸馏的方法由于其连续表征特性，对这类噪声具有更强的鲁棒性（Bao 等，2025）。

2.3 语义查询

语义特征查询旨在从语义增强的三维高斯中，根据文本或图像等查询指令检索符合语义目标的三维实体或空间区域。其核心思想是计算查询指令的嵌入表示与三维高斯语义特征之间的相似度。根据特征匹配的操作空间域不同，现有方法可分为基于 2D 空间与基于 3D 空间两类的查询，分别如图 7 与图 8 所示。

2.3.1 基于 2D 空间的查询

该类方法首先将具有语义特征的三维高斯渲染

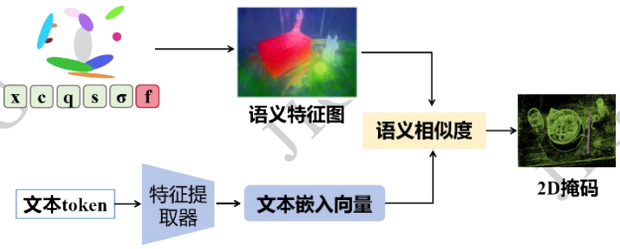


图7　基于2D空间查询的一般性流程图

Fig. 7　General pipeline of 2D space semantic querying

为二维特征图，并在图像空间内进行语义匹配。借助预训练VLMs的语义感知能力，实现像素级的高分辨率语义定位。

给定查询文本，经CLIP文本编码器得到特征向量$\boldsymbol{t}$，渲染得到的二维特征图I_s，以及逐像素的语义相似度图$\boldsymbol{S}$定义为：

$$S(x,y)=\left(\frac{t\cdot I_s(x,y)}{//\ t\ //\ \cdot\ //\ I_s(x,y)\ //}\right) \tag{5}$$

其中(x,y)为像素坐标。通过阈值函数生成二值掩码$\boldsymbol{M}$：

$$\boldsymbol{M}(x,y)=\begin{cases}1, & S(x,y)\geqslant\tau\\ 0, & S(x,y)<\tau\end{cases} \tag{6}$$

LangSplat（Qin等，2023）与LEGaussians（Shi等，2023）是该范式的代表。它们通过上述过程生成“语义相关性热力图”，并通过可学习阈值τ完成二值化与反投影，实现开放词汇条件下的查询。Feature Splatting（Qiu等，2024）引入可学习参数α将相似度映射为概率分布，实现更平滑的分割结果：

$$P(x,y)=\frac{\exp\left(\alpha\cdot S(x,y)\right)}{\Sigma_{x',y'}\exp\left(\alpha\cdot S(x',y')\right)} \tag{7}$$

GOI（Qu等，2024）将阈值选择问题转化为可优化的语义超平面（optimizable semantic hyperplane，OSH）学习任务，从而实现动态阈值判定。其决策函数定义为：

$$f(\boldsymbol{x})=\mathrm{sign}(\boldsymbol{w}\cdot\boldsymbol{x}+b) \tag{8}$$

其中，$\boldsymbol{w}$为可优化的权值向量，$\boldsymbol{x}$为输入的语义特征向量，b为偏置项。

SLGaussian（Chen等，2025）通过构建多视图语言记忆库$\boldsymbol{M}$，在查询阶段直接计算文本嵌入$\boldsymbol{t}$与记忆特征的相似度，实现高效语义检索。econSG（Zhang等，2025）则通过可学习阈值机制实现三维语义掩码生成。

该范式充分利用VLMs的语义表达能力，实现了高分辨率定位。然而，其性能受限于渲染过程，且查询结果与视角相关性强，质量高度依赖于语义绑定阶段的特征对齐效果。

2.3.2　基于3D空间的查询

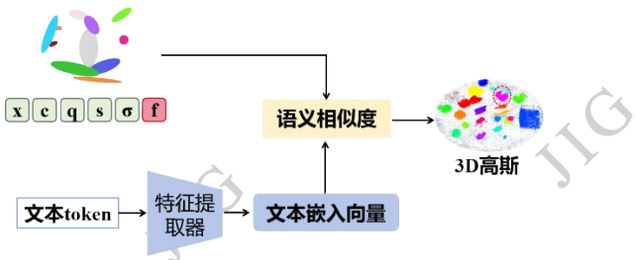


图8　基于3D空间查询的一般性流程图

Fig. 8　General pipeline of 3D space semantic querying

该类方法直接在3D空间中进行语义相似度计算与筛选，有效避免了二维投影过程的信息损失，更适用于全局场景检索及复杂空间结构下的语义查询任务。其核心计算公式为：

$$S_i=\frac{\boldsymbol{t}\cdot\boldsymbol{g}_i}{//\ \boldsymbol{t}\ //\cdot//\ \boldsymbol{g}_i\ //} \tag{9}$$

其中$\boldsymbol{t}$为查询文本特征向量，$\boldsymbol{g}_i$为第i个三维高斯的语义特征向量。随后根据预设阈值τ筛选出满足相似条件的三维高斯集合$\{\boldsymbol{g}_i\}$。

OpenGaussian（Wu等，2024）为该方向的代表性工作，其通过计算实例级语义向量与文本嵌入的相似度，实现三维开放词汇检索。然而，在复杂场景下，仅依赖余弦相似度易受特征分布重叠的影响，导致类别区分性指标不足。为此，Dr. Splat（Kim等，2025）引入了基于空间邻域的重排序机制。该方法首先计算初始相似度S_i，再结合邻域信息$\boldsymbol{N}_i$优对得分进行优化校准：

$$\tilde{S}_i=R(S_i,\boldsymbol{N}_i) \tag{10}$$

其中$\boldsymbol{N}_i$为第i个高斯点所属的时空掩码块（masklet）特征信息，$R(\cdot)$为语义校准与重排序函数。该机制利用了“同一物体的特征在不同视角下应保持稳定”的先验，通过跨视角追踪确保了物体在3D空间与时间维度的一致性。Dr. Splat（Kim等，2025）通过语义对比增强了特征的区分度，而seconGS（Yin等，2025）则通过真值锚定（GT-anchored）策略显著提升了点级查询的准确性，有效解决了全局阈值失效和局部语义噪声问题，从而增强了3D场景理解的鲁

棒性。

语义查询将高维语义表征转化为可交互的下游应用接口，形成语义信息流三阶段的完整技术闭环。

2.4 技术范式综合比较与权衡

基于前文对语义感知、语义绑定与语义查询三阶段的系统梳理，本节旨在构架一个多维度的综合比较框架，对现有代表性方法进行全局性的横向剖析与归纳。如表1所示，该框架不仅明确了各类方法在“语义信息流”各阶段的核心特征，更通过对比其核心技术机制、设计动机及性能表现，揭示了不同技术范式在语义保真度、计算效率、泛化能力与实现复杂度等核心维度上存在的根本性权衡。

三大主流语义绑定范式呈现出鲜明的特点与局限。基于特征蒸馏的范式（如LangSplat、LEGaussians）能够继承VLMs的强语义先验，但面临显存占用大、优化耗时的挑战；基于掩码提升的范式（如OpenGaussian、Seg-Gaussian）通过显式2D-3D几何对齐实现语义一致性，在实例级分割上表现优异，但其高度依赖2D分割模型的质量与鲁棒性；基于单次前馈预测的范式（如SLGaussian）通过前馈网络实现高效推理，在实时性方面优势显著，然而在跨场景泛化能力与多视角一致性方面仍存在提升空间。

通过系统性比较，从“语义信息流”的宏观视角揭示了3DGS语义扩展领域技术演进的内在逻辑与共性挑战。不仅为研究者理解和评估现有方法提供了清晰的坐标系，也为第4章深入探讨该领域面临的瓶颈与未来发展方向奠定了基础。

表1 3DGS语义扩展代表性方法的多维度特性综合对比

Table 1 Comprehensive multi-dimensional comparison of representative 3DGS semantic augmentation methods

绑定范式	文献	核心感知模型	核心创新/机制	特征对齐机制	语义粒度	查询空间	主要特性	核心挑战
基于特征蒸馏	Feature 3DGS（Zhou等，2023）	CLIP	逐场景梯度优化	插值特征场蒸馏	点/块级	2D	语义丰富支持细粒度查询	存储与计算开销大
	LangSplat（Qin等，2023）	CLIP+SAM	场景自编码压缩	分层语义光栅化	点级	2D	显著降低内存占用	压缩导致语义信息损失
	LEGaussians（Shi等，2023）	CLIP+DINO	离散码本量化	不确定性加权蒸馏	点级	2D	高压缩比缓解语义偏差	存在量化误差
	CLIP-GS（Liao等，2024）	CLIP+SAM	掩码引导语义紧凑化	对象级特征紧凑封装	点级	2D	提升对象级感知的清晰度	依赖掩码质量处理复杂背景难
	GOI（Qu等，2024）	CLIP+APE	可优化语义超平面	自适应阈值学习	点级	2D	查询鲁棒性强无需手动调参	超平面优化增加模型复杂性
	DF-3DGS（Dai等，2025）	LSeg	外观-语义解耦建模	解耦场联合优化	点级	2D	提升存储与渲染效率	解耦设计增加优化复杂度
	Occam’s LGS（Cheng等，2024）	CLIP	免训练全局优化	几何引导	点级	2D	避免反向传播计算效率高	对输入特征噪声敏感
	SpareLGS（Hu等，2024）	CLIP+SAM	显示约束双射关系建立	基于掩码的稀疏特征选择	点级	2D	聚焦关键区域减少冗余计算	稀疏性定义与通用性平衡困难
	seconGS（Yin等，2025）	SAM+SAM2	真值锚定	跨帧掩码	混合粒度	3D	支持动态场景时序语义一致性	动态场景下的掩码漂移问题
	Semantic-GS（Guo等，2024）	SAM+CLIP	掩码增强特征蒸馏	掩码引导特征聚合与对齐	点级	2D	结合实例感知开放词汇能力	计算流程复杂效率折中

表1续表

绑定范式	文献	核心感知模型	核心创新/机制	特征对齐机制	语义粒度	查询空间	主要特性	核心挑战
	FMGS（Zuo等，2024）	CLIP+DINO	多模态特征融合蒸馏	几何语义特征联合对齐	点级	2D	融合几何先验增强特征判别力	多模态特征对齐难度大
	DGD（Labe，2024）	SAM+CLIP+DINOv2	密集几何-语义蒸馏	多尺度几何语义对齐	点级	2D	几何感知强细节丰富	模型大训练推理成本高
	Feature-4X（Zhou等，2025）	SAM2+CLIP-Seg	4D动态特征场建模	时空一致性蒸馏	实例级	2D	支持4D场景实例级语义理解	4D数据获取标注困难
	4D LangSplat（Li等，2025）	CLIP+SAM	4D场景自编码压缩	时空分层语义光栅化	实例级	2D	扩展至动态场景保持高效压缩	动态下压缩质量与一致性
	EgoSplat（Li等，2025）	SAM2	第一人称视角感知绑定	视角自适应特征投影与对齐	点级	2D	针对第一人称视频优化实时性强	视角剧烈变化语义稳定性差
基于2D-3D掩码提升	GS-Grouping（Ye等，2024）	SAM	掩码反向投影	多视图多数投票	实例级	2D	实现分组关联	依赖准确实例分割结果
	Gaga（Lyu等，2024）	SAM	几何引导掩码聚合	几何感知掩码融合	点级	2D	几何信息提升掩码融合质量	对场景几何重建质量敏感
	SeG-Gaussian（Zhang等，2025）	SAM	分割引导几何分裂	语义-几何联合优化	实例级	3D	几何与语义显式协同优化	优化过程复杂易局部最优
	Seg-Wild（Bao等，2025）	SAM+DINO	不确定性建模	交互式引导	实例级	2D	支持弱监督交互式标注	需人工交互自动化程度低
	GAGS（Wang等，2024）	SAM+CLIP	几何感知掩码语义提升	几何约束掩码-语义对齐	混合粒度	2D	结合几何开放词汇语义	两阶段流程效率受限
	VoteSplat（Jiang等，2025）	SAM	多视图概率投票融合	概率掩码投票与加权融合	点级	2D	提升跨视角一致性	投票机制需要足够有效视角
	FlashSplat（Shen等，2024）	CLIP	快速掩码生成	轻量级特征渲染与掩码关联	点级	2D	追求速度	语义精度速度权衡
	Unified-Lift（Zhu等，2025）	SAM	统一掩码提升框架	通用掩码投影与优化模块	点级	2D	框架通用性强易扩展	性能依赖基础模块
	OpenGaussian（Wu等，2024）	SAM+CLIP	掩码平滑约束与码本聚类	掩码反向投影与3D聚类	实例级	3D	实例边界清晰几何一致性强	对遮挡敏感依赖SAM性能
	CCL-LGS（Tian等，2025）	CLIP	对比上下文	上下文感知掩码优化	实例级	2D	上下文信息增强辨别力	场景上下文依赖性高
	econSG（Zhang等，2025）	SAM+OpenSeg	掩码提升	高效掩码处理	块级	2D	有限资源下实现有效语义绑定	为效率牺牲部分精度与粒度

表1续表

绑定范式	文献	核心感知模型	核心创新/机制	特征对齐机制	语义粒度	查询空间	主要特性	核心挑战
	ILGS（Jang等，2025）	SAM	交互式语言引导分割提升	语言指令驱动掩码迭代优化	实例级	2D	支持自然语言交互式细化	实时交互延迟
	Scaffold-GS（Lu等，2023）	SAM	掩码引导三维高斯初始化	基于掩码反向投影	块级	2D	对高质量初始掩码依赖低	优化过程长
	InstanceGaussian（Li等，2025）	SAM	纯实例掩码驱动绑定	掩码到三维高斯的直接映射	实例级	2D	实现端到端实例级语义注入	跨视图实例关联困难
基于单次前馈预测	GSemSplat（Wang等，2024）	CLIP	端到端泛化推理	基线几何特征聚合	块级	2D	无需逐场景优化 泛化性强	未见场景适应性有限
	UniForward（Tian等，2025）	CLIP+ViT-B/32	统一前馈编码器	多视图特征Transformer聚合	点级	2D	统一框架处理多类任务	模型容量与效率平衡
	GaussionTR（Jiang等，2024）	CLIP+FeatUp	Transformer特征聚合	注意力机制融合多视图特征	点级	2D	上下文感知能力强特征分辨率高	网格参数量大效率折中
	SLGaussian（Chen等，2025）	CLIP	轻量前馈网络预测	跨视图记忆库聚合	实例级	2D	推理速度快适合实时交互	跨场景泛化能力弱需微调
	SemanticSplat（Li等，2025）	SAM+LSeg	SAM特征引导前馈预测	利用SAM特征为网络输入先验	点级	2D	实例级与开放词汇能力	网络设计复杂训练数据需求大
	Dr.Splat（Kim等，2025）	CLIP+SAM	空间邻域重排序	邻域信息优化	点级	3D	伪影少 几何感知强	相似度计算复杂度高
	LangScene-X（Liu等，2025）	FC-CLIP+SAM	冻结CLIP场景编码器	冻结特征与可学习适配器对齐	场景级	2D	训练成本低	适配器能力有限
	SceneSplat（Li等，2025）	CLIP+SAM	轻量级重排序网络	跨视图自注意力机制	场景级	2D	实现场景级语义理解	场景级语义评估量化困难

3 应用任务

3.1 开放词汇理解

开放词汇理解是3DGS语义扩展的核心应用。以OpenGaussian（Wu 等，2024）、SceneSplat（Li 等，2025）为代表的方法，通过将VLMs的语义先验嵌入三维高斯，使用户能够借助自然语言描述（如“找到角落里的黑色椅子”）直接查询和定位场景中的对象。GaussianGrasper（Zheng 等，2024）、SparseGrasp（Yu 等，2024）等研究，在3DGS点级理解的基础上，针对不同应用场景进行了深入探索。这不仅支持对常见类别的识别，更能有效处理长尾类别和模糊语义指令，极大地增强了三维场景的语义丰富性与交互智能性，为自动驾驶、机器人导航等领域的开放世界环境理解提供了技术基石。

3.2 高精度实例分割

在开放词汇理解的基础上，研究者进一步实现了高精度的实例与语义分割应用。例如，SeG-Gaussian（Zhang 等，2025）通过分割引导的优化策略显式分解物体，而Seg-Wild（Bao 等，2025）则针对非受限场景对语义不确定性进行建模，二者均显著提升了复杂场景下的分割鲁棒性与边界精度。GS-Grouping（Ye 等，2023）通过引入可学习的实例身份

编码(identity encoding),实现了端到端的实例级语义分组。DCSEG(Wiedmann 等,2025)等框架将开放词汇能力与分割任务深度融合,使其不仅能识别预定义类别,还能根据新颖的文本描述完成像素级或点级的分割。这些技术为机器人抓取(如GraspSplats(Ji 等,2025))、AR/VR 中的物体交互等需要精确定位与操作的应用提供了关键的空间语义感知能力。

3.3 语义驱动编辑

用户可通过自然语言、语义标签、点/框提示等多种模态,对已绑定语义的三维高斯进行选中、删除、替换和属性修改。例如系统可响应用户指令"将桌上的红色杯子变为蓝色",通过编辑对应三维高斯的语义或外观属性,实现"语义即操作"的直观编辑效果。这种基于语义的编辑范式,在数字孪生、虚拟内容创作和智能设计等领域展现出巨大的应用潜力。

3.4 综合集成应用:具身智能与机器人交互

具身智能与机器人交互并非单一的下游任务,而是3DGS语义扩展技术能力的综合体现与高阶集成。它深度融合了前文所述的开放词汇理解(3.1节)、高精度实例分割(3.2节)以及语义驱动编辑能力(3.3节),构建了一个可供机器人感知、理解与操作的可交互数字孪生环境。例如,GaussianGrasper(Zheng 等, 2024)利用语义场的连续性解决遮挡条件下的抓取问题,而SparseGrasp (Yu 等,2024) 则结合开放词汇查询实现对长尾物体的鲁棒识别。这标志着3DGS从一种单纯的渲染技术,演进为连接视觉感知与机器人行动决策的重要桥梁。

4 结论与展望

3DGS技术正经历从高保真三维重建向具备语义理解与交互能力的场景基础模型范式转变。本文构建了以语义信息流为主线的三阶段分类框架,将复杂的研究进展系统划分为语义感知、语义绑定与语义查询三个阶段,并揭示了各阶段的关键技术权衡。语义感知强调开放性与一致性的平衡;语义绑定需兼顾物理效率与重建精度;语义查询则体现了泛化性与结果鲁棒性的权衡。这一系统化分析为理解3DGS语义扩展技术演进提供了清晰、统一的视角。

尽管取得了显著进展,但仍面临从基础理论到实际应用的全链条挑战。基于本文提出的三阶段框架,对当前研究中的关键瓶颈进行分析,未来的研究方向包括:

4.1 突破存储墙:语义特征的轻量化与解耦

针对特征蒸馏带来的显存爆炸问题(如Feature 3DGS(Zhou 等,2024)显存占用相较于传统3DGS(Kerbl 等, 2023)增加约2至4倍,未来应探索超越简单的量化(quantization)的高级压缩技术。例如,根据三维高斯的几何重要性或空间分布,动态分配不同比特率的语义特征;或引入可学习的特征码本,在场景层面进行聚类,实现更紧凑、更具辨别力的语义表示。借鉴DF-3DGS(Dai 等,2025)的思路,研究基于属性解耦的表示方法,将颜色、材质与语义特征进行物理层面的分离。利用神经隐式编码对三维高斯的属性进行二级压缩,通过可学习的特征码本(codebook)在场景层面进行空间聚类,实现更紧凑且具辨别力的语义表示。

4.2 解决动态一致性:时空关联的语义流建模

目前的掩码提升(mask lifting)方法在动态场景中常因拓扑改变而失效。Seg-Wild (Bao 等, 2025)的研究指出,在非受限场景中,光照变化和动态遮挡会显著影响掩码提升方法的稳定性。未来的研究可以引入时序先验与约束:将光流估计、场景流等运动信息作为强先验,引导跨帧语义特征的传播与对齐,构建轻量级的时序语义关联图,为动态绑定提供显式的正则化约束,提升鲁棒性,使语义标签能够稳定地附着在形变的几何表面上,而非漂浮在空间中。探索高效的四维高斯表示:发展轻量化的4DGS框架,在扩展时空维度的同时,严格控制参数增长。核心在于设计高效的时序参数化方式(如位移场、形变场)和联合优化策略,实现动态语义的"一次优化,全程一致"。利用跨模态信息增强一致性:结合事件相机、惯性测量单元等多模态传感器数据,为动态语义建模提供异步、高时间分辨率的触发信号,辅助解决快速运动或遮挡导致的语义断裂问题。

4.3 迈向实用化:端侧轻量化与实时交互

受限于计算资源,当前大多数方法难以直接部署在移动端VR头显或嵌入式机器人端侧。基于单次前馈预测的通用模型是极具潜力的方向,即训练一个通用的"图像-三维高斯"编码器,在推理时直接从多视图图像回归出带语义属性的三维高斯,从而

规避逐场景优化(per-scene optimization)过程。GSemSplat(Wang 等，2024)和 SceneSplat(Li 等，2025)的初步探索表明,这一方向在零样本泛化和实时推理方面具有巨大潜力。

参考文献(References)

Alayrac J, Donahue J, Luc P, Miech A, Barr I, Hasson Y, et al. 2022. Flamingo: a visual language model for few-shot learning//Advances in Neural Information Processing Systems 35. New Orleans, USA: Curran Associates, Inc.: 23716-23736 [DOI: 10.52202/068431-1723]

Bao Y, Ding T, Huo J, Liu Y, Li Y, Li W, Gao Y and Luo J. 2025. 3D Gaussian splatting: survey, technologies, challenges and opportunities. IEEE Transactions on Circuits and Systems for Video Technology, 35(7): 6832-6852 [DOI: 10.1109/TCSVT.2025.3538684]

Bao Y, Tang C, Wang Y and Li H. 2025. Seg-wild: interactive segmentation based on 3D Gaussian splatting for unconstrained image collections//Proceedings of the 33rd ACM International Conference on Multimedia. Dublin, Ireland: ACM: 8567-8576 [DOI: 10.1145/3746027.3755567]

Caron M, Touvron H, Misra I, Jégou H, Mairal J, Bojanowski P and Joulin A. 2021. Emerging properties in self-supervised vision transformers//Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 9630-9640 [DOI: 10.1109/ICCV48922.2021.00951]

Cen J Z, Fang J, Yang C, Xie L, Zhang X, Shen W, et al. 2025. Segment any 3D Gaussians//Proceedings of the 39th AAAI Conference on Artificial Intelligence. Philadelphia, USA: AAAI Press: 1971-1979 [DOI: 10.1609/aaai.v39i2.32193]

Chen G and Wang W. 2024. A survey on 3D Gaussian splatting [EB/OL]. [2026-03-18]. https://arxiv.org/pdf/2401.03890.pdf

Chen K, Dai B, Qin M, Zhang D, Li P, Zou Y and Wang H. 2025. SLGaussian: fast language Gaussian splatting in sparse views//Proceedings of the 33rd ACM International Conference on Multimedia. Dublin, Ireland: ACM: 3047-3056 [DOI: 10.1145/3746027.3754964]

Chen Q, Shu S and Bai X. 2024. Thermal3D-GS: physics-induced 3D Gaussians for thermal infrared novel-view synthesis//Proceedings of the 18th European Conference on Computer Vision. Milan, Italy: Springer: 247-263 [DOI: 10.1007/978-3-031-73383-3_15]

Chen Z, Wu J, Wang W, Su W, Chen G, Xing S, et al. 2024. InternVL: scaling up vision foundation models and aligning for generic visual-linguistic tasks//Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 24185-24198 [DOI: 10.1109/CVPR52733.2024.02283]

Cheng B, Schwing A G and Kirillov A. 2021. Per-pixel classification is not all you need for semantic segmentation//Proceedings of the 35th International Conference on Neural Information Processing Systems. Online: Curran Associates Inc.: 17864-17875

Cheng J, Zaech J N, Van Gool L and Paudel D P. 2025. Occam's LGS: an efficient approach for language Gaussian splatting//Proceedings of the 36th British Machine Vision Conference. Sheffield, UK: BMVA: 1-12

Cherti M, Beaumont R, Wightman R, Wortsman M, Ilharco G, Gordon C, et al. 2023. Reproducible scaling laws for contrastive language-image learning//Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 2818-2829 [DOI: 10.1109/CVPR52729.2023.00276]

Dai Z, Liu T and Zhang Y. 2025. Efficient decoupled feature 3D Gaussian splatting via hierarchical compression//Proceedings of the 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 11156-11166 [DOI: 10.1109/CVPR52734.2025.01042]

Guo J, Ma X, Fan Y, Liu H and Li Q. 2024. Semantic Gaussians: open-vocabulary scene understanding with 3D Gaussian splatting// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE: 16080-16090.

He K, Gkioxari G, Dollár P and Girshick R. 2017. Mask R-CNN//Proceedings of the 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 2980-2988 [DOI: 10.1109/ICCV.2017.322]

He S, Ji P, Yang Y, Wang C, Ji J, Wang Y, et al. 2025. A survey on 3D Gaussian splatting applications: segmentation, editing and generation[EB/OL]. [2026-03-18]. https://arxiv.org/pdf/2508.09977.pdf

Hu J, Chen Z, Li Z, Xu Y and Zhang J. 2024. SparseLGS: sparse view language embedded Gaussian splatting[EB/OL]. [2026.03.18]. https://arxiv.org/pdf/2412.02245.pdf

Jang S M and Kim W J. 2025. Identity-aware language Gaussian splatting for open-vocabulary 3D semantic segmentation//Proceedings of the 2025 IEEE/CVF International Conference on Computer Vision. Seoul, South Korea: IEEE: 12345-12355 [DOI: 10.1109/ICCV51070.2025]

Ji M, Qiu R, Zou X and Wang X. 2025. GraspSplats: efficient manipulation with 3D feature splatting//Proceedings of the 8th Conference on Robot Learning. Munich, Germany: PMLR: 1443-1460

Jia C, Yang Y, Xia Y, Chen Y, Parekh Z, Pham H, et al. 2021. Scaling up visual and vision-language representation learning with noisy text supervision[EB/OL]. [2026-03-18] https://arxiv.org/pdf/2102.05918.pdf

Jia Di, Zhao Chen, Zhang Huaxiu and Song Huilun. 2026. Lightweight RGB-D semantic segmentation network incorporating frequency-domain guidance. Journal of Image and Graphics, 31

(2):0479-0498 (贾迪, 赵辰, 张华修, 宋慧伦. 2026. 融合频域引导的RGB-D轻量级语义分割网络. 中国图象图形学报, 31(2):0479-0498)[DOI: 10.11834/jig.250212]

Jiang H, Liu L, Cheng T, Wang X, Lin T, Su Z, et al. 2025. GaussTR: foundation model-aligned Gaussian transformer for self-supervised 3D spatial understanding//Proceedings of the 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 11960-11970 [DOI: 10.1109/CVPR52734.2025.01117]

Jiang M, Jia S, Gu J, Lu X, Zhu G, Dong A, et al. 2025. VoteSplat: Hough voting Gaussian splatting for 3D scene understanding//Proceedings of the 2025 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 6456-6465

Jose C, Moutakanni T, Kang D, Baldassarre F, Darcet T, Xu H, et al. 2025. DINOv2 meets text: a unified framework for image- and pixel-level vision-language alignment//Proceedings of the 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 24905-24916 [DOI: 10.1109/CVPR52734.2025.02319]

Kerbl B, Kopanas G, Leimkuehler T and Drettakis G. 2023. 3D Gaussian splatting for real-time radiance field rendering[J/OL]. ACM Transactions on Graphics, 42(4): 139 [DOI: 10.1145/3592433]

Kerr J, Kim C, Goldberg K, Kanazawa A and Tancik M. 2023. LERF: language embedded radiance fields//Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 19672-19682 [DOI: 10.1109/ICCV51070.2023.01807]

Kim G, Kwon T and Ye J. 2022. DiffusionCLIP: text-guided diffusion models for robust image manipulation//Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 2416-2425 [DOI: 10.1109/CVPR52688.2022.00246]

Kim J, Kim G, Kim Y, Wang Y, Choe J and Oh T. 2025. Dr. Splat: directly referring 3D Gaussian splatting via direct language embedding registration//Proceedings of the 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 14137-14146 [DOI: 10.1109/CVPR52734.2025.01319]

Kirillov A, Mintun E, Ravi N, Mao H, Rolland C, Gustafson L, et al. 2023. Segment anything[C]//Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV). Paris: IEEE: 3992-4003. [DOI: 10.1109/ICCV51070.2023.00371]

Kobayashi S, Matsumoto E and Sitzmann V. 2022. Decomposing NeRF for editing via feature field distillation//Proceedings of the 35th International Conference on Neural Information Processing Systems. New Orleans, USA: NeurIPS: 23311-23330 [DOI: 10.52202/068431-1694]

Labe I, Issachar N, Lang I and Benaim S. 2025. DGD: dynamic 3D Gaussians distillation//Proceedings of the 18th European Conference on Computer Vision. Milan, Italy: Springer: 345-361 [DOI: 10.1007/978-3-031-73113-6_21]

Levy Y, Shavin D, Lang I and Benaim S. 2025. Structurally disentangled feature fields distillation for 3D understanding and editing [EB/OL]. [2026-03-18]. https:// arxiv.org/pdf/2502.14789.pdf

Li B Y, Weinberger K Q, Belongie S J, Koltun V and Ranftl R. 2022. Language-driven semantic segmentation [EB/OL]. [2026-03-18]. https://arxiv.org/pdf/2201.03546.pdf

Li D, Feng J, Chen J, et al. 2025. EgoSplat: open-vocabulary egocentric scene understanding with language embedded 3D Gaussian splatting [EB/OL]. [2026-03-18]. https://arxiv.org/pdf/2503.11345.pdf

Li H, Wu Y, Meng J, Gao Q, Zhang Z, Wang R, et al. 2025. InstanceGaussian: appearance-semantic joint Gaussian representation for 3D instance-level perception//Proceedings of the 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 14078-14088 [DOI: 10.1109/CVPR52734.2025.01314]

Li J, Li D, Savarese S and Hoi S. 2023. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models//Proceedings of the 40th International Conference on Machine Learning. Honolulu, USA: PMLR: 19730-19742 [DOI: 10.5555/3618408.3619222]

Li J, Li D, Xiong C and Hoi S. 2022. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation//Proceedings of the 39th International Conference on Machine Learning. Baltimore, USA: PMLR: 12888-12900

Li Q, Sun J, An L, Su Z, Zhang H and Liu Y. 2025. SemanticSplat: feed-forward 3D scene understanding with language-aware Gaussian fields [EB/OL]. [2026-03-18]. https://arxiv.org/pdf/2506.09565.pdf

Li W, Zhou R, Zhou J, et al. 2025. 4D LangSplat: 4D language Gaussian splatting via multimodal large language models//Proceedings of the 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 22001-22011 [DOI: 10.1109/CVPR52734.2025.02049]

Li Y, Ma Q, Yang R, Li H, Ma M J, Ren B, et al. 2025. Scenesplat: Gaussian splatting-based scene understanding with vision-language pretraining//Proceedings of the IEEE/CVF International Conference on Computer Vision. Kyoto, Japan: IEEE: 4961-4972

Liao G B, Li J K, Bao Z Y, Ye X Q, Li Q and Liu K L. 2025. CLIP-GS: CLIP-informed Gaussian splatting for real-time and view-consistent 3D semantic understanding. ACM Transactions on Multimedia Computing, Communications, and Applications, 21(8): 240 [DOI: 10.1145/3746284]

Liu F, Li H, Chi J, Wang H, Yang H, Yang M, et al. 2025. Langscene-x: reconstruct generalizable 3D language-embedded scenes with trimap video diffusion//Proceedings of the IEEE/CVF International Conference on Computer Vision. Kyoto, Japan: IEEE: 29010-29020

Liu H, Li C, Wu Q and Lee Y J. 2023. Visual instruction tuning// Advances in Neural Information Processing Systems 36. New Orleans, USA: NeurIPS: 34892-34916

Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, et al. 2021. Swin transformer: hierarchical vision transformer using shifted windows//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 9992-10002 [DOI: 10.1109/ICCV48922.2021.00986]

Lu T, Li Z, Liu Y, Sima C, Wang S, Wang J, et al. 2023. Scaffold-GS: structured 3D Gaussians for view-adaptive rendering//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 20654-20664 [DOI: 10.1109/CVPR52733.2023.01986]

Lyu W, Li X, Kundu A, Tsai Y H and Yang M H. 2024. Gaga: group any gaussians via 3D-aware memory bank[EB/OL]. [2024-04-11]. https://arxiv.org/pdf/2404.07977.pdf

Mildenhall B, Srinivasan P P, Tancik M, Barron J T, Ramamoorthi R and Ng R. 2020. NeRF: representing scenes as neural radiance fields for view synthesis//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 405-421 [DOI: 10.1007/978-3-030-58452-8_24]

Oquab M, Darcet T, Moutakanni T, Vo H V, Szafraniec M, Khalidov V, et al. 2023. DINOv2: Learning Robust Visual Features without Supervision. [EB/OL]. [2026-03-18]. https://arxiv.org/pdf/2304.07193.pdf

Peng Y, Wang H, Liu Y, Wen C, Dong Z and Yang B. 2024. Gags: granularity-aware feature distillation for language Gaussian splatting [EB/OL]. [2026-03-18]. https://arxiv.org/pdf/2412.13654.pdf

Qi C R, Yi L, Su H and Guibas L J. 2017. PointNet++: deep hierarchical feature learning on point sets in a metric space//Advances in Neural Information Processing Systems 30. Long Beach, USA: NIPS: 5099-5108 [DOI: 10.48550/arXiv.1706.02413]

Qin M H, Li W H, Zhou J, Wang H Q and Pfister H. 2024. LangSplat: 3D language Gaussian splatting//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 20051-20060 [DOI: 10.1109/CVPR52733.2024.01895]

Qiu R Z, Yang G, Zeng W J and Wang X. 2024. Feature splatting: language-driven physics-based scene synthesis and editing [EB/OL]. [2026-03-18]. https://arxiv.org/pdf/2404.01223.pdf

Qu Y S, Song Z, Li Z, Wang P and Lin D. 2024. GOI: find 3D Gaussians of interest with an optimizable open-vocabulary semantic-space hyperplane//Proceedings of the 32nd ACM International Conference on Multimedia. Melbourne, Australia: ACM: 5328-5337 [DOI: 10.1145/3664647.3680852]

Radford A, Kim J W, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. 2021. Learning transferable visual models from natural language supervision//Proceedings of the 38th International Conference on Machine Learning. Virtual Event: PMLR: 8748-8763.

Rao Y M, Chen X, Lu J, Zhou J and Yan S. 2022. Denseclip: language-guided dense prediction with context-aware prompting//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 18061-18070 [DOI: 10.1109/CVPR52688.2022.01754]

Ravi N, Gabeur V, Hu Y T, Hu R, Ryali C K, Ma T, et al. 2024. SAM 2: segment anything in images and videos [EB/OL]. [2026-03-18]. https://arxiv.org/pdf/2408.00714.pdf

Safa A, Mohamed A, Issam B and Mohamed-Yassine H. 2023. Segformer: semantic segmentation based transformers for corrosion detection//Proceedings of the 2023 International Conference on Networking and Advanced Systems. Annaba, Algeria: IEEE: 1-6 [DOI: 10.1109/ICNAS59892.2023.10330461]

Shen Q, Yang X, Wang X. 2024. Flashsplat: 2d to 3d Gaussian splatting segmentation solved optimally//Proceedings of the 18th European Conference on Computer Vision. Milan, Italy: Springer: 437-454 [DOI: 10.1007/978-3-031-72670-5_26]

Shi J C, Wu Y, Liu Y, Wang X, Qin M H, Li W H, et al. 2023. Language embedded 3D Gaussians for open-vocabulary scene understanding//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE: 5333-5343 [DOI: 10.1109/CVPR52733.2023. 00516]

Siddiqui Y, Porzi L, Bulo S R, Müller N, Nießner M, Dai A, et al. 2023. Panoptic lifting for 3d scene understanding with neural fields//Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 9043-9052 [DOI: 10.1109/CVPR52729.2023.00873]

Tian L, Li X, Ma L, Huang H, Zheng Z, Yin H, et al. 2025. CCL-LGS: contrastive codebook learning for 3D language Gaussian splatting [EB/OL]. [2026-03-18]. https://arxiv.org/pdf/2505.20469.pdf

Tian Q J, Li Z, Liu Y, Wang Y and Zhang Y. 2025. UniForward: unified 3D scene and semantic field reconstruction via feed-forward Gaussian splatting from only sparse-view images [EB/OL]. [2026-03-18]. https://arxiv.org/pdf/2506. 09378.pdf

Wang X R, Li Y, Zhang Z, Liu Y and Chen D. 2024. GSemSplat: generalizable semantic 3D Gaussian splatting from uncalibrated image pairs [EB/OL]. [2026-03-18]. https://arxiv.org/pdf/2412.16932.pdf

Wang Xinjing, Gao Ying, Zou Yaqi, Zhu Zhengyu, Xu Chunxue, Zhao Qi. Semantic Alignment and Locality-Driven Open-Vocabulary Semantic Segmentation [J/OL]. Journal of Image and Graphics, 2026, 1-11 (王馨静，高颖，邹亚琦，朱政宇，徐春雪，赵琦. 语义对齐与局部感知驱动的开放词汇语义分割[J/OL]. 中国图象图形学报，2026,1-11[DOI: 10.11834/jig.250540]

Wiedmann L, Schult J, Engelmann A and Leibe B. 2025. DCSEG: decoupled 3d open-set segmentation using Gaussian splatting//Proceedings of the 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Nashville, USA: IEEE: 5217-5226 [DOI: 10.1109/CVPRW67362.2025.00517]

Wu T, Yuan Y J, Zhang L X, Yang J, Cao Y P, Yan L Q, et al. 2024. Recent advances in 3d Gaussian splatting//Computational Visual Media. 10(4): 613-642 [DOI: 10.1007/s41095-024-0436-y]

Wu Y M, Li H J, Meng J R, Gao Q K, Zhang Z Y, Wang R G, et al. 2024. OpenGaussian: towards point-level 3D Gaussian-based open vocabulary understanding. Neural Information Processing Systems Foundation, Inc: 19114-19138[DOI:10.52202/079017-0604]

Xie E, Wang W, Yu Z, Anandkumar A, Alvarez J M and Luo P. 2021. Segformer: simple and efficient design for semantic segmentation with transformers//Advances in Neural Information Processing Systems. 34: 12071-12081

Ye M, Wu Y, Liu Y, Chen X, Zhang Y and Wang Y. 2024. Gaussian grouping: segment and edit anything in 3d scenes//Proceedings of the 2024 European Conference on Computer Vision. Milan, Italy: Springer: 234-251 [DOI: 10.1007/978-3-031-93735-7_12]

Yin H, Zhan H Y, Xu Y, Chen J and Yeh R A. 2025. Semantic consistent language Gaussian splatting for point-level open-vocabulary querying [EB/OL]. [2026-03-18]. https://arxiv.org/pdf/2503.21767.pdf

Yu J Q, Ren X L, Gu Y C, Lin H T, Wang T Y, Zhu Y, et al. 2024. SparseGrasp: robotic grasping via 3D semantic Gaussian splatting from sparse multi-view RGB images. [EB/OL]. [2026-03-18]. https://arxiv.org/pdf/2412.02140.pdf

Zhai X H, Mustafa B, Kolesnikov A and Beyer L. 2023. Sigmoid loss for language image pre-training//Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 11941-11952 [DOI: 10.1109/ICCV51070.2023.01102]

Zhang C and Lee G H. 2025. econSG: efficient and multi-view consistent open-vocabulary 3D semantic Gaussians [EB/OL]. [2026-03-18]. https://arxiv.org/pdf/2504.06003.pdf

Zhang L X, Jiang C, Lai Y K and Gao L. 2025. SeG-Gaussian: segmentation-guided 3d Gaussian optimization for novel view synthesis. IEEE Transactions on Visualization and Computer Graphics [DOI: 10.1109/TVCG.2025.3615421]

Zheng Y H, Li H J, Wu Y M, Meng J R, Gao Q K, Zhang Z Y, et al. 2024. GaussianGrasper: 3D language Gaussian splatting for open-vocabulary robotic grasping. IEEE Robotics and Automation Letters, 9(8): 7827-7834 [DOI: 10.1109/LRA.2024.3428931]

Zhi S, Laidlow T, Leutenegger S and Davison A J. 2021. In-place scene labelling and understanding with implicit scene representation//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 15818-15827 [DOI: 10.1109/ICCV48922.2021.01554]

Zhou S J, Wu Y, Liu Y, Wang X, Li W H and Pfister H. 2023. Feature 3DGS: supercharging 3D Gaussian splatting to enable distilled feature fields//Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 21676-21685 [DOI: 10.1109/CVPR52733.2023.02091]

Zhou S J, Wu Y, Liu Y, Wang X, Li W H and Pfister H. 2025. Feature4X: bridging any monocular video to 4D agentic AI with versatile Gaussian feature fields//Proceedings of the 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 14179-14190 [DOI: 10.1109/CVPR52734.2025.01320]

Zhou Z Q, Wei Y, Shi Y, Liu Y and Huang T. 2022. ZegCLIP: towards adapting CLIP for zero-shot semantic segmentation//Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 11175-11185 [DOI: 10.1109/CVPR52688.2022.01092]

Zhu R, Qiu S, Liu Z, Hui K H, Wu Q, Heng P A and Fu C W. 2025. Rethinking end-to-end 2D to 3D scene segmentation in Gaussian splatting//Proceedings of the 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 12345-12355 [DOI: 10.1109/CVPR52734.2025.00346]

Zuo X X, Samangouei P Y, Li Y, Zhou Y W, Di Y and Li M Y. 2025. FMGS: foundation model embedded 3d Gaussian splatting for holistic 3d scene understanding. International Journal of Computer Vision, 133(2): 611-627 [DOI: 10.1007/s11263-024-02183-8]

Zwicker M, Pfister H, Van Baar J and Gross M. 2001. EWA volume splatting//Proceedings Visualization, 2001. San Diego, USA: IEEE: 29-36 [DOI: 10.1109/VISUAL.2001.964490]

作者简介

张思佳,女,硕士研究生,主要研究方向为三维视觉和场景理解。E-mail:1697346032@qq.com

张荣华,通信作者,男,高级工程师,主要研究方向为计算机图形学、3D AIGC和数字孪生等。E-mail:zronghua88@aliyun.com

李丽芬,女,副教授,主要研究方向为智能电网与电力信息化、计算机视觉等,发表学术论文20余篇。E-mail:52150760@ncepu.edu.cn