GaussianObject: High-Quality 3D Object Reconstruction from Four Views with Gaussian Splatting

CHEN YANG*, MoE Key Lab of Artificial Intelligence, AI Institute, SJTU, China
SIKUANG LI*, MoE Key Lab of Artificial Intelligence, AI Institute, SJTU, China
JIEMIN FANG†, Huawei Inc., China
RUOFAN LIANG, University of Toronto, Canada
LINGXI XIE, Huawei Inc., China
XIAOPENG ZHANG, Huawei Inc., China
WEI SHEN‡, MoE Key Lab of Artificial Intelligence, AI Institute, SJTU, China
QI TIAN, Huawei Inc., China

(a) 4 Input Views (b) Ground Truth (c) DietNeRF (d) FreeNeRF (e) 3DGS (f) Ours

Fig. 1. We introduce GaussianObject, a framework capable of reconstructing high-quality 3D objects from only 4 images with Gaussian splatting. GaussianObject demonstrates superior performance over previous state-of-the-art (SOTA) methods on challenging objects.

Reconstructing and rendering 3D objects from highly sparse views is of critical importance for promoting applications of 3D vision techniques and improving user experience. However, images from sparse views only contain very limited 3D information, leading to two significant challenges: 1) Difficulty in building multi-view consistency as images for matching are too few; 2) Partially omitted or highly compressed object information as view coverage is insufficient. To tackle these challenges, we propose GaussianObject, a framework to represent and render the 3D object with Gaussian splatting that achieves high rendering quality with only 4 input images. We first introduce techniques of visual hull and floater elimination, which explicitly inject structure priors into the initial optimization process to help build multi-view consistency, yielding a coarse 3D Gaussian representation. Then we construct a Gaussian repair model based on diffusion models to supplement the omitted object information, where Gaussians are further refined. We design a self-generating strategy to obtain image pairs for training the repair model. We further design a COLMAP-free variant, where pre-given accurate camera poses are not required, which achieves competitive quality and facilitates wider applications. GaussianObject is evaluated on several challenging datasets, including MipNeRF360, OmniObject3D, OpenIllumination, and our-collected unposed images, achieving superior performance from only four views and significantly outperforming previous SOTA methods.

CCS Concepts: • **Computing methodologies** → **Reconstruction**; **Rendering**; **Point-based models**.

*Equal contributions.
†Project lead.
‡Corresponding author.

Authors' addresses: Chen Yang, MoE Key Lab of Artificial Intelligence, AI Institute, SJTU, Shanghai, China, ycyangchen@sjtu.edu.cn; Sikuang Li, MoE Key Lab of Artificial Intelligence, AI Institute, SJTU, Shanghai, China, uranusits@sjtu.edu.cn; Jiemin Fang, Huawei Inc., Wuhan, China, jaminfong@gmail.com; Ruofan Liang, University of Toronto, Toronto, Canada, ruofan@cs.toronto.edu; Lingxi Xie, Huawei Inc., Beijing, China, 198808xc@gmail.com; Xiaopeng Zhang, Huawei Inc., Shanghai, China, zxphistory@gmail.com; Wei Shen, MoE Key Lab of Artificial Intelligence, AI Institute, SJTU, Shanghai, China, wei.shen@sjtu.edu.cn; Qi Tian, Huawei Inc., Shenzhen, China, tian.qi1@huawei.com.

ACM 0730-0301/2024/12-ART
https://doi.org/10.1145/3687759

Additional Key Words and Phrases: Sparse view reconstruction, 3D Gaussian Splatting, ControlNet, Visual hull, Novel view synthesis

ACM Reference Format:
Chen Yang, Sikuang Li, Jiemin Fang, Ruofan Liang, Lingxi Xie, Xiaopeng Zhang, Wei Shen, and Qi Tian. 2024. GaussianObject: High-Quality 3D Object Reconstruction from Four Views with Gaussian Splatting. *ACM Trans. Graph.* 43, 6 (December 2024), 12 pages. https://doi.org/10.1145/3687759

1 INTRODUCTION

Reconstructing and rendering 3D objects from 2D images has been a long-standing and important topic, which plays critical roles in a vast range of real-life applications. One key factor that impedes users, especially ones without expert knowledge, from widely using these techniques is that usually dozens of multi-view images need to be captured, which is cumbersome and sometimes impractical. Efficiently reconstructing high-quality 3D objects from highly sparse captured images is of great value for expediting downstream applications such as 3D asset creation for game/movie production and AR/VR products.

In recent years, a series of methods [Guangcong et al. 2023; Jain et al. 2021; Niemeyer et al. 2022; Shi et al. 2024b; Song et al. 2023b; Yang et al. 2023; Zhou and Tulsiani 2023; Zhu et al. 2024] have been proposed to reduce reliance on dense captures. However, it is still challenging to produce high-quality 3D objects when the views become **extremely sparse**, *e.g.* only 4 images in a 360° range, as shown in Fig. 1. We delve into the task of sparse-view reconstruction and discover two main challenges behind it. The first one lies in the difficulty of building multi-view consistency from highly sparse input. The 3D representation is easy to overfit the input images and degrades into fragmented pixel patches of training views without reasonable structures. The other challenge is that with sparse captures in a 360° range, some content of the object can be inevitably omitted or severely compressed when observed from extreme views[1]. The omitted or compressed information is impossible or hard to be reconstructed in 3D only from the input images.

To tackle the aforementioned challenges, we introduce GaussianObject, a novel framework designed to reconstruct high-quality 3D objects from as few as 4 input images. We choose 3D Gaussian splatting (3DGS) [Kerbl et al. 2023] as the basic representation as it is fast and, more importantly, explicit enough. Benefiting from its point-like structure, we design several techniques for introducing object structure priors, *e.g.* the basic/rough geometry of the object, to help build multi-view consistency, including visual hull [Laurentini 1994] to locate Gaussians within the object outline and floater elimination to remove outliers. To erase artifacts caused by omitted or highly compressed object information, we propose a Gaussian repair model driven by 2D large diffusion models [Rombach et al. 2022], translating corrupted rendered images into high-fidelity ones. As normal diffusion models lack the ability to repair corrupted images, we design self-generating strategies to construct image pairs to tune the diffusion models, including rendering images from leave-one-out training models and adding 3D noises to Gaussian attributes. Images generated from the repair model can be used to refine the 3D Gaussians optimized with structure priors, where the rendering quality can be further improved. To further extend GaussianObject to practical applications, we introduce a COLMAP-free variant of GaussianObject (CF-GaussianObject), which achieves competitive reconstruction performance on challenging datasets with only four input images without inputting accurate camera parameters.

Our contributions are summarized as follows:

- We optimize 3D Gaussians from highly sparse views using explicit structure priors, introducing techniques of visual hull for initialization and floater elimination for training.
- We propose a Gaussian repair model based on diffusion models to remove artifacts caused by omitted or highly compressed information, where the rendering quality can be further improved.
- The overall framework GaussianObject consistently outperforms current SOTA methods on several challenging real-world datasets, both qualitatively and quantitatively. A COLMAP-free variant is further presented for wider applications, weakening the requirement of accurate camera poses.

2 RELATED WORK

Vanilla NeRF struggles in sparse settings. Techniques like Deng et al. [2022]; Roessle et al. [2022]; Somraj et al. [2024, 2023]; Somraj and Soundararajan [2023] use Structure from Motion (SfM) [Schönberger and Frahm 2016] derived visibility or depth and mainly focus on closely aligned views. Xu et al. [2022] uses ground truth depth maps, which are costly to obtain in real-world images. Some methods [Guangcong et al. 2023; Song et al. 2023b] estimate depths with monocular depth estimation models [Ranftl et al. 2021, 2022] or sensors, but these are often too coarse. Jain et al. [2021] uses a vision-language model [Radford et al. 2021] for unseen view rendering, but the semantic consistency is too high-level to guide low-level reconstruction. Shi et al. [2024b] combines a deep image prior with factorized NeRF, effectively capturing overall appearance but missing fine details in input views. Priors based on information theory [Kim et al. 2022], continuity [Niemeyer et al. 2022], symmetry [Seo et al. 2023], and frequency regularization [Song et al. 2023a; Yang et al. 2023] are only effective for specific scenarios, limiting their further applications. Besides, there are some methods [Jang and Agapito 2024; Jiang et al. 2024; Xu et al. 2024c; Zou et al. 2024] that employ Vision Transformer (ViT) [Dosovitskiy et al. 2021] to reduce the requirements for constructing NeRFs and Gaussians.

The recent progress in diffusion models has spurred notable advancements in 3D applications. Dreamfusion [Poole et al. 2023] proposes Score Distillation Sampling (SDS) for distilling NeRFs with 2D priors from a pre-trained diffusion model for 3D object generation from text prompts. It has been further refined for text-to-3D [Chen et al. 2023; Lin et al. 2023; Metzer et al. 2023; Shi et al. 2024a; Tang et al. 2024b; Wang et al. 2023a,b; Yi et al. 2024] and 3D/4D editing [Haque et al. 2023; Shao et al. 2024] by various studies, demonstrating the versatility of 2D diffusion models in 3D contexts. Burgess et al. [2024]; Chan et al. [2023]; Liu et al. [2023c]; Müller et al. [2024]; Pan et al. [2024]; Zhu and Zhuang [2024] have adapted these methods for 3D generation and view synthesis from a single

[1]When the view is orthogonal to the surface of the object, the observed information attached to the surface can be largely preserved; On the contrary, the information will be severely compressed.

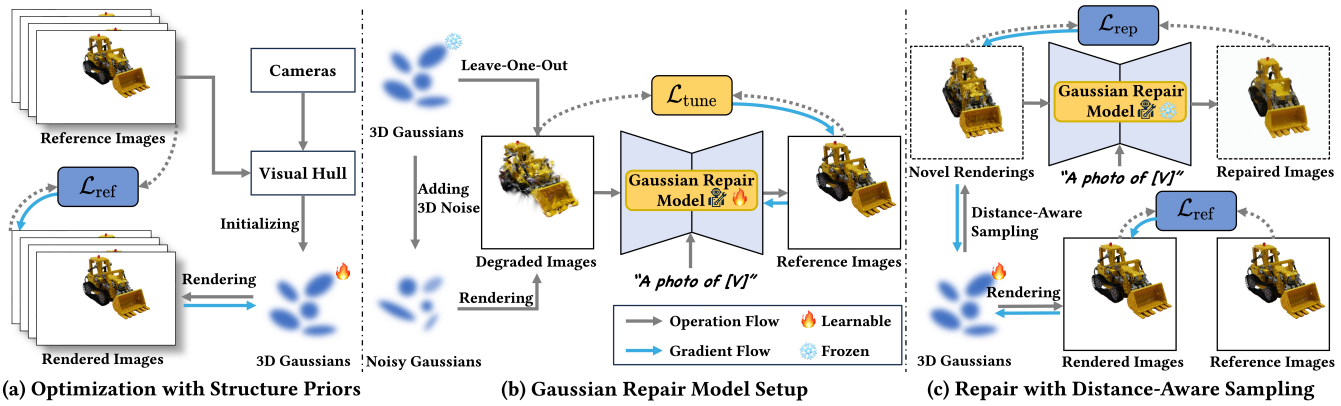

Fig. 2. Overview of GaussianObject. (a) We initialize 3D Gaussians by constructing a visual hull with camera parameters and masked images, which are optimized with $\mathcal{L}_{\text{ref}}$ and refined through floater elimination. (b) We use a novel 'leave-one-out' strategy and add 3D noise to Gaussians to generate corrupted Gaussian renderings. These renderings, paired with their corresponding reference images, facilitate the training of the Gaussian repair model employing $\mathcal{L}_{\text{tune}}$. For details please refer to Fig. 3. (c) Once trained, the Gaussian repair model is frozen and used to correct views that need to be rectified. These views are identified through distance-aware sampling. The repaired images and reference images are used to further optimize 3D Gaussians with $\mathcal{L}_{\text{rep}}$ and $\mathcal{L}_{\text{ref}}$.

image, while they often have strict input requirements and can produce overly saturated images. In sparse reconstruction, approaches like DiffusioNeRF [Wynn and Turmukhambetov 2023], SparseFusion [Zhou and Tulsiani 2023], Deceptive-NeRF [Liu et al. 2023b], ReconFusion [Wu et al. 2024] and CAT3D [Gao et al. 2024] integrate diffusion models with NeRFs. Recently, Large reconstruction models (LRMs) [Hong et al. 2024; Li et al. 2024; Tang et al. 2024a; Wang et al. 2024b; Wei et al. 2024; Weng et al. 2023; Xu et al. 2024a,b; Zhang et al. 2024] also achieve 3D reconstruction from highly sparse views. Though effective in generating images fast, these methods encounter issues with large pretraining, strict requirements on view distribution and object location, and difficulty in handling real-world captures.

While 3DGS shows strong power in novel view synthesis, it struggles with sparse 360° views similar to NeRF. Inspired by few-shot NeRFs, methods [Charatan et al. 2024; Chung et al. 2023; Paliwal et al. 2024; Xiong et al. 2023; Zhu et al. 2024] have been developed for sparse 360° reconstruction. However, they still severely rely on the SfM points. Our GaussianObject proposes structure-prior-aided Gaussian initialization to tackle this issue, drastically reducing the required input views to only 4, a significant improvement compared with over 20 views required by FSGS [Zhu et al. 2024].

3 METHOD

The subsequent sections detail the methodology: Sec. 3.1 reviews foundational techniques; Sec. 3.2 introduces our overall framework; Sec. 3.3 describes how we apply the structure priors for initial optimization; Sec. 3.4 details the setup of our Gaussian repair model; Sec. 3.5 illustrates the repair of 3D Gaussians using this model and Sec. 3.6 elucidates the COLMAP-free version of GaussianObject. To facilitate a better understanding, all key mathematical symbols and their corresponding meanings are listed in Table 1.

Table 1. List of Key Mathematical Symbols

Symbol	Meaning
$X^{\text{ref}} = \{x_i\}_{i=1}^{N}$	Reference images
$K^{\text{ref}} = \{k_i\}_{i=1}^{N}$	Intrinsics of X^{ref}
$\hat{K}^{\text{ref}}$	Estimated intrinsics of X^{ref}
$\hat{K}$	Estimated shared intrinsics of X^{ref}
$\Pi^{\text{ref}} = \{\pi_i\}_{i=1}^{N}$	Extrinsics of X^{ref}
Π^{nov}	Extrinsics of viewpoints in repair path
$\hat{\Pi}^{\text{ref}}$	Estimated extrinsics of X^{ref}
$M^{\text{ref}} = \{m_i\}_{i=1}^{N}$	Masks of X^{ref}
μ	Center location of Gaussian
q	Rotation quaternion of Gaussian
s	Scale vector of Gaussian
σ	Opacity of Gaussian
sh	Spherical harmonic coefficients of Gaussian
$\mathcal{G}_c$	Coarse 3D Gaussians
$\mathcal{R}$	Diffusion based Gaussian repair model
$\mathcal{E}$	Latent diffusion encoder of $\mathcal{R}$
$\mathcal{D}$	Latent diffusion decoder of $\mathcal{R}$
x'	Degraded rendering
$\hat{x}$	Image repaired by $\mathcal{R}$
ϵ_s	3D Noise added to attributes of $\mathcal{G}_c$
ϵ	2D Gaussian noise for fine-tuning
ϵ_θ	2D Noise predicted by $\mathcal{R}$
c^{tex}	Object-specific language prompt
$\mathcal{P}$	Coarse point cloud predicted by DUSt3R

3.1 Preliminary

3D Gaussian Splatting. 3D Gaussian Splatting [Kerbl et al. 2023] represents 3D scenes with 3D Gaussians. Each 3D Gaussian is composed of the center location μ, rotation quaternion q, scaling vector s, opacity σ, and spherical harmonic (SH) coefficients sh. Thus, a scene is parameterized as a set of Gaussians $\mathcal{G} = \{G_i : \mu_i, q_i, s_i, \sigma_i, sh_i\}_{i=1}^{P}$.

ControlNet. Diffusion models are generative models that sample from a data distribution $q(X_0)$, beginning with Gaussian noise ϵ and using various sampling schedulers. They operate by reversing a discrete-time stochastic noise addition process $\{X_t\}_{t=0}^{T}$ with a diffusion model $p_\theta(X_{t-1}|X_t)$ trained to approximate $q(X_{t-1}|X_t)$, where $t \in [0, T]$ is the noise level and θ is the learnable parameters. Substituting X_0 with its latent code Z_0 from a Variational Autoencoder (VAE) [Kingma and Welling 2014] leads to the development of Latent Diffusion Models (LDM) [Rombach et al. 2022]. ControlNet [Zhang et al. 2023a] further enhances the generative process with additional image conditioning by integrating a network structure similar to the diffusion model, optimized with the loss function:

$$\mathcal{L}_{Cond} = \mathbb{E}_{Z_0,t,\epsilon}[\|\epsilon_\theta(\sqrt{\bar{\alpha}_t}Z_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t, c^{\text{tex}}, c^{\text{img}}) - \epsilon\|_2^2], \quad (1)$$

where c^{tex} and c^{img} denote the text and image conditioning respectively, and ϵ_θ is the Gaussian noise inferred by the diffusion model with parameter θ, $\bar{\alpha}_{1:T} \in (0, 1]^T$ is a decreasing sequence associated with the noise-adding process.

3.2 Overall Framework

Given a sparse collection of N reference images $X^{\text{ref}} = \{x_i\}_{i=1}^{N}$, captured within a 360° range and encompassing one object, along with the corresponding camera intrinsics[2] $K^{\text{ref}} = \{k_i\}_{i=1}^{N}$, extrinsics $\Pi^{\text{ref}} = \{\pi_i\}_{i=1}^{N}$ and masks $M^{\text{ref}} = \{m_i\}_{i=1}^{N}$ of the object, our target is to obtain a 3D representation $\mathcal{G}$, which can achieve photo-realistic rendering $x = \mathcal{G}(\pi|\{x_i, \pi_i, m_i\}_{i=1}^{N})$ from any viewpoint. To achieve this, we employ the 3DGS model for its simplicity for structure priors embedding and fast rendering capabilities. The process begins with initializing 3D Gaussians using a visual hull [Laurentini 1994], followed by optimization with floater elimination, enhancing the structure of Gaussians. Then we design self-generating strategies to supply sufficient image pairs for constructing a Gaussian repair model, which is used to rectify incomplete object information. The overall framework is shown in Fig. 2.

3.3 Initial Optimization with Structure Priors

Sparse views, especially for only 4 images, provide limited 3D information for reconstruction. In this case, SfM points, which are the key for 3DGS initialization, are often absent. Besides, insufficient multi-view consistency leads to ambiguity among shape and appearance, resulting in many floaters during reconstruction. We propose two techniques to initially optimize the 3D Gaussian representation, which take full advantage of structure priors from the limited views and result in a satisfactory outline of the object.

Initialization with Visual Hull. To better leverage object structure information from limited reference images, we utilize the view frustums and object masks to create a visual hull as a geometric scaffold for initializing our 3D Gaussians. Compared with the limited number of SfM points in extremely sparse settings, the visual hull provides more structure priors that help build multiview consistency by excluding unreasonable Gaussian distributions. The cost of the visual hull is just several masks derived from sparse 360° images, which can be easily acquired using current segmentation models such as SAM [Kirillov et al. 2023]. Specifically, points are randomly initialized within the visual hull using rejection sampling: we project uniformly sampled random 3D points onto image planes and retain those within the intersection of all image-space masks. Point colors are averaged from bilinearly interpolated pixel colors across reference image projections. Then we transform these 3D points into 3D Gaussians. For each point, we assign its position as μ and convert its color into sh. The mean distance between adjacent points forms the scale s, while the rotation q is set to a unit quaternion as default. The opacity σ is initialized to a constant value. This initialization strategy relies on the initial masks. Despite potential inaccuracies in these masks or unrepresented concavities by the visual hull, we observed that subsequent optimization processes reliably yield high-quality reconstructions.

Floater Elimination. While the visual hull builds a coarse estimation of the object geometry, it often contains regions that do not belong to the object due to the inadequate coverage of reference images. These regions usually appear to be floaters, damaging the quality of novel view synthesis. These floaters are problematic as the optimization process struggles to adjust them due to insufficient observational data regarding their position and appearance.

To mitigate this issue, we utilize the statistical distribution of distances among the 3D Gaussians to distinguish the primary object and the floaters. This is implemented by the K-Nearest Neighbors (KNN) algorithm, which calculates the average distance to the nearest $\sqrt{P}$ Gaussians for each element in $\mathcal{G}_c$. We then establish a normative range by computing the mean and standard deviation of these distances. Based on statistical analysis, we exclude Gaussians whose mean neighbor distances exceed the adaptive threshold $\tau = \text{mean} + \lambda_e \text{std}$. This thresholding process is repeated periodically throughout optimization, where λ_e is linearly decreased to 0 to refine the scene representation progressively.

Initial Optimization The optimization of $\mathcal{G}_c$ incorporates color, mask, and monocular depth losses. The color loss combines L1 and D-SSIM losses from 3D Gaussian Splatting:

$$\mathcal{L}_1 = \|x - x^{\text{ref}}\|_1, \quad \mathcal{L}_{\text{D-SSIM}} = 1 - \text{SSIM}(x, x^{\text{ref}}), \quad (2)$$

where x is the rendering and x^{ref} is the corresponding reference image. A binary cross entropy (BCE) loss [Jadon 2020] is applied as mask loss:

$$\mathcal{L}_{\text{m}} = -(m^{\text{ref}} \log m + (1 - m^{\text{ref}}) \log(1 - m)), \quad (3)$$

where m denotes the object mask. A shift and scale invariant depth loss is utilized to guide geometry:

$$\mathcal{L}_{\text{d}} = \|D^* - D^*_{\text{pred}}\|_1, \quad (4)$$

where D^* and D^*_{pred} are per-frame rendered depths and monocularly estimated depths [Bhat et al. 2023] respectively. The depth values are computed following a normalization strategy [Ranftl et al. 2020]:

$$D^* = \frac{D - \text{median}(D)}{\frac{1}{M}\sum_{i=1}^{M}|D - \text{median}(D)|}, \quad (5)$$

where M denotes the number of valid pixels. The overall loss combines these components:

$$\mathcal{L}_{\text{ref}} = (1 - \lambda_{\text{SSIM}})\mathcal{L}_1 + \lambda_{\text{SSIM}}\mathcal{L}_{\text{D-SSIM}} + \lambda_{\text{m}}\mathcal{L}_{\text{m}} + \lambda_{\text{d}}\mathcal{L}_{\text{d}}, \quad (6)$$

[2]Given that the camera intrinsics are known and fixed, we exclude them from the rendering function for simplicity.

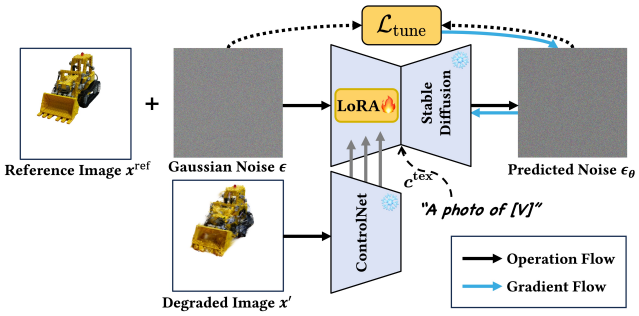

Fig. 3. Illustration of Gaussian repair model setup. First, we add Gaussian noise ϵ to a reference image x^{ref} to form a noisy image. Next, this noisy image along with x^{ref}'s corresponding degraded image x' are passed to a pre-trained fixed ControlNet with learnable LoRA layers to predict a noise distribution ϵ_θ. We use the differences among ϵ and ϵ_θ to fine-tune the parameters in LoRA layers.

where λ_{SSIM}, λ_{m}, and λ_{d} control the magnitude of each term. Thanks to the efficient initialization, our training speed is remarkably fast. It only takes 1 minute to train a coarse Gaussian representation $\mathcal{G}_c$ at a resolution of 779×520.

3.4 Gaussian Repair Model Setup

Combining visual hull initialization and floater elimination significantly enhances 3DGS performance for NVS in sparse 360° contexts. While the fidelity of our reconstruction is generally passable, $\mathcal{G}_c$ still suffers in regions that are poorly observed, regions with occlusion, or even unobserved regions. These challenges loom over the completeness of the reconstruction, like the sword of Damocles.

To mitigate these issues, we introduce a Gaussian repair model $\mathcal{R}$ designed to correct the aberrant distribution of $\mathcal{G}_c$. Our $\mathcal{R}$ takes corrupted rendered images $x'(\mathcal{G}_c, \pi^{\text{nov}})$ as input and outputs photo-realistic and high-fidelity images $\hat{x}$. This image repair capability can be used to refine the 3D Gaussians, leading to learning better structure and appearance details.

Sufficient data pairs are essential for training $\mathcal{R}$ but are rare in existing datasets. To this end, we adopt two main strategies for generating adequate image pairs, *i.e.*, **leave-one-out training** and **adding 3D noises**. For leave-one-out training, we build N subsets from the N input images, each containing $N-1$ reference images and 1 left-out image x^{out}. Then we train N 3DGS models with reference images of these subsets, termed as $\{\mathcal{G}_c^i\}_{i=0}^{N-1}$. After specific iterations, we use the left-out image x^{out} to continue training each Gaussian model $\{\mathcal{G}_c^i\}_{i=0}^{N-1}$ into $\{\hat{\mathcal{G}}_c^i\}_{i=0}^{N-1}$. Throughout this process, the rendered images from the left-out view at different iterations are stored to form the image pairs along with left-out image x^{out} for training the repair model. Note that training these left-out models costs little, with less than N minutes in total. The other strategy is to add 3D noises ϵ_s onto Gaussian attributes. The ϵ_s are derived from the mean μ_Δ and variance σ_Δ of attribute differences between $\{\mathcal{G}_c^i\}_{i=0}^{N-1}$ and $\{\hat{\mathcal{G}}_c^i\}_{i=0}^{N-1}$. This allows us to render more degraded images $x'(\mathcal{G}_c(\epsilon_s), \pi^{\text{ref}})$ at all reference views from the created noisy Gaussians, resulting in extensive image pairs (X', X^{ref}).

We inject LoRA weights and fine-tune a pre-trained ControlNet [Zhang et al. 2023b] using the generated image pairs as our Gaussian repair model. The training procedure is shown in Fig. 3. The loss function, based on Eq. 1, is defined as:

$$\mathcal{L}_{\text{tune}} = \mathbb{E}_{x^{\text{ref}},t,\epsilon,x'}\left[\|(\epsilon_\theta(x_t^{\text{ref}}, t, x', c^{\text{tex}}) - \epsilon)\|_2^2\right], \tag{7}$$

where c^{tex} denotes an object-specific language prompt, defined as "a photo of [V]," as per Dreambooth [Ruiz et al. 2023]. Specifically, we inject LoRA layers into the text encoder, image condition branch, and U-Net for fine-tuning. Please refer to the Appendix for details.

3.5 Gaussian Repair with Distance-Aware Sampling

After training $\mathcal{R}$, we distill its target object priors into $\mathcal{G}_c$ to refine its rendering quality. The object information near the reference views is abundant. This observation motivates designing distance as a criterion in identifying views that need rectification, leading to distance-aware sampling.

Specifically, we establish an elliptical path aligned with the training views and focus on a central point. Arcs near Π^{ref}, where we assume $\mathcal{G}_c$ renders high-quality images, form the reference path. The other arcs, yielding renderings, need to be rectified and define the repair path, as depicted in Fig. 4. In each iteration, novel viewpoints, $\pi_j \in \Pi^{\text{nov}}$, are randomly sampled among the repair path. For each π_j, we render the corresponding image $x_j(\mathcal{G}_c, \pi_j)$, encode it to be $\mathcal{E}(x_j)$ by the latent diffusion encoder $\mathcal{E}$ and pass $\mathcal{E}(x_j)$ to the image conditioning branch of $\mathcal{R}$. Simultaneously, a cloned $\mathcal{E}(x_j)$ is disturbed into a noisy latent z_t:

$$z_t = \sqrt{\bar{\alpha}_t}\mathcal{E}(x_j) + \sqrt{1-\bar{\alpha}_t}\epsilon, \text{ where } \epsilon \sim \mathcal{N}(0, I), t \in [0, T], \tag{8}$$

which is similar to SDEdit [Meng et al. 2022]. We then generate a sample $\hat{x}_j$ from $\mathcal{R}$ by running DDIM sampling [Song et al. 2021] over $k = \lfloor 50 \cdot \frac{t}{T} \rfloor$ steps and forwarding the diffusion decoder $\mathcal{D}$:

$$\hat{x}_j = \mathcal{D}(\text{DDIM}(z_t, \mathcal{E}(x_j))), \tag{9}$$

where $\mathcal{E}$ and $\mathcal{D}$ are from the VAE model used by the diffusion model. The distances from π_j to Π^{ref} is used to weight the reliability of $\hat{x}_j$, guiding the optimization with a loss function:

$$\mathcal{L}_{\text{rep}} = \mathbb{E}_{\pi_j,t}\left[w(t)\lambda(\pi_j)\left(\|x_j - \hat{x}_j\|_1 + \|x_j - \hat{x}_j\|_2 + L_p(x_j, \hat{x}_j)\right)\right],$$
$$\text{where } \lambda(\pi_j) = \frac{2 \cdot \min_{i=1}^{N}(\|\pi_j - \pi_i\|_2)}{d_{\max}}. \tag{10}$$

Here, L_p denotes the perceptual similarity metric LPIPS [Zhang et al. 2018], $w(t)$ is a noise-level modulated weighting function from DreamFusion [Poole et al. 2023], $\lambda(\pi_j)$ denotes a distance-based weighting function, and $d_{\max}$ is the maximal distance among neighboring reference viewpoints. To ensure coherence between 3D Gaussians and reference images, we continue training $\mathcal{G}_c$ with $\mathcal{L}_{\text{ref}}$ during the whole Gaussian repair procedure.

3.6 COLMAP-Free GaussianObject (CF-GaussianObject)

Current SOTA sparse view reconstruction methods rely on precise camera parameters, including intrinsics and poses, obtained through an SfM pipeline with dense input, limiting their usability in daily applications. This process can be cumbersome and unreliable in sparse-view scenarios where matched features are insufficient for accurate reconstruction.

To overcome this limitation, we introduce an advanced sparse matching model, DUSt3R [Wang et al. 2024a], into GaussianObject

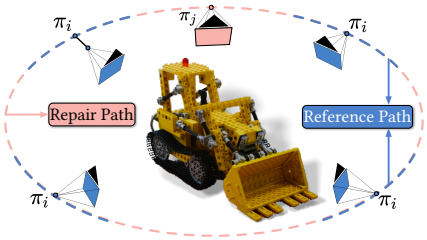

Fig. 4. Illustration of our distance-aware sampling. Blue and red indicate the reference and repair path, respectively.

to enable COLMAP-free sparse 360° reconstruction. Given reference input images X^{ref}, DUSt3R is formulated as:

$$\mathcal{P}, \hat{\Pi}^{\text{ref}}, \hat{K}^{\text{ref}} = \text{DUSt3R}(X^{\text{ref}}), \tag{11}$$

where $\mathcal{P}$ is an estimated coarse point cloud of the scene, and $\hat{\Pi}^{\text{ref}}$, $\hat{K}^{\text{ref}}$ are the predicted camera poses and intrinsics of X^{ref}, respectively. For CF-GaussianObject, we modify the intrinsic recovery module within DUSt3R, allowing $x_i \in X^{\text{ref}}$ to share the same intrinsic $\hat{K}$. This adaption enables the retrieval of $\mathcal{P}$, $\hat{\Pi}^{\text{ref}}$, and $\hat{K}$. Besides, we apply structural priors with a visual hull to $\mathcal{P}$ to initialize 3D Gaussians. After initialization, we optimize $\hat{\Pi}^{\text{ref}}$ and the initialized 3D Gaussians using X^{ref} and depth maps rendered from $\mathcal{P}$ simultaneously. Besides, we introduce a regularization loss to constrain deviations from $\hat{\Pi}^{\text{ref}}$, enhancing the robustness of the optimization. After optimization, the 3D Gaussians and camera parameters are used for constructing the Gaussian repair model and Gaussian repairing process as described in Sec. 3.4 and Sec. 3.5. Refer to the Appendix for more details.

4 EXPERIMENTS

4.1 Implementation Details

Our framework, illustrated in Fig. 2, is based on 3DGS [Kerbl et al. 2023] and threestudio [Guo et al. 2023]. The 3DGS model is trained for 10k iterations in the initial optimization, with periodic floater elimination every 500 iterations. The monocular depth for $\mathcal{L}_{\text{d}}$ is predicted by ZoeDepth [Bhat et al. 2023]. We use a ControlNet-Tile [Zhang et al. 2023b] model based on stable diffusion v1.5 [Rombach et al. 2022] as our repair model's backbone. LoRA [Hu et al. 2022] weights, injected into the text-encoder and transformer blocks using minLoRA [Chang 2023], are trained for 1800 steps at a LoRA rank of 64 and a learning rate of 10^{-3}. $\mathcal{G}_c$ is trained for another 4k iterations during distance-aware sampling. For the first 2800 iterations, optimization involves both a reference image and a repaired novel view image, with the weight of $\mathcal{L}_{\text{rep}}$ progressively decayed from 1.0 to 0.1. The final 1200-step training only involves reference views. The whole process of GaussianObject takes about 30 minutes on a GeForce RTX 3090 GPU for 4 input images at a 779 × 520 resolution. For more details, please refer to the Appendix.

4.2 Datasets

We evaluate GaussianObject on three datasets suited for sparse-view 360° object reconstruction with varying input views, including Mip-NeRF360 [Barron et al. 2021], OmniObject3D [Wu et al. 2023], and OpenIllumination [Liu et al. 2023a]. Additionally, we use an iPhone 13 to capture four views of some daily-life objects to show the COLMAP-free performance. SAM [Kirillov et al. 2023] is used to obtain masks of the target objects.

4.3 Evaluation

Sparse 360° Reconstruction Performance. We evaluate the performance of GaussianObject against several reconstruction baselines, including the vanilla 3DGS [Kerbl et al. 2023] with random initialization and DVGO [Sun et al. 2022], and various few-view reconstruction models on the three datasets. Compared methods of RegNeRF [Niemeyer et al. 2022], DietNeRF [Jain et al. 2021], SparseNeRF [Guangcong et al. 2023], and ZeroRF [Shi et al. 2024b] utilize a variety of regularization techniques. Besides, FSGS [Zhu et al. 2024] is also built upon Gaussian splatting with SfM-point initialization. Note that we supply extra SfM points to FSGS so that it can work with the highly sparse 360° setting. Since camera pose estimation often suffers from scale and positional errors compared to ground truth, we adopt the evaluation used for COLMAP-free methods under dense view settings [Fu et al. 2024; Wang et al. 2021]. All models are trained using publicly released codes.

Table 2 and 3 present the view-synthesis performance of GaussianObject compared to existing methods on the MipNeRF360, OmniObject3D, and OpenIllumination datasets. Experiments show that GaussianObject consistently achieves SOTA results in all datasets, especially in the perceptual quality – LPIPS. Although GaussianObject is designed to address extremely sparse input views, it still outperforms other methods with more input views, *i.e.* 6 and 9, further proving the effectiveness. Notably, GaussianObject excels with as few as 4 views and significantly improves LPIPS over FSGS from 0.0951 to 0.0498 on MipNeRF360. This improvement is critical, as LPIPS is a key indicator of perceptual quality [Park et al. 2021].

Fig. 5 and Fig. 6 illustrate rendering results of various methods across different datasets with only 4 input views. We observe that GaussianObject achieves significantly better visual quality and fidelity than the competing models. We find that implicit representation based methods and random initialized 3DGS fail in extremely sparse settings, typically reconstructing objects as fragmented pixel patches. This confirms the effectiveness of integrating structure priors with explicit representations. Although ZeroRF exhibits competitive PSNR and SSIM on OpenIllumination, its renderings are blurred and lack details, as shown in Fig. 6. In contrast, GaussianObject demonstrates fine-detailed reconstruction. This superior perceptual quality highlights the effectiveness of the Gaussian repair model. It is highly suggested to refer to comprehensive video comparisons included in supplementary materials.

Comparison with LRMs. We further compare GaussianObject to recently popular LRM-like feed-forward reconstruction methods, *i.e.* LGM [Tang et al. 2024a] and TriplaneGaussian (TGS) [Zou et al. 2024] which are publicly available. The comparisons are shown in Table 4 on the challenging MipNeRF360 dataset. Given that TriplaneGaussian accommodates only a single image input, we feed it with frontal views of objects. LGM requires placing the target object at the world coordinate origin with cameras oriented towards it at an elevation of 0° and azimuths of 0°, 90°, 180°, and 270°. Therefore, we report two versions of LGM – LGM-4 which uses

Fig. 5. Qualitative examples on the MipNeRF360 and OmniObject3D dataset with 4 input views. Many methods fail to reach a coherent 3D representation, resulting in floaters and disjoint pixel patches. A pure white image indicates a total miss of the object by the corresponding method, usually caused by overfitting the input images.

Table 2. Comparisons with varying input views. LPIPS* = LPIPS × 10^2 throughout this paper. Best results are highlighted as 1st, 2nd and 3rd.

	Method	4-view			6-view			9-view		
		LPIPS* ↓	PSNR ↑	SSIM ↑	LPIPS* ↓	PSNR ↑	SSIM ↑	LPIPS* ↓	PSNR ↑	SSIM ↑
MipNeRF360	DVGO [Sun et al. 2022]	24.43	14.39	0.7912	26.67	14.30	0.7676	25.66	14.74	0.7842
	3DGS [Kerbl et al. 2023]	10.80	20.31	0.8991	8.38	22.12	0.9134	6.42	24.29	0.9331
	DietNeRF [Jain et al. 2021]	11.17	18.90	0.8971	6.96	22.03	0.9286	5.85	23.55	0.9424
	RegNeRF [Niemeyer et al. 2022]	20.44	13.59	0.8476	20.72	13.41	0.8418	19.70	13.68	0.8517
	FreeNeRF [Yang et al. 2023]	16.83	13.71	0.8534	6.84	22.26	0.9332	5.51	27.66	0.9485
	SparseNeRF [Guangcong et al. 2023]	17.76	12.83	0.8454	19.74	13.42	0.8316	21.56	14.36	0.8235
	ZeroRF [Shi et al. 2024b]	19.88	14.17	0.8188	8.31	24.14	0.9211	5.34	27.78	0.9460
	FSGS [Zhu et al. 2024]	9.51	21.07	0.9097	7.69	22.68	0.9264	6.06	25.31	0.9397
	GaussianObject (Ours)	4.98	24.81	0.9350	3.63	27.00	0.9512	2.75	28.62	0.9638
	CF-GaussianObject (Ours)	8.47	21.39	0.9014	5.71	24.06	0.9269	5.50	24.39	0.9300
OmniObject3D	DVGO [Sun et al. 2022]	14.48	17.14	0.8952	12.89	18.32	0.9142	11.49	19.26	0.9302
	3DGS [Kerbl et al. 2023]	8.60	17.29	0.9299	7.74	18.29	0.9378	6.50	20.26	0.9483
	DietNeRF [Jain et al. 2021]	11.64	18.56	0.9205	10.39	19.07	0.9267	10.32	19.26	0.9258
	RegNeRF [Niemeyer et al. 2022]	16.75	15.20	0.9091	14.38	15.80	0.9207	10.17	17.93	0.9420
	FreeNeRF [Yang et al. 2023]	8.28	17.78	0.9402	7.32	19.02	0.9464	7.25	20.35	0.9467
	SparseNeRF [Guangcong et al. 2023]	17.47	15.22	0.8921	21.71	15.86	0.8935	23.76	17.16	0.8947
	ZeroRF [Shi et al. 2024b]	4.44	27.78	0.9615	3.11	31.94	0.9731	3.10	32.93	0.9747
	FSGS [Zhu et al. 2024]	6.25	24.71	0.9545	6.05	26.36	0.9582	4.17	29.16	0.9695
	GaussianObject (Ours)	2.07	30.89	0.9756	1.55	33.31	0.9821	1.20	35.49	0.9870
	CF-GaussianObject (Ours)	2.62	28.51	0.9669	2.03	30.73	0.9738	2.08	31.23	0.9757

Table 3. Quantitative comparisons on the OpenIllumination dataset. Methods with † means the metrics are from the ZeroRF paper [Shi et al. 2024b].

Method	4-view			6-view		
	LPIPS* ↓	PSNR ↑	SSIM ↑	LPIPS* ↓	PSNR ↑	SSIM ↑
DVGO	11.84	21.15	0.8973	8.83	23.79	0.9209
3DGS	30.08	11.50	0.8454	29.65	11.98	0.8277
DietNeRF†	10.66	23.09	0.9361	9.51	24.20	0.9401
RegNeRF†	47.31	11.61	0.6940	30.28	14.08	0.8586
FreeNeRF†	35.81	12.21	0.7969	35.15	11.47	0.8128
SparseNeRF	22.28	13.60	0.8808	26.30	12.80	0.8403
ZeroRF†	9.74	24.54	0.9308	7.96	26.51	0.9415
Ours	6.71	24.64	0.9354	5.44	26.54	0.9443

Fig. 6. Qualitative results on the OpenIllumination dataset. Although ZeroRF shows competitive PSNR and SSIM, its renderings often appear blurred. While GaussianObject outperforms in restoring fine details, achieving a significant perceptual quality advantage.

four sparse captures as input views directly, and LGM-1 which uses MVDream [Shi et al. 2024a] to generate images that comply with LGM's setup requirements following its original manner. Results show that the strict requirements among input views significantly hinder the sparse reconstruction performance of LRM-like models with in-the-wild captures. In contrast, GaussianObject does not require extensive pre-training, has no restrictions on input views, and can reconstruct any complex object in daily life.

Performance of CF-GaussianObject. CF-GaussianObject is evaluated on the MipNeRF360 and OmniObject3D datasets, with results detailed in Table 2 and Fig. 5. Though CF-GaussianObject exhibits some performance degradation, it eliminates the need for precise camera parameters, significantly enhancing its practical utility. Its performance remains competitive compared to other SOTA methods that depend on accurate camera parameters. Notably, we observe that the performance degradation correlates with an increase in the number of input views, primarily due to declines in the accuracy of DUSt3R's estimates as the number of views rises. As demonstrated in Fig. 7, comparative experiments on smartphone-captured images confirm the superior reconstruction capabilities and visual quality of CF-GaussianObject. More visualization of CF-GaussianObject can be found in our appendix and supplementary materials.

4.4 Ablation Studies

Key Components. We conduct a series of experiments to validate the effectiveness of each component. The following experiments are performed on MipNeRF360 with 4 input views, and averaged metric values are reported. We disable visual hull initialization, floater elimination, Gaussian repair model setup, and Gaussian repair process

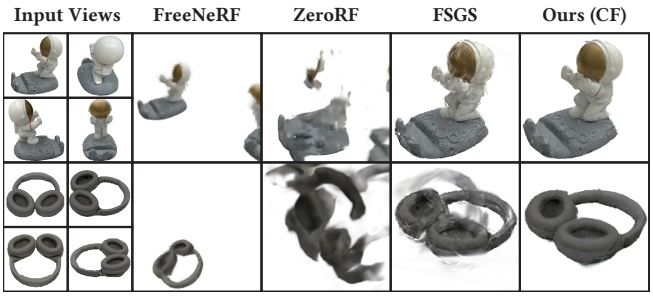

Fig. 7. Qualitative results on our-collected images captured by an iPhone 13. We equip other SOTAs with camera parameters predicted by DUSt3R for fair comparison. The results demonstrate the superior performance of our CF-GaussianObject among casually captured images, with fine details and higher visual quality.

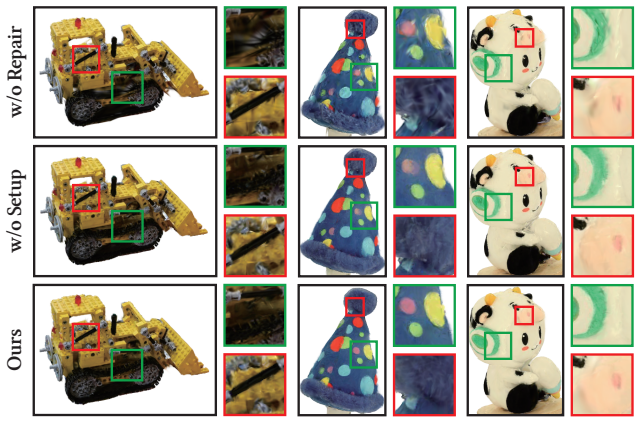

Fig. 8. Importance of our Gaussian repair model setup. Without the Gaussian repair process or the finetuning of the ControlNet, the renderings exhibit noticeable artifacts and lack of details, particularly in areas with insufficient view coverage. Zoom in for better comparison.

once at a time to verify their effectiveness. The Gaussian repair loss is further compared with the Score Distillation Sampling (SDS) loss [Poole et al. 2023], and the depth loss is ablated. The results, presented in Table 5 and Fig. 9, indicate that each element significantly contributes to performance, with their absence leading to a decline in results. Particularly, omitting visual hull initialization results in a marked decrease in performance. Gaussian repair model setup and the Gaussian repair process significantly enhance visual quality, and the absence of either results in a substantial decline in perceptual quality as shown in Fig. 8. Contrary to its effectiveness in text-to-3D or single image-to-3D, SDS results in unstable optimization and diminished performance in our context. The depth loss shows marginal promotion, mainly for LPIPS and SSIM. We apply it to enhance the robustness of our framework.

Structure of Repair Model. Our repair model is designed to generate photo-realistic and 3D-consistent views of the target object. This is achieved by leave-one-out training and perturbing the attributes of 3D Gaussians to create image pairs for fine-tuning a pre-trained

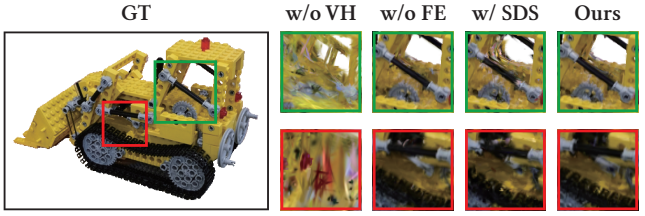

Fig. 9. Ablation study on different components. "VH" denotes for visual hull and "FE" is floater elimination. The "GT" image is from a test view.

Table 4. Quantitative comparisons with LRM-like methods on MipNeRF360.

Method	LPIPS* ↓	PSNR ↑	SSIM ↑
TGS [Zou et al. 2024]	9.14	18.07	0.9073
LGM-4 [Tang et al. 2024a]	9.20	17.97	0.9071
LGM-1 [Tang et al. 2024a]	9.13	17.46	0.9071
GaussianObject (Ours)	**4.99**	**24.81**	**0.9350**

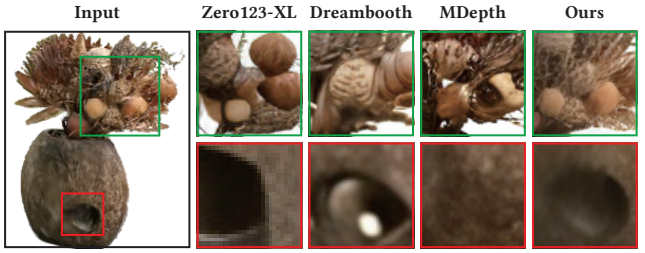

Fig. 10. Qualitative comparisons by ablating different Gaussian repair model setup methods. "MDepth" denotes the repair model with masked monocular depth estimation as the condition.

Table 5. Ablation study on key components.

Method	LPIPS* ↓	PSNR ↑	SSIM ↑
Ours w/o Visual Hull	12.72	15.95	0.8719
Ours w/o Floater Elimination	4.99	24.73	0.9346
Ours w/o Setup	5.53	24.28	0.9307
Ours w/o Gaussian Repair	5.55	24.37	0.9297
Ours w/o Depth Loss	5.09	**24.84**	0.9341
Ours w/ SDS [Poole et al. 2023]	6.07	22.42	0.9188
GaussianObject (Ours)	**4.98**	24.81	**0.9350**

image-conditioned ControlNet. Similarities can be found in Dreambooth [Ruiz et al. 2023], which aims to generate specific subject images from limited inputs. To validate the efficacy of our design, we evaluate the samples generated by our Gaussian repair model and other alternative structures. The first is implemented with Dreambooth [Raj et al. 2023; Ruiz et al. 2023], which embeds target object priors with semantic modifications. To make the output corresponding to the target object, we utilize SDEdit [Meng et al. 2022] to guide the image generation. Inspired by Song et al. [2023b], the second introduces a monocular depth conditioning ControlNet [Zhang et al. 2023a], which is fine-tuned using data pair generation as in Sec. 3.4. We also assess the performance using masked depth conditioning.

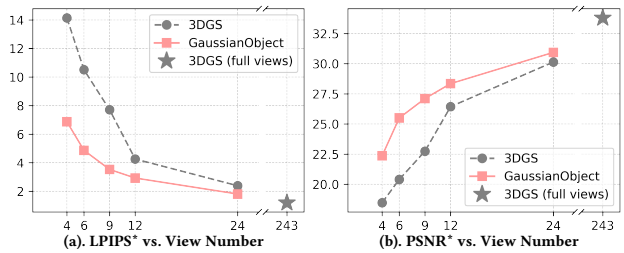

Fig. 11. Ablation on Training View Number. Experiments are conducted on scene *kitchen* in the MipNeRF360 dataset.

Table 6. Ablation study about alternatives of the Gaussian repair model.

Method	LPIPS* ↓	PSNR ↑	SSIM ↑
Zero123-XL [Liu et al. 2023c]	13.97	17.71	0.8921
Dreambooth [Ruiz et al. 2023]	6.58	21.85	0.9093
Depth Condition	7.00	21.87	0.9112
Depth Condition w/ Mask	6.87	21.92	0.9117
GaussianObject (Ours)	**5.79**	**23.55**	**0.9220**

Furthermore, we consider Zero123-XL [Deitke et al. 2023; Liu et al. 2023c], a well-known single-image reconstruction model requiring object-centered input images with precise camera rotations and positions. Here, we manually align the coordinate system and select the closest image to the novel viewpoint as its reference.

The results, as shown in Table 6 and Fig. 10, reveal that semantic modifications proposed by Dreambooth alone fail in 3D-coherent synthesis. Monocular depth conditioning, whether with or without masks, despite some improvements, still struggles with depth roughness and artifacts. Zero123-XL, while generating visually acceptable images, the multi-view structure consistency is lacking. In contrast, our model excels in both 3D consistency and detail fidelity, outperforming others qualitatively and quantitatively.

Effect of View Numbers. We design experiments to evaluate the advantage of our method over different training views. As shown in Fig. 11, GaussianObject consistently outperforms vanilla 3DGS in varying numbers of training views. Besides, GaussianObject with 24 training views achieves performance comparable to that of 3DGS trained on all views (243).

4.5 Limitations and Future Work

GaussianObject demonstrates notable performance in sparse 360° object reconstruction, yet several avenues for future research exist. In regions completely unobserved or insufficiently observed, our repair model may generate hallucinations, *i.e.*, it may produce non-existent details, as shown in Fig. 12. However, these regions are inherently non-deterministic in information, and other methods also struggle in these areas. Additionally, due to the high sparsity level, our model is currently limited in capturing view-dependent effects. With such sparse data, our method cannot differentiate whether the appearance is view-dependent or inherent. Consequently, it 'bakes in' the view-dependent features (like reflected white light) onto the surface, resulting in an inability to display view-dependent appearance from novel viewpoints correctly and leading to some unintended artifacts as demonstrated in Fig. 13. Fine-tuning diffusion models with more

Fig. 12. Hallucinations of non-existent details. GaussianObject may fabricate visually reasonable details in areas with little information. For instance, the hole in the stone vase is filled in.

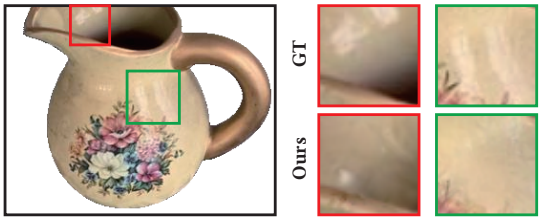

Fig. 13. Comparative visualization highlighting the challenge of reconstructing view-dependent appearance with only four input images.

view-dependent data may be a promising direction. Besides, integrating GaussianObject with surface reconstruction methods like 2DGS [Huang et al. 2024] and GOF [Yu et al. 2024] is a promising direction. Furthermore, CF-GaussianObject achieves competitive performance, but there is still a performance gap compared to precise camera parameters. An interesting exploration is to leverage confidence maps from matching methods for more accurate pose estimation.

5 CONCLUSION

In summary, GaussianObject is a novel framework designed for high-quality 3D object reconstruction from extremely sparse 360° views, based on 3DGS with real-time rendering capabilities. We design two main methods to achieve this goal: structure-prior-aided optimization for facilitating the multi-view consistency construction and a Gaussian repair model to remove artifacts caused by omitted or highly compressed object information. We also provide a COLMAP-free version that can be easily applied in real life with competitive performance. We sincerely hope that GaussianObject can advance daily-life applications of reconstructing 3D objects, markedly reducing capture requirements and broadening application prospects.

ACKNOWLEDGMENTS

This work was supported by the NSFC under Grant 62322604 and 62176159, and in part by the Shanghai Municipal Science and Technology Major Project under Grant 2021SHZDZX0102. The authors express gratitude to the anonymous reviewers for their valuable feedback and to Deyu Wang for his assistance with figure drawing and Blender support.

REFERENCES

Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman. 2021. Mip-NeRF 360: Unbounded Anti-Aliased Neural Radiance Fields. *2022 IEEE/CVF*

Conference on Computer Vision and Pattern Recognition (CVPR) (2021), 5460–5469.
Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. 2023. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288* (2023).
James Burgess, Kuan-Chieh Wang, and Serena Yeung. 2024. Viewpoint Textual Inversion: Unleashing Novel View Synthesis with Pretrained 2D Diffusion Models. *ECCV* (2024).
Eric R Chan, Koki Nagano, Matthew A Chan, Alexander W Bergman, Jeong Joon Park, Axel Levy, Miika Aittala, Shalini De Mello, Tero Karras, and Gordon Wetzstein. 2023. GeNVS: Generative novel view synthesis with 3D-aware diffusion models.
Jonathan Chang. 2023. minLoRA. https://github.com/cccntu/minLoRA.
David Charatan, Sizhe Lester Li, Andrea Tagliasacchi, and Vincent Sitzmann. 2024. pixelsplat: 3d gaussian splats from image pairs for scalable generalizable 3d reconstruction. In *CVPR*. 19457–19467.
Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. 2023. Fantasia3D: Disentangling Geometry and Appearance for High-quality Text-to-3D Content Creation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 22246–22256.
Jaeyoung Chung, Jeongtaek Oh, and Kyoung Mu Lee. 2023. Depth-regularized optimization for 3d gaussian splatting in few-shot images. *arXiv preprint arXiv:2311.13398* (2023).
Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, Eli VanderBilt, Aniruddha Kembhavi, Carl Vondrick, Georgia Gkioxari, Kiana Ehsani, Ludwig Schmidt, and Ali Farhadi. 2023. Objaverse-XL: A Universe of 10M+ 3D Objects. *arXiv preprint arXiv:2307.05663* (2023).
Kangle Deng, Andrew Liu, Jun-Yan Zhu, and Deva Ramanan. 2022. Depth-supervised NeRF: Fewer Views and Faster Training for Free. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 12872–12881. https://doi.org/10.1109/CVPR52688.2022.01254
Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*. https://openreview.net/forum?id=YicbFdNTTy
Yang Fu, Sifei Liu, Amey Kulkarni, Jan Kautz, Alexei A Efros, and Xiaolong Wang. 2024. Colmap-free 3d gaussian splatting. *CVPR* (2024).
Ruiqi Gao, Aleksander Holynski, Philipp Henzler, Arthur Brussee, Ricardo Martin-Brualla, Pratul Srinivasan, Jonathan T Barron, and Ben Poole. 2024. CAT3D: Create Anything in 3D with Multi-View Diffusion Models. *arXiv preprint arXiv:2405.10314* (2024).
Guangcong, Zhaoxi Chen, Chen Change Loy, and Ziwei Liu. 2023. SparseNeRF: Distilling Depth Ranking for Few-shot Novel View Synthesis. *IEEE/CVF International Conference on Computer Vision (ICCV)* (2023).
Yuan-Chen Guo, Ying-Tian Liu, Ruizhi Shao, Christian Laforte, Vikram Voleti, Guan Luo, Chia-Hao Chen, Zi-Xin Zou, Chen Wang, Yan-Pei Cao, and Song-Hai Zhang. 2023. threestudio: A unified framework for 3D content generation. https://github.com/threestudio-project/threestudio.
Ayaan Haque, Matthew Tancik, Alexei Efros, Aleksander Holynski, and Angjoo Kanazawa. 2023. Instruct-NeRF2NeRF: Editing 3D Scenes with Instructions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.
Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. 2024. Lrm: Large reconstruction model for single image to 3d. *ICLR* (2024).
Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. *ICLR* (2022).
Binbin Huang, Zehao Yu, Anpei Chen, Andreas Geiger, and Shenghua Gao. 2024. 2D Gaussian Splatting for Geometrically Accurate Radiance Fields. In *SIGGRAPH 2024 Conference Papers*. Association for Computing Machinery. https://doi.org/10.1145/3641519.3657428
Shruti Jadon. 2020. A survey of loss functions for semantic segmentation. In *2020 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*. 1–7. https://doi.org/10.1109/CIBCB48159.2020.9277638
Ajay Jain, Matthew Tancik, and Pieter Abbeel. 2021. Putting NeRF on a Diet: Semantically Consistent Few-Shot View Synthesis. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. 5865–5874. https://doi.org/10.1109/ICCV48922.2021.00583
Wonbong Jang and Lourdes Agapito. 2024. NViST: In the Wild New View Synthesis from a Single Image with Transformers. *CVPR* (2024).
Hanwen Jiang, Zhenyu Jiang, Yue Zhao, and Qixing Huang. 2024. LEAP: Liberate Sparse-view 3D Modeling from Camera Poses. *ICLR* (2024).
Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 2023. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Transactions on Graphics* 42, 4 (2023).
Mijeong Kim, Seonguk Seo, and Bohyung Han. 2022. Infonerf: Ray entropy minimization for few-shot neural volume rendering. In *CVPR*. 12912–12921.
Diederik P. Kingma and Max Welling. 2014. Auto-Encoding Variational Bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, Yoshua Bengio and Yann LeCun (Eds.). http://arxiv.org/abs/1312.6114
Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. 2023. Segment anything. *ICCV* (2023).
A. Laurentini. 1994. The visual hull concept for silhouette-based image understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 16, 2 (1994), 150–162. https://doi.org/10.1109/34.273735
Jiahao Li, Hao Tan, Kai Zhang, Zexiang Xu, Fujun Luan, Yinghao Xu, Yicong Hong, Kalyan Sunkavalli, Greg Shakhnarovich, and Sai Bi. 2024. Instant3d: Fast text-to-3d with sparse-view generation and large reconstruction model. *ICLR* (2024).
Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. 2023. Magic3D: High-Resolution Text-to-3D Content Creation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
Isabella Liu, Linghao Chen, Ziyang Fu, Liwen Wu, Haian Jin, Zhong Li, Chin Ming Ryan Wong, Yi Xu, Ravi Ramamoorthi, Zexiang Xu, and Hao Su. 2023a. OpenIllumination: A Multi-Illumination Dataset for Inverse Rendering Evaluation on Real Objects. *NeuRIPS 2023*.
Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. 2023c. Zero-1-to-3: Zero-shot One Image to 3D Object. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 9298–9309.
Xinhang Liu, Shiu-hong Kao, Jiaben Chen, Yu-Wing Tai, and Chi-Keung Tang. 2023b. Deceptive-NeRF: Enhancing NeRF Reconstruction using Pseudo-Observations from Diffusion Models. *arXiv preprint arXiv:2305.15171* (2023).
Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. 2022. Sdedit: Guided image synthesis and editing with stochastic differential equations. *ICLR* (2022).
Gal Metzer, Elad Richardson, Or Patashnik, Raja Giryes, and Daniel Cohen-Or. 2023. Latent-NeRF for Shape-Guided Generation of 3D Shapes and Textures. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 12663–12673. https://doi.org/10.1109/CVPR52729.2023.01218
Norman Müller, Katja Schwarz, Barbara Rössle, Lorenzo Porzi, Samuel Rota Bulò, Matthias Nießner, and Peter Kontschieder. 2024. MultiDiff: Consistent Novel View Synthesis from a Single Image. In *CVPR*. 10258–10268.
Michael Niemeyer, Jonathan T. Barron, Ben Mildenhall, Mehdi S. M. Sajjadi, Andreas Geiger, and Noha Radwan. 2022. RegNeRF: Regularizing Neural Radiance Fields for View Synthesis from Sparse Inputs. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 5470–5480. https://doi.org/10.1109/CVPR52688.2022.00540
Avinash Paliwal, Wei Ye, Jinhui Xiong, Dmytro Kotovenko, Rakesh Ranjan, Vikas Chandra, and Nima Khademi Kalantari. 2024. CoherentGS: Sparse Novel View Synthesis with Coherent 3D Gaussians. *ECCV* (2024).
Zijie Pan, Zeyu Yang, Xiatian Zhu, and Li Zhang. 2024. Fast Dynamic 3D Object Generation from a Single-view Video. *arXiv preprint arXiv 2401.08742* (2024).
Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T. Barron, Sofien Bouaziz, Dan B. Goldman, Ricardo Martin-Brualla, and Steven M. Seitz. 2021. HyperNeRF: a higher-dimensional representation for topologically varying neural radiance fields. *ACM Trans. Graph.* 40, 6 (2021), 238:1–238:12.
Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. 2023. Dreamfusion: Text-to-3d using 2d diffusion. *ICLR* (2023).
Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning*. PMLR, 8748–8763. https://proceedings.mlr.press/v139/radford21a.html
Amit Raj, Srinivas Kaza, Ben Poole, Michael Niemeyer, Nataniel Ruiz, Ben Mildenhall, Shiran Zada, Kfir Aberman, Michael Rubinstein, Jonathan Barron, et al. 2023. Dreambooth3d: Subject-driven text-to-3d generation. *ICCV* (2023).
René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. 2021. Vision Transformers for Dense Prediction. *ICCV* (2021).
René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. 2020. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence* 44, 3 (2020), 1623–1637.
René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. 2022. Towards Robust Monocular Depth Estimation: Mixing Datasets for Zero-Shot Cross-Dataset Transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 3 (2022).
Barbara Roessle, Jonathan T Barron, Ben Mildenhall, Pratul P Srinivasan, and Matthias Nießner. 2022. Dense depth priors for neural radiance fields from sparse input views. In *CVPR*. 12892–12901.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 10674–10685. https://doi.org/10.1109/CVPR52688.2022.01042

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 2023. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *CVPR*. 22500–22510.

Johannes Lutz Schönberger and Jan-Michael Frahm. 2016. Structure-from-Motion Revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Seunghyeon Seo, Yeonjin Chang, and Nojun Kwak. 2023. Flipnerf: Flipped reflection rays for few-shot novel view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 22883–22893.

Ruizhi Shao, Jingxiang Sun, Cheng Peng, Zerong Zheng, Boyao Zhou, Hongwen Zhang, and Yebin Liu. 2024. Control4D: Efficient 4D Portrait Editing with Text. *CVPR* (2024).

Ruoxi Shi, Xinyue Wei, Cheng Wang, and Hao Su. 2024b. ZeroRF: Fast Sparse View 360° Reconstruction with Zero Pretraining. *CVPR* (2024).

Yichun Shi, Peng Wang, Jianglong Ye, Long Mai, Kejie Li, and Xiao Yang. 2024a. MVDream: Multi-view Diffusion for 3D Generation. *ICLR* (2024).

Nagabhushan Somraj, Adithyan Karanayil, Sai Harsha Mupparaju, and Rajiv Soundararajan. 2024. Simple-RF: Regularizing Sparse Input Radiance Fields with Simpler Solutions. *arXiv preprint arXiv:2404.19015* (2024).

Nagabhushan Somraj, Adithyan Karanayil, and Rajiv Soundararajan. 2023. SimpleNeRF: Regularizing Sparse Input Neural Radiance Fields with Simpler Solutions. In *SIGGRAPH Asia 2023 Conference Papers (SA '23)*. Association for Computing Machinery, New York, NY, USA, 1–11. https://doi.org/10.1145/3610548.3618188

Nagabhushan Somraj and Rajiv Soundararajan. 2023. ViP-NeRF: Visibility Prior for Sparse Input Neural Radiance Fields. (August 2023). https://doi.org/10.1145/3588432.3591539

Jiaming Song, Chenlin Meng, and Stefano Ermon. 2021. Denoising diffusion implicit models. *ICLR* (2021).

Jiuhn Song, Seonghoon Park, Honggyu An, Seokju Cho, Min-Seop Kwak, Sungjin Cho, and Seungryong Kim. 2023b. DaRF: Boosting Radiance Fields from Sparse Inputs with Monocular Depth Adaptation. *2023 NIPS* (2023).

Liangchen Song, Zhong Li, Xuan Gong, Lele Chen, Zhang Chen, Yi Xu, and Junsong Yuan. 2023a. Harnessing Low-Frequency Neural Fields for Few-Shot View Synthesis. *arXiv preprint arXiv:2303.08370* (2023).

Cheng Sun, Min Sun, and Hwann-Tzong Chen. 2022. Direct Voxel Grid Optimization: Super-fast Convergence for Radiance Fields Reconstruction. In *CVPR*.

Jiaxiang Tang, Zhaoxi Chen, Xiaokang Chen, Tengfei Wang, Gang Zeng, and Ziwei Liu. 2024a. LGM: Large Multi-View Gaussian Model for High-Resolution 3D Content Creation. *ECCV* (2024).

Jiaxiang Tang, Jiawei Ren, Hang Zhou, Ziwei Liu, and Gang Zeng. 2024b. DreamGaussian: Generative Gaussian Splatting for Efficient 3D Content Creation. *ICLR* (2024).

Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A. Yeh, and Greg Shakhnarovich. 2023a. Score Jacobian Chaining: Lifting Pretrained 2D Diffusion Models for 3D Generation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 12619–12629. https://doi.org/10.1109/CVPR52729.2023.01214

Peng Wang, Hao Tan, Sai Bi, Yinghao Xu, Fujun Luan, Kalyan Sunkavalli, Wenping Wang, Zexiang Xu, and Kai Zhang. 2024b. PF-LRM: Pose-Free Large Reconstruction Model for Joint Pose and Shape Prediction. *ICLR* (2024).

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. 2024a. DUSt3R: Geometric 3D Vision Made Easy. *CVPR* (2024).

Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. 2023b. ProlificDreamer: High-Fidelity and Diverse Text-to-3D Generation with Variational Score Distillation. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Zirui Wang, Shangzhe Wu, Weidi Xie, Min Chen, and Victor Adrian Prisacariu. 2021. NeRF–: Neural radiance fields without known camera parameters. *arXiv preprint arXiv:2102.07064* (2021).

Xinyue Wei, Kai Zhang, Sai Bi, Hao Tan, Fujun Luan, Valentin Deschaintre, Kalyan Sunkavalli, Hao Su, and Zexiang Xu. 2024. MeshLRM: Large Reconstruction Model for High-Quality Mesh. *arXiv preprint arXiv:2404.12385* (2024).

Zhenzhen Weng, Jingyuan Liu, Hao Tan, Zhan Xu, Yang Zhou, Serena Yeung-Levy, and Jimei Yang. 2023. Template-Free Single-View 3D Human Digitalization with Diffusion-Guided LRM. *Preprint* (2023).

Rundi Wu, Ben Mildenhall, Philipp Henzler, Keunhong Park, Ruiqi Gao, Daniel Watson, Pratul P Srinivasan, Dor Verbin, Jonathan T Barron, Ben Poole, et al. 2024. ReconFusion: 3D Reconstruction with Diffusion Priors. *CVPR* (2024).

Tong Wu, Jiarui Zhang, Xiao Fu, Yuxin Wang, Jiawei Ren, Liang Pan, Wayne Wu, Lei Yang, Jiaqi Wang, Chen Qian, Dahua Lin, and Ziwei Liu. 2023. OmniObject3D: Large-Vocabulary 3D Object Dataset for Realistic Perception, Reconstruction and Generation. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2023), 803–814. https://api.semanticscholar.org/CorpusID:255998491

Jamie Wynn and Daniyar Turmukhambetov. 2023. DiffusioNeRF: Regularizing Neural Radiance Fields with Denoising Diffusion Models. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 4180–4189. https://doi.org/10.1109/CVPR52729.2023.00407

Haolin Xiong, Sairisheek Muttukuru, Rishi Upadhyay, Pradyumna Chari, and Achuta Kadambi. 2023. SparseGS: Real-Time 360° Sparse View Synthesis using Gaussian Splatting. *Arxiv* (2023).

Dejia Xu, Yifan Jiang, Peihao Wang, Zhiwen Fan, Humphrey Shi, and Zhangyang Wang. 2022. SinNeRF: Training Neural Radiance Fields on Complex Scenes from a Single Image. In *Computer Vision – ECCV 2022 (Lecture Notes in Computer Science)*, Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (Eds.). Springer Nature Switzerland, Cham, 736–753. https://doi.org/10.1007/978-3-031-20047-2_42

Dejia Xu, Ye Yuan, Morteza Mardani, Sifei Liu, Jiaming Song, Zhangyang Wang, and Arash Vahdat. 2024c. AGG: Amortized Generative 3D Gaussians for Single Image to 3D. *arXiv preprint 2401.04099* (2024).

Yinghao Xu, Zifan Shi, Wang Yifan, Hansheng Chen, Ceyuan Yang, Sida Peng, Yujun Shen, and Gordon Wetzstein. 2024a. Grm: Large gaussian reconstruction model for efficient 3d reconstruction and generation. *ECCV* (2024).

Yinghao Xu, Hao Tan, Fujun Luan, Sai Bi, Peng Wang, Jiahao Li, Zifan Shi, Kalyan Sunkavalli, Gordon Wetzstein, Zexiang Xu, and Kai Zhang. 2024b. DMV3D: Denoising Multi-View Diffusion using 3D Large Reconstruction Model. *ICLR* (2024).

Jiawei Yang, Marco Pavone, and Yue Wang. 2023. FreeNeRF: Improving Few-Shot Neural Rendering with Free Frequency Regularization. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 8254–8263. https://doi.org/10.1109/CVPR52729.2023.00798

Taoran Yi, Jiemin Fang, Junjie Wang, Guanjun Wu, Lingxi Xie, Xiaopeng Zhang, Wenyu Liu, Qi Tian, and Xinggang Wang. 2024. GaussianDreamer: Fast Generation from Text to 3D Gaussians by Bridging 2D and 3D Diffusion Models. *CVPR* (2024).

Zehao Yu, Torsten Sattler, and Andreas Geiger. 2024. Gaussian Opacity Fields: Efficient High-quality Compact Surface Reconstruction in Unbounded Scenes. *arXiv:2404.10772* (2024).

Kai Zhang, Sai Bi, Hao Tan, Yuanbo Xiangli, Nanxuan Zhao, Kalyan Sunkavalli, and Zexiang Xu. 2024. GS-LRM: Large Reconstruction Model for 3D Gaussian Splatting. *arXiv* (2024).

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023a. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3836–3847.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023b. ControlNet-v1-1-nightly. https://github.com/lllyasviel/ControlNet-v1-1-nightly.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 586–595.

Zhizhuo Zhou and Shubham Tulsiani. 2023. SparseFusion: Distilling View-Conditioned Diffusion for 3D Reconstruction. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 12588–12597. https://doi.org/10.1109/CVPR52729.2023.01211

Junzhe Zhu and Peiye Zhuang. 2024. HiFA: High-fidelity Text-to-3D Generation with Advanced Diffusion Guidance. *ICLR* (2024).

Zehao Zhu, Zhiwen Fan, Yifan Jiang, and Zhangyang Wang. 2024. FSGS: Real-Time Few-shot View Synthesis using Gaussian Splatting. *ECCV* (2024).

Zi-Xin Zou, Zhipeng Yu, Yuan-Chen Guo, Yangguang Li, Ding Liang, Yan-Pei Cao, and Song-Hai Zhang. 2024. Triplane Meets Gaussian Splatting: Fast and Generalizable Single-View 3D Reconstruction with Transformers. *CVPR* (2024).