InstaScene: Towards Complete 3D Instance Decomposition and Reconstruction from Cluttered Scenes

Zesong Yang[1,2*] Bangbang Yang[2] Wenqi Dong[1,2] Chenxuan Cao[1]
Liyuan Cui[1] Yuewen Ma[2] Zhaopeng Cui[1] Hujun Bao[1†]

[1] State Key Lab of CAD & CG, Zhejiang University [2] ByteDance

Project Page: https://zju3dv.github.io/instascene/

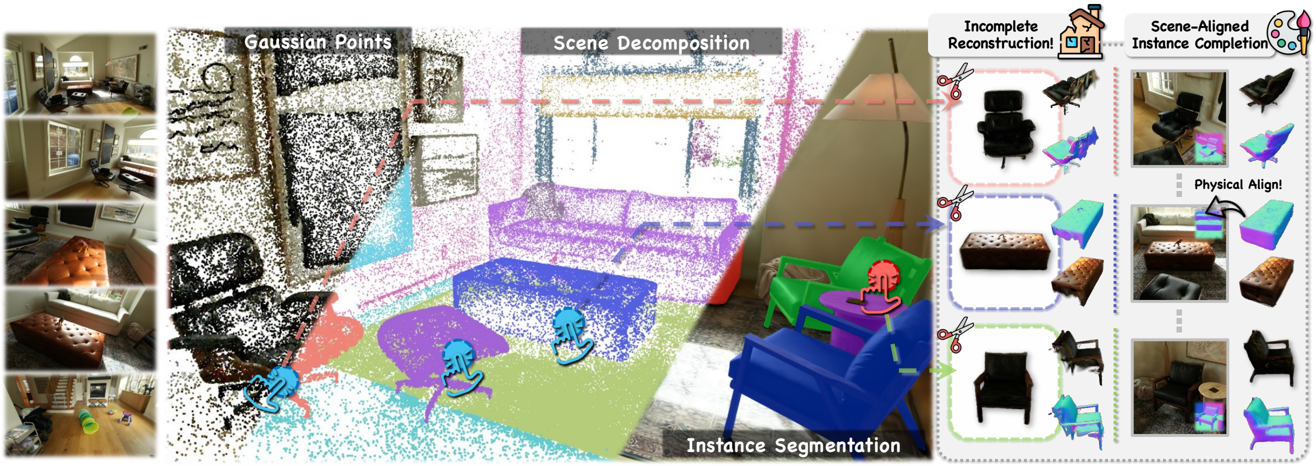


Figure 1. **InstaScene** allows users to pick up and decompose arbitrary instances from cluttered environments, while automatically reconstructing them into complete objects with intact geometry and appearance that align with the physical world.

Abstract

Humans can naturally identify and mentally complete occluded objects in cluttered environments. However, imparting similar cognitive ability to robotics remains challenging even with advanced reconstruction techniques, which models scenes as undifferentiated wholes and fails to recognize complete object from partial observations. In this paper, we propose ***InstaScene****, a new paradigm towards holistic 3D perception of complex scenes with a primary goal: decomposing arbitrary instances while ensuring complete reconstruction. To achieve precise decomposition, we develop a novel spatial contrastive learning by tracing rasterization of each instance across views, significantly enhancing semantic supervision in cluttered scenes. To overcome incompleteness from limited observations, we introduce in-situ generation that harnesses valuable observations and geometric cues, effectively guiding 3D generative models to reconstruct complete instances that seamlessly align with the real world. Experiments on scene decomposition and object completion across complex real-world and synthetic scenes demonstrate that our method achieves superior decomposition accuracy while producing geometrically faithful and visually intact objects.*

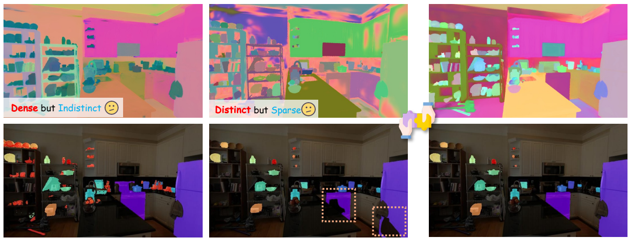
1. Introduction

As humans, we possess an innate ability to understand cluttered 3D scenes and interact with diverse objects without deliberate observation. Imagine walking into a crowded kitchen: we can immediately recognize each piece of furniture and every dish, and without a second thought you might reach out to pick up a specific utensil from a busy countertop. Over the years, numerous efforts have been made in computer vision and robotics to achieve similar skills [18], developing series of capabilities like efficient reconstruction and rendering [14], 3D scene understanding [33], and 3D object generation [9, 34, 61]. However, the gap for downstream scene editing and simulation applications [13, 43, 55, 63] still remains.

Several key limitations in existing approaches hinder this goal. First, most generic 3D reconstruction methods [14, 28, 39, 47] usually treat the scene as a whole model, which hinders delicate instance-level tasks such as manipulation and rearrangement; Second, open-set scene understanding methods [33, 46] enables object-level query and segmentation but fails to produce complete instances from cluttered environments; Third, category-specific generative approaches can predict complete 3D shapes from partial observations, but they often struggle to generalize to diverse objects in cluttered scenes, or ensure physical and visual alignment with the real world [21], *i.e.*, matching the object's actual size, shape, and texture, leading to physical and

[*] Work done during an internship at PICO, ByteDance.

[†] Corresponding author.



(a) Noisy 2D Segmentation (b) Coarse 3D Segmentation (c) Ours Feat. & Segmentation

Figure 2. **Motivation of spatial contrastive learning with mutual guidance.** In complex scenes, naively supervising feature field with noisy 2D segmentation masks results in indistinct features (see (a)), and only using 3D masks from spatial trackers would result in sparse Gaussian points (see (b)). Observed that the former provides dense features while the latter offers a robust reference, we leverage the interplay to mutually guide each other and achieve a dense and distinct feature field (see (c)).

visual inconsistencies.

To enable robotic systems with human-like perception for understanding and interacting with surrounding environment, a new paradigm is required to seamlessly integrate 3D understanding, reconstruction, and generation into a unified system with two essential capabilities: **1) Precise instance decomposition**: the system should precisely segment and isolate arbitrary object instances in a complex scene, irrespective of their categories. **2) Complete instance-aware scene reconstruction**: for every object that is extracted, the system should reconstruct the object's complete geometry and appearance while adhering to the physical world, even if parts of it were never directly observed due to occlusion or limited viewpoints.

In this work, we propose **InstaScene**, a framework that realizes the above vision by integrating instance segmentation and complete reconstruction in a unified pipeline. At a high level, our method takes a captured cluttered scene (*i.e*., in the form of Gaussian Splatting [14]), and decomposes it into complete instances that possess faithful geometry and appearance. By addressing segmentation and reconstruction together, **InstaScene** ensures that the object in the scene can be individually obtained as a complete 3D model **aligned** with the scene's context. To achieve this goal, we first develop a novel spatial contrastive learning scheme guided by traced Gaussian clustering for fine-grained instance decomposition. Specifically, we construct spatial trackers that identify and cluster the Gaussian points primarily contributing to each mask's rasterization, enabling consistent instance identification across multiple views. These trackers provide reliable supervision for the instance-level feature field through spatial contrastive learning, yielding highly distinguishable features as shown in Fig. 2 (c). This enables precise scene decomposition even in challenging environments (*e.g*., closely spaced bottles in Fig. 6).

To recover complete objects that align with real-world scenes, a primary challenge is that clean object views required by typical 3D generative models [9, 34, 61] are rarely available in complex scenes, where objects are often partially occluded or captured from suboptimal views. To tackle this challenge, we propose in-situ generation, which harnesses a 3D generative model to leverage valuable information from known observations and partially reconstructed geometry as omni-conditions. Specifically, for each instance, we employ occlusion-aware viewpoint selection to render optimal views that serve as alternated conditions for the diffusion process. We further incorporate geometry hints to complement known latent features via feature warping, enhancing multi-view consistency between observations and generations. The generated views, together with source observations, are then used to finetune the object's Gaussian model, achieving complete reconstruction that can be directly placed back into the scene (referred to as ”in-situ” or ”in place”). Through these comprehensive designs, our method effectively emulates the human cognitive ability to identify and mentally complete partially observed objects in complex environments, thereby facilitating downstream tasks like robot-object interaction.

Our main contributions can be summarized as follows:

- We introduce **InstaScene**, a novel framework, which decomposes arbitrary objects from complex scenes while ensuring complete geometry and appearance reconstruction of each instance with limited observation.
- We propose a novel spatial contrastive learning scheme based on a traced Gaussian clustering technique, achieving accurate scene decomposition.
- We design a new in-situ generation pipeline, which aggregates all available observations and geometric hints as omni-conditions to control the 3D generative prior, yielding complete instance modeling that seamlessly aligns with the physical world.
- Extensive validations on various complex datasets demonstrate the superior performance of our method in fine-grained scene decomposition and faithful complete object reconstruction, with promising applications for downstream tasks like scene editing and manipulation.

2. Related Work

Instance-aware Scene Reconstruction. Generic scene reconstruction approaches [14, 28, 40] usually model the entire scene as a single representation. For further scene understanding tasks, recent works tend to conduct reconstruction at instance-level granularity [19, 21, 49, 50, 56]. However, existing methods either require laborious manual mask annotation [19, 49, 50, 56] for instance decomposition and usually encode instances through multiple MLPs, which struggle with the scene containing numerous objects due to the limited scalability. Another line of work learns a generative shape prior from predefined CAD models or large-scale synthetic datasets [1, 4, 8, 21]. Nevertheless, due to the limitation of 3D data collection, these methods often face difficulties in generalizing to complex real-world scenes with numerous categories and diverse shapes, which leads to domain gaps and shape misalignment between reconstructed results and real-world objects. Besides, most of these works only focus on geometric modeling while ignoring realistic appearance rendering, which restricts the applications in vision-based simulation and scene editing.

3D Segmentation with Scene Reconstruction. Recent

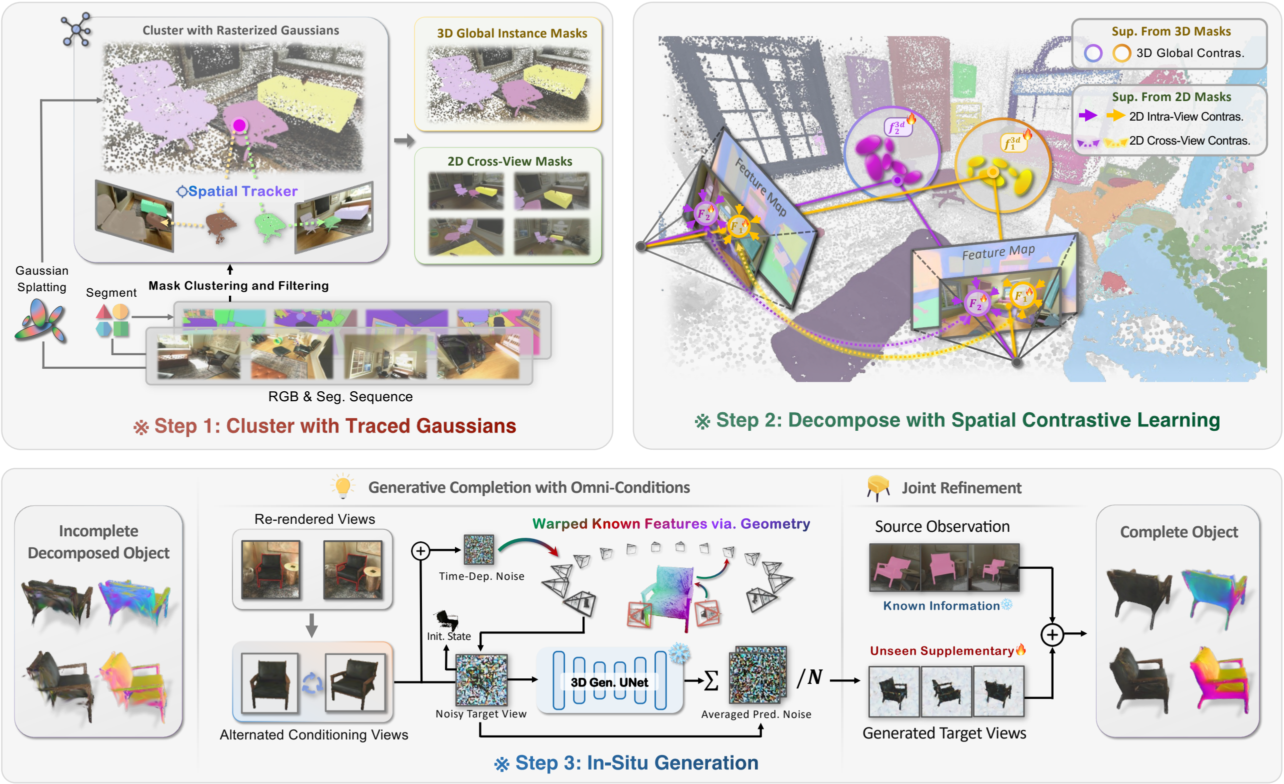


Figure 3. **System Overview.** Given a reconstructed Gaussian Splatting scene, our method first clusters and filters 2D segmentation masks by tracing the rasterization of Gaussian Splatting, which yields 2D and 3D instance masks. Then, we use spatial contrastive learning with mask supervision to train a feature field that achieves high-quality scene decomposition. Finally, for each decomposed incomplete object, we conduct an in-situ generation pipeline that takes all known observations and geometric cues as omni-conditions to the 3D generative model to obtain supplemented views, which will be jointly fine-tuned with source training views to obtain a complete object.

advances in large-scale vision, language, and segmentation models [3, 17, 32, 38] have enabled encoding rich visual features into volumetric fields or 3D Gaussians, facilitating versatile 3D instance segmentation over existing scene reconstruction. These works either reduce the dimension of pre-trained features with quantization and distillation [36, 37, 41, 62], or conduct video tracking for cross-view masks association [58]. However, VLM-based semantic features struggle with complex, cluttered scenes containing duplicate objects (Fig. 6), while video tracking becomes unreliable under heavy occlusion, which inevitably messes up the feature learning. Some approaches [5, 16, 20, 51, 60] adopt contrastive learning scheme, which enables object extraction from well-observed data but still struggles under cluttered scenes due to inconsistent 2D priors (see Fig. 2 in the experiments). Besides, none of these works consider the completeness of the extracted instance, which limits the application in downstream scene editing or simulation tasks.

3D Inpainting and Amodal Reconstruction. To obtain complete instances under insufficient observation or severe occlusion, one line of works is to apply 3D inpainting. However, most related works focus on scene-level inpainting using 2D inpainting tools [25, 27, 29, 44], which excels at removing objects from a clean background but cannot recover occluded parts of complex objects. For object-level inpainting tasks, [11, 48] use 2D diffusion-based inpainting tools but require user-provided mask sketches and prompts, and the result is not stable. Apart from that, one promising approach is to utilize recent image-to-3D models [6, 10, 23, 24, 42] for generative reconstruction, but generic generation pipelines cannot ensure alignment with real-world scenes or properly handle incomplete/occluded observation (see Sec. 4.2), limiting the applicability for precise and faithful reconstruction. A similar concurrent work, DP-Recon [31], employs generative priors to improve sparse and occluded regions—geometry completion then texture refinement, while our method jointly completes both in a single step with realistic rendering.

3. Methods

We present a novel framework for instance-aware reconstruction in complex scenes, jointly addressing segmentation and instance completion in a unified pipeline. Unlike prior approaches that treat segmentation and generation as disjoint tasks, our segmentation stage provides essential spatial priors (e.g., geometry, object-centric views, masks) that directly guide the subsequent generative module. Our key components include a spatial contrastive learning for fine-grained scene decomposition (Sec. 3.2), and a new in-situ generation for complete instance modeling (Sec. 3.3).

3.1. Preliminary

Given posed RGB image sequences as input, we first reconstruct the scene with point-based rendering method 2D Gaussian Splatting [12]. Building upon 3DGS [14], 2DGS collapses the 3D ellipsoid volumes into a set of 2D oriented

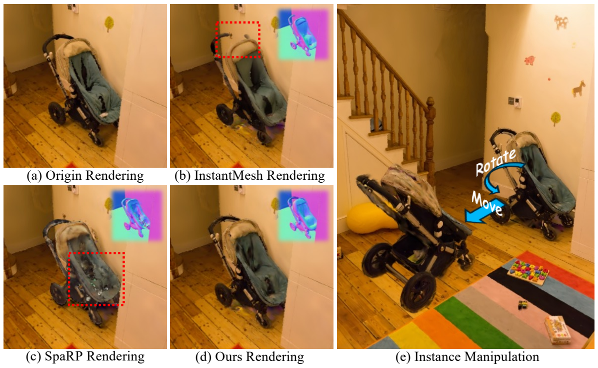

(a) Origin Rendering
(b) InstantMesh Rendering
(c) SpaRP Rendering
(d) Ours Rendering
(e) Instance Manipulation

Figure 4. **In-Situ Generation vs. Generic Image-to-3D.** We perform different instance-level complete reconstruction approaches on the baby carriage and put it back on the scene. Generic image-to-3D methods [52, 53] struggle to maintain consistent reconstruction with the original scene (a) and suffer from misalignment (such as broken handles or seats in (b) and (c)). Our in-situ generation ensures the appearance and geometric consistency to the original scene while faithfully completing the unseen regions (see (d)). We also show the application of scene manipulation on the decomposed complete instance in (e).

planar Gaussian disks, enhancing view-consistent rendering while improving the geometric quality. A 2D Gaussian is parameterized as a local tangent plane, which is defined as:

$$P(u,v) = \mathbf{p}_k + \mathbf{s}_u \mathbf{t}_u u + \mathbf{s}_v \mathbf{t}_v v, \tag{1}$$

where the center $\mathbf{p}_k$, scaling $(\mathbf{s}_u, \mathbf{s}_v)$, the rotation $(\mathbf{t}_u, \mathbf{t}_v)$, opacity α and view-dependent appearance $\mathbf{c}$ with spherical harmonics are learnable parameters. For a point $\mathbf{u} = (u, v)$ in uv space, its 2D Gaussian value can be evaluated as:

$$\mathcal{G}(\mathbf{u}) = \exp\left(-\frac{u^2 + v^2}{2}\right). \tag{2}$$

Through rasterization, Gaussians passed through by the rays $\mathbf{x}$ emitted from the uv space are composed into pixel appearance with depth-sorted alpha blending, as:

$$\mathbf{c}(\mathbf{x}) = \sum_{i=1} \mathbf{c}_i \alpha_i \mathcal{G}_i(\mathbf{u}(\mathbf{x})) \prod_{j=1}^{i-1} (1 - \alpha_j \mathcal{G}_j(\mathbf{u}(\mathbf{x}))). \tag{3}$$

We augment each Gaussian with a D-dimensional randomly initialized embedding $\mathbf{f}_i^{3d} \in \mathbb{R}^D$ as the instance-aware feature. With the rasterizer akin to Eq. 3, we obtain pixel-level features $\mathbf{f}$, as:

$$\mathbf{f}(\mathbf{x}) = \sum_{i=1} \mathbf{f}_i^{3d} \alpha_i \mathcal{G}_i(\mathbf{u}(\mathbf{x})) \prod_{j=1}^{i-1} (1 - \alpha_j \mathcal{G}_j(\mathbf{u}(\mathbf{x}))). \tag{4}$$

In our setup, we set $D = 16$ and freeze other attributes.

3.2. Scene Decomposition with Spatial Contrastive Learning

Mask Clustering with Spatial Gaussian Tracker. We first employ the off-the-shelf mask predictor EntitySeg [35] to generate class-agnostic instance-level 2D segmentation masks. To obtain 3D segmentation of the scene, existing methods [5, 16, 20, 51, 60] generally lift 2D segmentation priors to a unified 3D space. However, 2D segmentation priors are usually noisy (*e.g.*, cross-view inconsistency, under-segmentation), which introduces ambiguity and suboptimal supervision in feature field learning (see Fig. 2). To extract robust supervision from 2D masks, inspired by previous works that exploit the re-projected spatial consistency of 2D segmentations to achieve robust 3D segmentation [26, 30, 54, 57, 59], we apply a trustful mask clustering strategy with traced Gaussian points, which back-projects the segmentation masks into 3D space and utilizes the spatial relationships among the corresponding Gaussian points to achieve cross-view mask matching, which yields a robust global semantic prior for scene decomposition.

Due to the inherently noisy distribution of reconstructed 3DGS point clouds, naive depth projection discards potentially valid points associated with 2D segmentations. Given the frame I_i along with its semantic map M_i, we render $\bar{I}_i$ with its corresponding pose and define the Gaussians that contribute to the rasterization of each pixel within the j-th mask $m_{i,j}$, with transmittance exceeding 0.5, as individual spatial tracker $P_{i,j}$. To determine whether two masks belong to the same instance, we utilize the view consensus rate as proposed in [54]. Specifically, a spatial tracker $P_{i,j}$ is considered visible at frame $I_{i'}$ if its 30% points contribute to the rasterization of $I_{i'}$, and $P_{i,j}$ is contained within frame $I_{i''}$ if 80% of its points appear within a tracker $P_{i'',j''}$. Given any two mask trackers $P_{i,j}$ and $P_{k,l}$, the view consensus rate $\mathcal{C}$ is defined as:

$$\mathcal{C}(P_{i,j}, P_{k,l}) = \frac{N_{contain}(P_{i,j}, P_{k,l})}{N_{vis}(P_{i,j}, P_{k,l})}, \tag{5}$$

where N_{vis} and $N_{contain}$ represent the number of frames where both trackers are visible and contained respectively. If $\mathcal{C}$ exceeds 0.9, the corresponding masks $m_{i,j}, m_{k,l}$, can be considered to belong to the same instance.

Additionally, we categorize $m_{i,j}$ as under-segmentation if its associated spatial tracker $P_{i,j}$ simultaneously intersects with multiple trackers $\{P_{k,j'}\}$ from the same frame I_k and is consistently present in its visible frames. Then we discard $m_{i,j}$. Thus, we achieve cross-view mask clustering and filter noisy segmentation by tracing rasterization of Gaussians. We denote the clustered cross-view segmentations for instance $\mathcal{I}_n$ as $\mathcal{M}_n^{2d} = \{m_{i,j}\}$. In addition to utilizing these multi-view consistent semantic priors, we merge Gaussian tracker points belonging to the same instance and employ DBSCAN[7] to filter out floaters. The fused Gaussians $\mathcal{P}_n$ serve as a robust 3D global instance mask $\mathcal{M}_n^{3d}$ for instance $\mathcal{I}_n$.

Spatial Contrastive Learning. Although DBSCAN is applied to remove floaters, it inadvertently discards semantically meaningful Gaussian points misclassified as noise (see Fig. 2.b). To address this, we propose a spatial contrastive learning framework that jointly leverages intra-view and cross-view consistent 2D segmentation masks, along with robust 3D global instance masks, to effectively guide the learning of a distinctive and semantically coherent feature field. While the 3D masks are sufficiently distinguishable to guide the learning of 2D features with corresponding labels at a global level, the 2D segmentation masks supplement the Gaussians that are filtered out as floaters in 3D masks as shown in Fig. 2. This interplay allows the 2D local mask supervision and the 3D global mask supervision to mutually guide each other in learning a distinctive feature field.

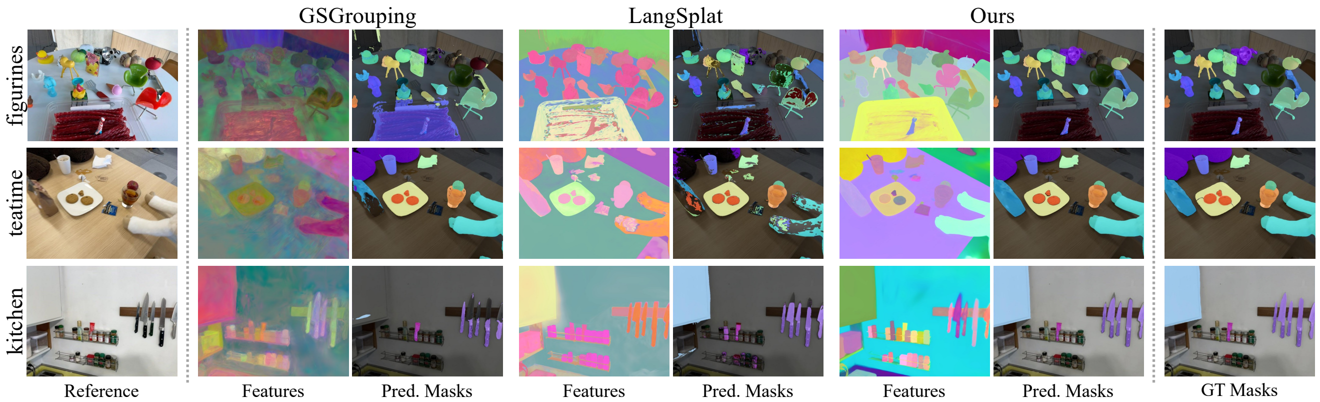


Figure 5. **Comparison on LERF-Mask Dataset.** We compare the segmentation results and visualized features (with PCA) of GSGrouping [58], LangSplat [36] and ours. Our method learns highly distinguishable features and achieves the most precise instance segmentation.

Methods	Figurines	Teatime	Kitchen	Average
Langsplat [36]	58.1	73.0	50.7	60.6
GSGrouping [58]	59.0	72.3	43.1	58.1
Ours	**85.7**	**93.7**	**77.3**	**85.6**

Table 1. **Quantitative comparison on LERF-Mask Dataset.** We report mIoU (%) metric across three scenes. Our method demonstrates a substantial improvement over baselines in instance-level segmentation, especially on complex scenes.

Model	Figurines	Teatime	Kitchen	Average
with m^{2d}_{noisy}	80.3	90.1	71.2	80.5
with M^{3d}	81.5	88.5	67.0	79.0
+ with m^{2d}_{filter}	83.9	91.4	75.4	83.6
+ with m^{2d}_{cv}	**85.7**	**93.7**	**77.3**	**85.6**

Table 2. **Ablation Study of spatial contrastive learning.** We gradually add 2D supervision into the spatial contrastive learning. m^{2d}_{filter} represents contrastive learning on the filtered 2D intra-view masks, and m^{2d}_{cv} denotes contrastive learning on the 2D cross-view masks.

Given features with instance labels $\mathcal{F} = \{f_i^j\}$, we use contrastive learning to maximize the similarity for features with the same semantic labels $\{f_i\}$ while distinguishing features from different labels with the following loss as:

$$\mathcal{L}_{CF}(\mathcal{F}) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{|\{f_i\}|}\log\frac{\exp(f_i^j \cdot \bar{f}_i/\phi_i)}{\sum_{k=1}^{N}\exp(f_i^j \cdot \bar{f}_k/\phi_k)}, \quad (6)$$

where N is the number of instances involved in $\mathcal{F}$, $\bar{f}_i$ is the mean value for f_i, and ϕ_i is the instance temperature. We apply $\mathcal{L}_{CF}$ to features with corresponding labels sampled from the single view $(\mathbf{F}_i, M_i)$, adjacent views $\bar{\mathbf{F}}_i = \{(\mathbf{F}_j, \mathcal{M}_j^{2d}) \mid j \in [i-k, i+k]\}$, and 3D Gaussian points $(\mathbf{f}_i^{3d}, \mathcal{M}_i^{3d})$, and train the feature field with loss:

$$\mathcal{L}_{\mathcal{F}} = \lambda_1\mathcal{L}_{CF}(\mathbf{F}_i) + \lambda_2\mathcal{L}_{CF}(\bar{\mathbf{F}}_i) + \lambda_3\mathcal{L}_{CF}(\mathbf{f}_i^{3d}), \quad (7)$$

where $\mathbf{f}_i^{3d}$ represents the features of visible Gaussians for frame I_i, and $\mathbf{F}_i$ is the rendered feature with Eq. 4. Upon completing the training, we perform instance segmentation by calculating cosine similarity between the average of coarse 3D instance features $\hat{f}_i^{3d}$ and each Gaussian feature, applying $\tau_{seg} = 0.9$ as the segmentation threshold. Please refer to supp. material for specific implementation details.

3.3. In-Situ Generation

We aim for the reconstruction to be not only complete, even with occlusions and partial observations, but also authentic, featuring realistic appearance and close shapes that align with the real-world (as demonstrated in Fig. 4). To fulfill all these demands, we propose a novel in-situ generation pipeline, which tames the generic 3D generative model with all known information to achieve realistic and complete instance reconstruction.

3D Generation with Omni-Conditions. We first formulate our diffusion process with omni-conditions. As the generative models are already capable of inferring 3D geometric distributions from a single image, given all known information of the partial reconstructed instance $\mathcal{I}$ (*e.g.*, views $\{y^k\}$ and depths $\{d^k\}$ rendered in visible viewpoints $\{\pi^k\}$), we aim to control the 3D diffusion model [10] with these available omni-conditions to predict the unseen regions $p(\{x^n\}|\{y^k, d^k, \hat{\pi}_n^k\})$ with precise alignment to real-world identity, where $\{x^n\}$ is the unseen views in $\{\pi^n\}$, and $\{\hat{\pi}_n^k\}$ is the relative pose between input and target viewpoint. To complement the generation model with known visual observations, we first design an alternated view conditioning strategy, which sequentially feeds optimal views re-rendered with instance's 2DGS as alternated conditions through the denoising process, as shown in the Step-3 of Fig. 3. For each timestep, we average the noise predictions across each target view $\{\epsilon_{\theta}{}_n^k\}$, as:

$$\begin{aligned} x_{t-1}^n &= \frac{1}{\sqrt{\alpha_t}}x_t^n - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\bar{\epsilon}_\theta^n, \\ \bar{\epsilon}_\theta^n &= \frac{1}{N_k}\textstyle\sum_{k=1}^{N_k}\epsilon_\theta^n(x_t^n, y^k, \hat{\pi}_n^k), \end{aligned} \quad (8)$$

where ϵ_θ is the noise predictor and $\bar{\epsilon}_\theta^n$ represents the averaged predicted noise for target view x^n.

Complement Known Features via Geometry Cues. Even with complemented views as conditions, the predicted views from the generative model may still deviate from the real observation due to the domain gap between training data and the real-world scene (see Fig. 9). To mitigate this, we leverage geometric cues from existing Gaussian scenes and design a geometry-aware feature warping strategy, which further enforces consistency from the known observation to the generative prediction. During each diffusion iteration, time-dependent noise is added to the input views' latent features, which are then projected onto visible pixels of target views using rendering depths. Since 2DGS yields a fused mesh, we use the surface normal of

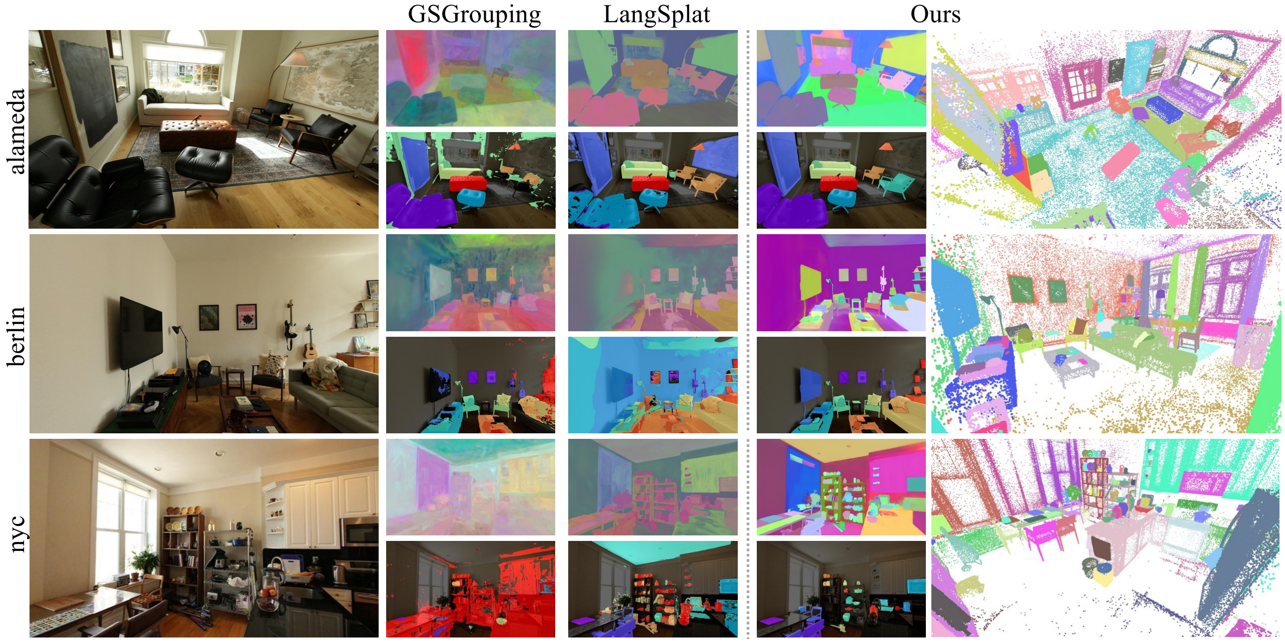


Figure 6. **Comparison on ZipNeRF Dataset.** The baselines yield highly noisy segmentation results in such complex scenes, while our method achieves fine-grained instance-level segmentation. We also show the segmented Gaussian points of our method.

the viewing-ray-mesh intersection to discard projections indicating back-facing surface. This enforces consistency in visible regions, while invisible parts are initialized as random noise and are gradually denoised through guidance from both alternated condition views and the warped latent features from visible pixels. See supp. for more details.

Occlusion-Aware Viewpoint Selection and Joint Refinement. Unlike the typical image-to-3D pipeline that takes a standard elevated view as input, occlusions in cluttered scenes make the optimal viewpoints for each instance not straightforwardly available. Hence, we set up the following principles that select the appropriate viewpoints as ideal input conditions for each instance. First, inspired by the target view setup in [24], we select 16 viewpoints centered around the segmented object. Among these viewpoints, those with the least scene occlusion between the viewpoint and the object are selected as the ideal input conditions, while the remaining viewpoints, subject to occlusion, are considered unseen and require supplementation from the 3D generative prior as explained above. We filter out the background in the selected rendering views with the 2D instance mask extracted from the rendered features. Finally, we jointly refine the instance's 2DGS with both the source observations and the generated views. The source observation ensures consistency for the visible parts, while the generated views are responsible for supplementing the unseen parts. Thus, we achieve a complete instance reconstruction guided by the known information. See more details in supp. material.

4. Experiments

We compare InstaScene for instance-aware scene reconstruction in two folds. The first section focuses on fine-grained instance decomposition from scene reconstruction. The second section analyzes the in-situ generation for instance-level complete reconstruction, which measures the quality and accuracy of recovering decomposed objects.

4.1. Fine-Grained Instance Decomposition

Datasets. To demonstrate our performance in 3D instance segmentation, we conduct a quantitative comparison using LERF-Mask Dataset [15]. We select three instance-rich scenes: figurines, waldo-kitchen, and teatime, and manually re-annotate them with instance-level ground truth masks. Following [5], we extract the instance features from a given reference view, generate the corresponding instance masks in the target view using cosine similarity, and calculate mIoU between the extracted masks and groundtruth masks as the evaluation metric. To further demonstrate the robustness of our method in complex environments, we present qualitative comparison results with three scenes selected from the ZipNeRF Dataset [2], which feature a variety of objects. We also conduct a comparison on 3D-OVS Dataset [22] in supp. material.

Baselines. We compare our approach with the state-of-the-art Gaussian Splatting-based 3D segmentation methods [36, 58]. For a fair comparison, all methods optimize the feature field based on our pre-trained Gaussian Splatting models and use the same instance segmentation masks [35] as 2D semantic priors for supervision.

Comparison Results. As shown in Fig. 5, the mask generated by GSGrouping [58] suffers from issues such as floaters and missing segments. This is mainly due to the inconsistency of its object-tracking, which is inherently sensitive to mask noise and frame discontinuity. LangSplat mislabels repeated instances of similar objects within a scene, such as the bottles in the kitchen in Fig. 5. Both methods exhibit poor feature smoothness and the predicted results contain significant noise, particularly in complex scenarios as shown in Fig. 6. In contrast, our method enhances the 2D semantic priors with minimal cost and achieves the most

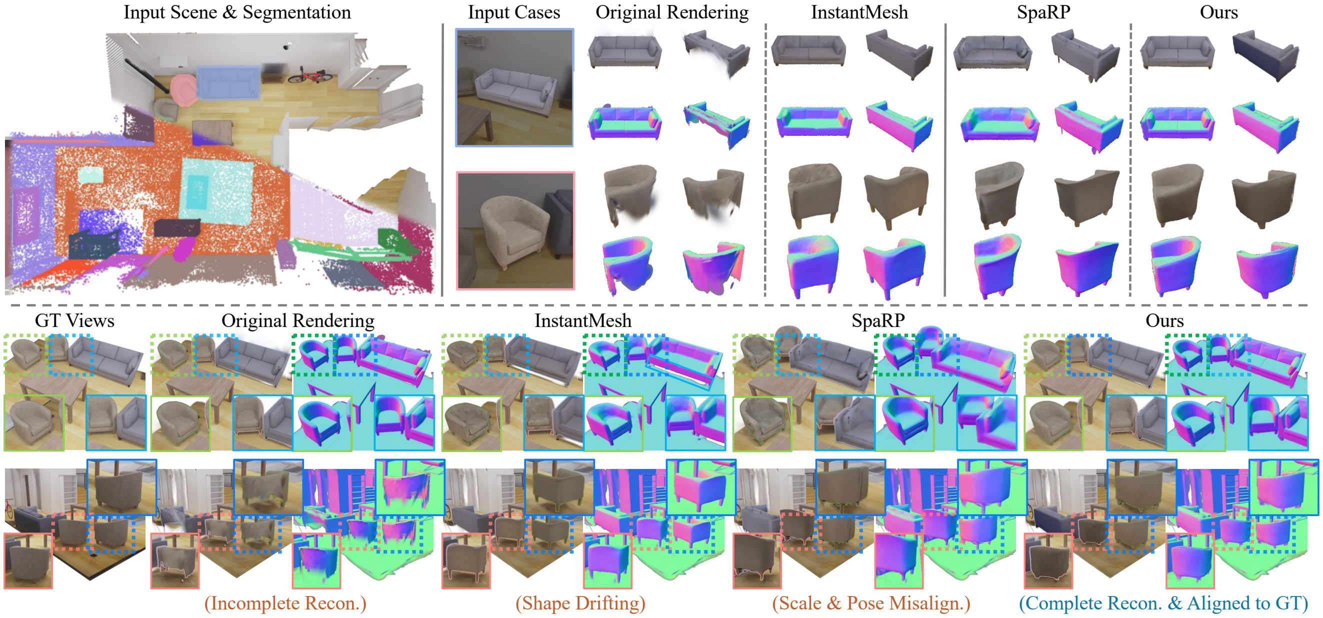


Figure 7. **Comparisons on Replica-CAD Dataset [45].** Instances extracted from the decomposition suffer from incomplete reconstruction, our method preserves the quality of the originally visible regions while achieving most plausible recovery of unknown regions.

Methods		PSNR↑		SSIM↑		LPIPS↓		CD↓	F1-Score↑	Volume IoU↑
		Known	Unknown	Known	Unknown	Known	Unknown			
Origin Recon.	2DGS [12]	31.67	27.44	0.976	0.918	0.034	0.093	0.028	0.734	0.361
Single-View	MVDFusion [10]	17.19	17.46	0.797	0.787	0.232	0.251	0.081	0.150	0.531
	InstantMesh [53]	23.05	22.83	0.853	0.862	0.129	0.139	0.045	0.382	0.570
Multi-View	SparP [52]	25.09	23.03	0.881	0.868	0.112	0.129	0.037	0.406	0.590
	Ours	**32.57**	**29.02**	**0.979**	**0.944**	**0.028**	**0.066**	**0.016**	**0.767**	**0.716**

Table 3. **Quantitative comparison of In-situ Generation.** Our in-situ generation achieves superior appearance and geometric alignment with the original scene while ensuring the highest quality recovery in the unknown region.

distinguishable feature field, enabling fine-grained instance segmentation even in challenging scenes with frequent adjacent and occluded objects (such as utensils on the shelf in nyc) as demonstrated in Fig. 6. Please refer to the supplementary material for more results.

Ablation Study. To demonstrate the effectiveness of spatial contrastive learning with mutual guidance, we conduct experiments by applying contrastive learning on raw noisy 2D segmentation masks and the preprocessed coarse 3D instance priors. The qualitative results are presented in Fig. 2, which shows the improvement of segmentation granularity and feature distinctiveness with mutual guidance. We also inspect the efficacy of filtered 2D masks and cross-view masks during the spatial contrastive learning in Tab. 2, which also proves that the design of mutual guidance significantly improves the distinctiveness of the feature field, leading to more accurate instance-level segmentation.

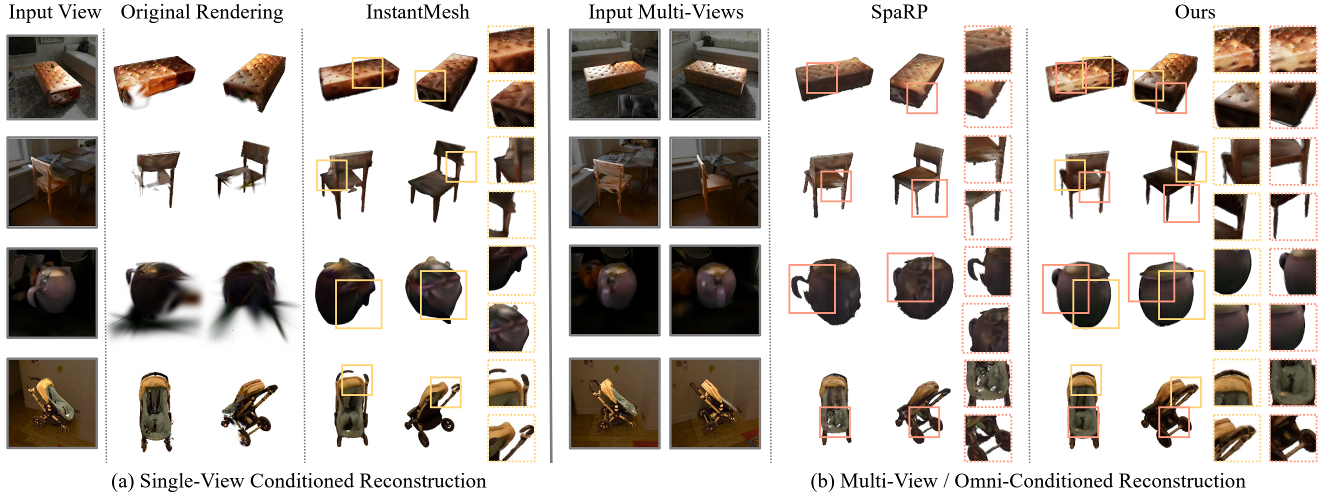
4.2. Instance-Level Complete Reconstruction

In this section, we analyze the instance completion quality of in-situ generation. As demonstrated in Fig. 4, different from generic reconstruction or image-to-3D tasks, for instance-level complete reconstruction in a certain environment, the reconstructed object should be complete and also align to the real-world scenes with precise geometric shape and close appearance.

Baselines. For single-view conditioned generation methods, we compare our approach with MVDFusion [10] and InstantMesh [53]. We select the rendering view with the least occlusion and the largest object coverage in our filtered viewpoints as input. For multi-view conditioned reconstruction methods, we compare our approach with SpaRP [52], which implicitly infers the 3D spatial relationships among the given sparse views and uses it to accomplish 3D reconstruction. For a fair comparison, we use the same optimal views filtered by our method as the multi-view inputs.

Metrics. We focus on two aspects, the completion quality and the alignment with the original scene, including both rendering and scale. Following existing reconstruction [12, 47] and generative methods [10, 24, 53], we use Chamfer Distance (CD), F1-Score and Volumetric Intersection over Union (Volume IoU) between GT shapes and instance reconstruction to evaluate the completion quality in unseen regions and assess the spatial alignment. Additionally, we compute PSNR, SSIM, and LPIPS for both known and unknown views (as set up in Viewpoint Selection) to evaluate the appearance alignment and completion quality.

Datasets. We conduct the quantitative comparison on the Replica-CAD Dataset [45], which is a synthetic scene composed of artist-recreated scanned objects. Furthermore, we conduct a qualitative comparison on diverse objects across various real-world complex scenes in the ZipNeRF Dataset [2]. We also conduct user studies to measure the



(a) Single-View Conditioned Reconstruction (b) Multi-View / Omni-Conditioned Reconstruction

Figure 8. **Comparison with different generation methods.** We show the complete reconstruction results for each method on diverse instances. Our method not only achieves faithful completion but also maintains consistency with the original scene rendering.

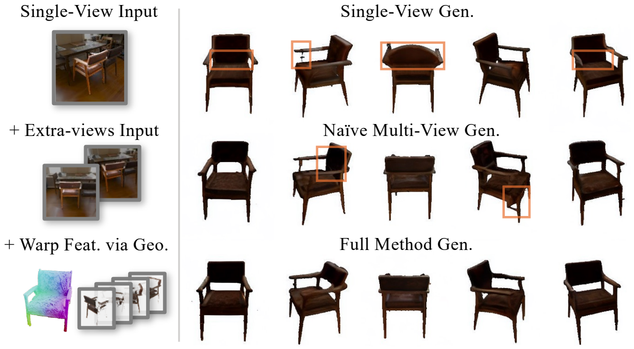


Figure 9. **Ablation studies of the omni-conditioned completion from in-situ generation.** Naïvely alternating the condition views resolves the unrealistic predictions of single-view input, and the geometry-aware feature warping further enhances the consistency of the generated views.

completion quality, please refer to the supplementary material for details.

Comparison Results. As shown in Fig. 8, although most methods recover geometry for common objects, such as the leather stool in the first row, only our approach successfully maintains consistent rendering results such as shiny leather texture. When reconstructing a partially observed object such as the chair in the second row, the results from InstantMesh exhibit significant misalignment and contain noticeable floaters. In the case of the teapot from the third row, even with multi-view conditions, SpaRP fails to recover the geometry, which is mainly due to the domain gap between the generative priors and real world. Our method, by employing the warped features from geometric cues, imposes a geometry-aware constraint and recovers reasonable geometry even under conditions of poor observations. Moreover, for unusual and complex objects like the baby carriage in the last row, our method recovers the unseen regions while preserving the most realistic rendering quality, whereas other methods suffer from issues such as holes or misalignment. Fig. 7 and metrics shown in Table. 3 further demonstrate that our method maintains alignment with the physical world while achieving the most plausible prediction for unseen regions. See more results in supp. material.

Ablation Study. We present the qualitative comparison of our omni-conditioned completion from the in-situ generation in Fig. 9. The original single-view conditioned method [10] suffers from inconsistencies between generated views and produces unreasonable results. The naive multi-view conditioned approach, which alternately uses additional input views as conditions, avoids unreasonable prediction but still results in floaters and inconsistencies. By further adding warped features via geometric cues, which enforces feature consistency from the known observation to the generation, our method significantly enhances the coherence and plausibility of the generated results.

5. Conclusion

We have proposed a novel open-set scene decomposition and reconstruction framework, **InstaScene**. InstaScene allows users to pick up arbitrary objects from cluttered Gaussian Splatting scenes, and produces complete geometry and appearance of the instance which aligns with the physical world. The key insight is to trace Gaussian rasterization during mask clustering, which derives spatial contrastive learning that produces highly distinguishable feature fields. To recover complete objects from occluded observation, we also proposed a novel in-situ generation which fully leverages known information to coordinate the 3D generative prior with faithful instance completion. As a limitation, we cannot decompose dynamic, transparent or highly reflective objects (see the supplementary material for details). A potential solution is to incorporate 4D or physically-based representation for these cases, which is left as future work.

6. Acknowledgments

We express our gratitude to all the anonymous reviewers for their professional and constructive comments. This work was partially supported by the NSFC (No. 62441222), “Innovation Yongjiang 2035” Key R&D Programme (No. 2025Z062), Information Technology Center, and State Key Lab of CAD&CG, Zhejiang University.

References

[1] Armen Avetisyan, Manuel Dahnert, Angela Dai, Manolis Savva, Angel X Chang, and Matthias Nießner. Scan2cad: Learning cad model alignment in rgb-d scans. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 2614–2623, 2019. 2

[2] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Zip-nerf: Anti-aliased grid-based neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19697–19705, 2023. 6, 7

[3] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 3

[4] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 2

[5] Seokhun Choi, Hyeonseop Song, Jaechul Kim, Taehyeong Kim, and Hoseok Do. Click-gaussian: Interactive segmentation to any 3d gaussians. In *European Conference on Computer Vision*, pages 289–305. Springer, 2025. 3, 4, 6

[6] Wenqi Dong, Bangbang Yang, Lin Ma, Xiao Liu, Liyuan Cui, Hujun Bao, Yuewen Ma, and Zhaopeng Cui. Coin3d: Controllable and interactive 3d assets generation with proxy-guided conditioning. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–10, 2024. 3

[7] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, pages 226–231, 1996. 4

[8] Huan Fu, Bowen Cai, Lin Gao, Ling-Xiao Zhang, Jiaming Wang, Cao Li, Qixun Zeng, Chengyue Sun, Rongfei Jia, Binqiang Zhao, et al. 3d-front: 3d furnished rooms with layouts and semantics. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10933–10942, 2021. 2

[9] Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. Lrm: Large reconstruction model for single image to 3d. *arXiv preprint arXiv:2311.04400*, 2023. 1, 2

[10] Hanzhe Hu, Zhizhuo Zhou, Varun Jampani, and Shubham Tulsiani. Mvd-fusion: Single-view 3d via depth-consistent multi-view generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9698–9707, 2024. 3, 5, 7, 8

[11] Yubin Hu, Sheng Ye, Wang Zhao, Matthieu Lin, Yuze He, Yu-Hui Wen, Ying He, and Yong-Jin Liu. Oˆ 2-recon: Completing 3d reconstruction of occluded objects in the scene with a pre-trained 2d diffusion model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2285–2293, 2024. 3

[12] Binbin Huang, Zehao Yu, Anpei Chen, Andreas Geiger, and Shenghua Gao. 2d gaussian splatting for geometrically accurate radiance fields. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024. 3, 7

[13] Mazeyu Ji, Ri-Zhao Qiu, Xueyan Zou, and Xiaolong Wang. Graspsplats: Efficient manipulation with 3d feature splatting. *arXiv preprint arXiv:2409.02084*, 2024. 1

[14] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023. 1, 2, 3

[15] Justin Kerr, Chung Min Kim, Ken Goldberg, Angjoo Kanazawa, and Matthew Tancik. Lerf: Language embedded radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19729–19739, 2023. 6

[16] Chung Min Kim, Mingxuan Wu, Justin Kerr, Ken Goldberg, Matthew Tancik, and Angjoo Kanazawa. Garfield: Group anything with radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21530–21539, 2024. 3, 4

[17] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 3

[18] Oliver Kroemer, Scott Niekum, and George Konidaris. A review of robot learning for manipulation: Challenges, representations, and algorithms. *Journal of machine learning research*, 22(30):1–82, 2021. 1

[19] Zizhang Li, Xiaoyang Lyu, Yuanyuan Ding, Mengmeng Wang, Yiyi Liao, and Yong Liu. Rico: Regularizing the unobservable for indoor compositional reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17761–17771, 2023. 2

[20] Anran Liu, Cheng Lin, Yuan Liu, Xiaoxiao Long, Zhiyang Dou, Hao-Xiang Guo, Ping Luo, and Wenping Wang. Part123: part-aware 3d reconstruction from a single-view image. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–12, 2024. 3, 4

[21] Haolin Liu, Chongjie Ye, Yinyu Nie, Yingfan He, and Xiaoguang Han. Lasa: Instance reconstruction from real scans using a large-scale aligned shape annotation dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20454–20464, 2024. 1, 2

[22] Kunhao Liu, Fangneng Zhan, Jiahui Zhang, Muyu Xu, Yingchen Yu, Abdulmotaleb El Saddik, Christian Theobalt, Eric Xing, and Shijian Lu. Weakly supervised 3d open-vocabulary segmentation. *Advances in Neural Information Processing Systems*, 36:53433–53456, 2023. 6

[23] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9298–9309, 2023. 3

[24] Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. Syncdreamer: Generating multiview-consistent images from a single-view image. *arXiv preprint arXiv:2309.03453*, 2023. 3, 6, 7

[25] Zhiheng Liu, Hao Ouyang, Qiuyu Wang, Ka Leong Cheng, Jie Xiao, Kai Zhu, Nan Xue, Yu Liu, Yujun Shen, and Yang Cao. Infusion: Inpainting 3d gaussians via learning depth completion from diffusion prior. *arXiv preprint arXiv:2404.11613*, 2024. 3

[26] Shiyang Lu, Haonan Chang, Eric Pu Jing, Abdeslam Boularias, and Kostas Bekris. Ovir-3d: Open-vocabulary 3d instance retrieval without training on 3d data. In *Conference on Robot Learning*, pages 1610–1620. PMLR, 2023. 4

[27] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11461–11471, 2022. 3

[28] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 1, 2

[29] Ashkan Mirzaei, Tristan Aumentado-Armstrong, Konstantinos G Derpanis, Jonathan Kelly, Marcus A Brubaker, Igor Gilitschenski, and Alex 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. 3

[30] Phuc Nguyen, Tuan Duc Ngo, Evangelos Kalogerakis, Chuang Gan, Anh Tran, Cuong Pham, and Khoi Nguyen. Open3dis: Open-vocabulary 3d instance segmentation with 2d mask guidance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4018–4028, 2024. 4

[31] Junfeng Ni, Yu Liu, Ruijie Lu, Zirui Zhou, Song-Chun Zhu, Yixin Chen, and Siyuan Huang. Decompositional neural scene reconstruction with generative diffusion prior. In *CVPR*, 2025. 3

[32] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 3

[33] Songyou Peng, Kyle Genova, Chiyu Jiang, Andrea Tagliasacchi, Marc Pollefeys, Thomas Funkhouser, et al. Openscene: 3d scene understanding with open vocabularies. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 815–824, 2023. 1

[34] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 1, 2

[35] Lu Qi, Jason Kuen, Weidong Guo, Tiancheng Shen, Jiuxiang Gu, Jiaya Jia, Zhe Lin, and Ming-Hsuan Yang. High-quality entity segmentation. *arXiv preprint arXiv:2211.05776*, 2022. 4, 6

[36] Minghan Qin, Wanhua Li, Jiawei Zhou, Haoqian Wang, and Hanspeter Pfister. Langsplat: 3d language gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20051–20060, 2024. 3, 5, 6

[37] Ri-Zhao Qiu, Ge Yang, Weijia Zeng, and Xiaolong Wang. Feature splatting: Language-driven physics-based scene synthesis and editing. *arXiv preprint arXiv:2404.01223*, 2024. 3

[38] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3

[39] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4104–4113, 2016. 1

[40] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision (ECCV)*, 2016. 2

[41] Jin-Chuan Shi, Miao Wang, Hao-Bin Duan, and Shao-Hua Guan. Language embedded 3d gaussians for open-vocabulary scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5333–5343, 2024. 3

[42] Ruoxi Shi, Hansheng Chen, Zhuoyang Zhang, Minghua Liu, Chao Xu, Xinyue Wei, Linghao Chen, Chong Zeng, and Hao Su. Zero123++: a single image to consistent multi-view diffusion base model. *arXiv preprint arXiv:2310.15110*, 2023. 3

[43] Olaolu Shorinwa, Johnathan Tucker, Aliyah Smith, Aiden Swann, Timothy Chen, Roya Firoozi, Monroe David Kennedy, and Mac Schwager. Splat-mover: multi-stage, open-vocabulary robotic manipulation via editable gaussian splatting. In *8th Annual Conference on Robot Learning*, 2024. 1

[44] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2149–2159, 2022. 3

[45] Andrew Szot, Alex Clegg, Eric Undersander, Erik Wijmans, Yili Zhao, John Turner, Noah Maestre, Mustafa Mukadam, Devendra Chaplot, Oleksandr Maksymets, Aaron Gokaslan, Vladimir Vondrus, Sameer Dharur, Franziska Meier, Wojciech Galuba, Angel Chang, Zsolt Kira, Vladlen Koltun, Jitendra Malik, Manolis Savva, and Dhruv Batra. Habitat 2.0: Training home assistants to rearrange their habitat. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 7

[46] Ayça Takmaz, Elisabetta Fedele, Robert W Sumner, Marc Pollefeys, Federico Tombari, and Francis Engelmann. Openmask3d: Open-vocabulary 3d instance segmentation. *arXiv preprint arXiv:2306.13631*, 2023. 1

[47] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *arXiv preprint arXiv:2106.10689*, 2021. 1, 7

[48] Ethan Weber, Aleksander Holynski, Varun Jampani, Saurabh Saxena, Noah Snavely, Abhishek Kar, and Angjoo Kanazawa. Nerfiller: Completing scenes via generative 3d inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20731–20741, 2024. 3

[49] Qianyi Wu, Xian Liu, Yuedong Chen, Kejie Li, Chuanxia Zheng, Jianfei Cai, and Jianmin Zheng. Object-compositional neural implicit surfaces. In *European Conference on Computer Vision*, pages 197–213. Springer, 2022. 2

[50] Qianyi Wu, Kaisiyuan Wang, Kejie Li, Jianmin Zheng, and Jianfei Cai. Objectsdf++: Improved object-compositional neural implicit surfaces. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21764–21774, 2023. 2

[51] Yanmin Wu, Jiarui Meng, Haijie Li, Chenming Wu, Yahao Shi, Xinhua Cheng, Chen Zhao, Haocheng Feng, Errui Ding, Jingdong Wang, et al. Opengaussian: Towards point-level 3d gaussian-based open vocabulary understanding. *arXiv preprint arXiv:2406.02058*, 2024. 3, 4

[52] Chao Xu, Ang Li, Linghao Chen, Yulin Liu, Ruoxi Shi, Hao Su, and Minghua Liu. Sparp: Fast 3d object reconstruction and pose estimation from sparse views. In *European Conference on Computer Vision*, pages 143–163. Springer, 2025. 4, 7

[53] Jiale Xu, Weihao Cheng, Yiming Gao, Xintao Wang, Shenghua Gao, and Ying Shan. Instantmesh: Efficient 3d mesh generation from a single image with sparse-view large reconstruction models. *arXiv preprint arXiv:2404.07191*, 2024. 4, 7

[54] Mi Yan, Jiazhao Zhang, Yan Zhu, and He Wang. Maskclustering: View consensus based mask graph clustering for open-vocabulary 3d instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28274–28284, 2024. 4

[55] Yunzhi Yan, Haotong Lin, Chenxu Zhou, Weijie Wang, Haiyang Sun, Kun Zhan, Xianpeng Lang, Xiaowei Zhou, and Sida Peng. Street gaussians for modeling dynamic urban scenes. *arXiv preprint arXiv:2401.01339*, 2024. 1

[56] Bangbang Yang, Yinda Zhang, Yinghao Xu, Yijin Li, Han Zhou, Hujun Bao, Guofeng Zhang, and Zhaopeng Cui. Learning object-compositional neural radiance field for editable scene rendering. In *International Conference on Computer Vision (ICCV)*, 2021. 2

[57] Yunhan Yang, Xiaoyang Wu, Tong He, Hengshuang Zhao, and Xihui Liu. Sam3d: Segment anything in 3d scenes. *arXiv preprint arXiv:2306.03908*, 2023. 4

[58] Mingqiao Ye, Martin Danelljan, Fisher Yu, and Lei Ke. Gaussian grouping: Segment and edit anything in 3d scenes. In *European Conference on Computer Vision*, pages 162–179. Springer, 2025. 3, 5, 6

[59] Yingda Yin, Yuzheng Liu, Yang Xiao, Daniel Cohen-Or, Jingwei Huang, and Baoquan Chen. Sai3d: Segment any instance in 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3292–3302, 2024. 4

[60] Haiyang Ying, Yixuan Yin, Jinzhi Zhang, Fan Wang, Tao Yu, Ruqi Huang, and Lu 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. 3, 4

[61] Longwen Zhang, Ziyu Wang, Qixuan Zhang, Qiwei Qiu, Anqi Pang, Haoran Jiang, Wei Yang, Lan Xu, and Jingyi Yu. Clay: A controllable large-scale generative model for creating high-quality 3d assets. *ACM Transactions on Graphics (TOG)*, 43(4):1–20, 2024. 1, 2

[62] Shijie Zhou, Haoran Chang, Sicheng Jiang, Zhiwen Fan, Zehao Zhu, Dejia Xu, Pradyumna Chari, Suya You, Zhangyang Wang, and Achuta Kadambi. Feature 3dgs: Supercharging 3d gaussian splatting to enable distilled feature fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21676–21685, 2024. 3

[63] Xiaoyu Zhou, Zhiwei Lin, Xiaojun Shan, Yongtao Wang, Deqing Sun, and Ming-Hsuan Yang. Drivinggaussian: Composite gaussian splatting for surrounding dynamic autonomous driving scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21634–21643, 2024. 1