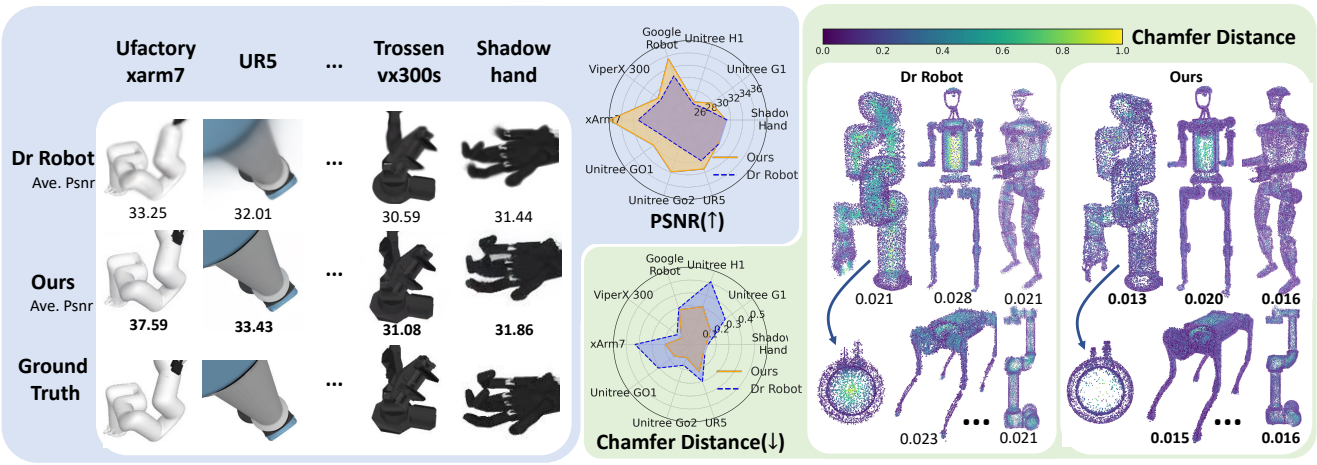
GRADRobot: Geometry-Aware Rendering with Articulation and Diffusion for Robot Modeling

Yunlong Li[1] Boyuan Chen[1] Chongjie Ye[2,7] Bohan Li[3,8]
Zhaoxi Chen[4] Shaocong Xu[5] Hao Tang[6] Hao Zhao[1,5]

[1]AIR, THU [2]CUHKSZ [3]SJTU [4]NTU [5]BAAI [6]PKU [7]FNii [8]EIT



GRADRobot: Sharper Rendering and Tighter Geometry Across Diverse Robots 1) Left. Rendered views of representative robots. Compared to Dr Robot (top row), GRADRobot (middle row) reconstructs finer surface details and achieves higher PSNR across instances. 2) Centre. Radar-plot comparisons on multiple robot models. GRADRobot (orange) yields higher PSNR (↑) and lower Chamfer Distance (↓) than Dr Robot (blue), indicating better rendering fidelity and geometry. 3) Right. Error-coloured point clouds visualize reconstruction accuracy (purple = low error, yellow = high). GRADRobot (right column) reduces reconstruction errors and Chamfer Distance across all models; see slice-based comparisons later in the paper.

Abstract

Gaussian fields are a promising representation for robot body modeling due to their differentiability and inherently low sim-to-real gap. However, existing methods like Dr-Robot overlook explicit geometric constraints, leading to artifacts under novel poses or views. Directly enforcing depth and normal supervision on articulated Gaussians is unstable due to entanglement between pose deformation and 3D appearance learning. To address this, we propose a two-stage training strategy: we first learn a canonical Gaussian field in a canonical pose using dense RGB, depth, and normal supervision, establishing a geometry-aware reconstruction. We then fine-tune the Gaussian parameters jointly with a deformation network conditioned on joint angles using only RGB losses, ensuring consistent geometry and appearance across poses. To further mitigate rendering artifacts in novel poses and viewpoints, we integrate a diffusion-based refinement module. This module conditions on both the initial Gaussian renderings and the target robot skeletons, and significantly enhances visual fidelity while preserving pose accuracy. Experiments across multiple robotic platforms show that GRADRobot outperforms DrRobot by a large margin in both rendering quality (PSNR) and geometric accuracy (Chamfer Distance).

https://github.com/liyunlooong/GRADRobot

1. Introduction

Accurate and differentiable rendering of *articulated robots* is a core enabler for simulation, perception, and control. An effective representation must (i) preserve fine surface geometry, (ii) synthesize view-dependent appearance, and (iii) remain stable under large pose changes. Gaussian splatting provides a continuous, optimizable scene parameterization with efficient rasterization, but—when trained only with RGB losses—it tends to oversmooth high-frequency details and yields pose-dependent artifacts, as observed in prior work such as DrRobot [30].

We introduce **GRADRobot**, a geometry-aware articulated Gaussian renderer that decouples *geometry reconstruction* from *articulation learning*, and refines appearance with structure-conditioned diffusion. Concretely, GRADRobot proceeds in three steps: (1) **Canonical reconstruction** learns surface-anchored Gaussians in a reference pose by supervising depth and surface normals, which stabilizes geometry and mitigates view/pose leakage; (2) **Pose-conditioned deformation** generalizes to novel joint configurations using Linear Blend Skinning (LBS) that is optimized on top of the canonical geometry, avoiding entanglement between geometry and motion; (3) **Diffusion refinement** applies a ControlNet-conditioned image prior [58] driven by the robot skeleton to recover thin structures and specular details while respecting kinematics.

Compare with DrRobot. GRADRobot adopts the articulated-Gaussian paradigm but differs in two key aspects: (i) it uses *explicit* geometric supervision (depth/normal) in a canonical space to anchor Gaussians to the physical surface; and (ii) it adds a *structure-conditioned* diffusion stage for high-frequency restoration, which improves RGB fidelity without modifying the underlying kinematic model.

Scope and practical considerations. Our evaluation focuses on controlled settings with known kinematics and camera parameters. We report rendering accuracy (e.g., PSNR) and geometric fidelity (e.g., Chamfer Distance), and analyze runtime of each stage, including diffusion refinement. We also discuss temporal consistency (single-frame refinement may introduce flicker) and provide ablations to quantify the contributions of geometry losses and refinement. Extending to real-robot imagery with sensor noise and complex materials is an important next step and is discussed as a limitation.

Our main contributions are:

- We propose a *decoupled* learning framework that first reconstructs canonical, surface-anchored Gaussians with depth/normal supervision, then learns pose-conditioned appearance via LBS for robust generalization to novel configurations.
- We introduce a *structure-conditioned diffusion* module that refines Gaussian renderings using skeleton guidance, recovering high-frequency details while preserving kinematic coherence.
- We conduct extensive experiments across diverse robot morphologies, with ablations on geometry losses and refinement. GRADRobot outperforms DrRobot by up to **37%** in Chamfer Distance and **1.2 dB** in PSNR under identical protocols.

Together, these design choices yield a renderer that pairs physically grounded geometry with high-fidelity appearance, offering a practical building block for scalable robotic digital twins.

2. Related Work

Differentiable rendering with Gaussian primitives has driven rapid progress in 3-D reconstruction and robotics [1, 4, 8, 14, 19, 21, 24, 26, 28, 30, 47, 50, 51, 61]. The seminal 3-D Gaussian Splatting (GS) by Kerbl [19] showed that thousands of anisotropic Gaussians, refined by interleaved density control, can render radiance fields in real time while preserving geometry. Subsequent work improves efficiency and scalability by training once and deploying across multiple scenes [28], handling ego-centric sparse views [47], extremely sparse and unposed 360-degree imagery [1, 50], online unposed image streams [24], and wide-coverage large scenes with reconstruction-oriented large models and self-distillation [4, 8]. Beyond first-order primitives, Quadratic Gaussian Splatting [61] introduces second-order geometric primitives for high-quality surface reconstruction, while StochasticSplats [21] proposes stochastic rasterisation for sorting-free, anti-aliased rendering.

Building on this foundation, GPS-Gaussian [62] extends GS to human novel-view synthesis, producing 2K-resolution images from sparse cameras without per-subject optimisation and demonstrating cross-identity generalisation in casual capture setups. Follow-up work explores more structured human and avatar representations: MeGA combines meshes and Gaussians for high-fidelity, editable head avatars [44]; Human Gaussian Model learns efficient and generalisable human representations [33]; and tetrahedron-constrained Gaussian splatting further regularises editable photorealistic avatars [27]. For geometry–aware editing, SuGaR [14] anchors splats to an underlying mesh for coherent large deformations, Mani-GS manipulates Gaussian fields via triangular meshes [12], BG-Triangle introduces Bézier Gaussian triangles for 3D vectorisation and rendering [49], and GauUpdate enables inserting new objects into 3D Gaussian fields under consistent global illumination [37]. Gaussian Opacity Fields [54] fur-

ther improves surface fidelity by surface-aware optimisation of Gaussian points.

In robotics, Differentiable Robot Rendering (DrRobot) [30] couples Gaussian splats with kinematics-aware deformation so that pixel-space gradients can be back-propagated to joint angles for pose recovery and vision–language alignment, enabling analysis-by-synthesis pipelines that close the loop between perception and control. Complementary work studies Gaussian-based scene representations for navigation, SLAM, and driving: DeSiRe-GS proposes 4D Street Gaussians with static–dynamic decomposition for urban driving scenes [36]; DeGauss performs dynamic–static decomposition for distractor-free 3D reconstruction [45]; and Street Gaussians without 3D Object Tracker removes the need for explicit object tracking in street-scale reconstructions [59]. MAGiC-SLAM [56] leverages Gaussians for globally consistent multi-agent SLAM, while EmbodiedSplat [5] uses Gaussian splats reconstructed from mobile devices for personalised real-to-sim-to-real navigation. EventSplat extends GS to moving event cameras for real-time rendering [57], indicating that Gaussian fields can also accommodate alternative sensing modalities.

For inverse rendering, GS-IR [26] adds depth-regularised normals and baked occlusion handling for photorealistic relighting and material estimation, while GaussianShader [18] aligns analytic shading with Gaussian normals to efficiently handle reflective surfaces and supports differentiable BRDF manipulation. RNG [9] further introduces relightable neural Gaussians, IRGS models inter-reflections via 2D Gaussian ray tracing [13], and Luminance-GS adapts GS to challenging lighting conditions through view-adaptive curve adjustment [6]. Ref-GS focuses on directional factorisation for 2D Gaussian splatting [60]. Other extensions combat aliasing or improve efficiency: Multi-Scale GS [51] adaptively enlarges Gaussians in low-resolution views, SA-GS introduces training-free scale-adaptive filtering for anti-aliasing in GS [41], and Rip-NeRF realises anti-aliasing via ripmap-encoded Platonic solids [29]. SlimmeRF enables test-time slimmability for radiance fields to trade accuracy for speed/memory on demand [55], and SUNDAE spectrally prunes Gaussian fields with a neural compensation head to reduce memory cost with minimal quality loss [52].

Beyond reconstruction, Gaussian fields are increasingly used as a representation for high-level tasks and perception. POp-GS formulates next-best-view planning directly on 3D Gaussian splats via P-optimality [48], and NeRF Is a Valuable Assistant for 3D Gaussian Splatting exploits NeRF as a complementary prior for GS optimisation [10]. SplatTalk performs 3D visual question answering on Gaussian scenes [42], Trace3D lifts 2D segmentations to consistent 3D instances via Gaussian instance tracing [39], and LUDVIG uplifts 2D visual features to Gaussian scenes in a learning-free manner [35], highlighting the role of Gaussians as a unifying 3D scaffold for downstream tasks. Complementary perception advances that benefit manipulation and reconstruction include joint semantic–affordance–attribute parsing with Cerberus Transformer [3], shape-aware zero-shot semantic segmentation [32], and unsupervised 3D keypoint discovery via SNAKE [63] and the 3D Implicit Transporter [64]; these modules, together with Gaussian-based VQA, segmentation lifting, and feature uplifting [35, 39, 42], offer supervision signals and structure priors that can be readily integrated into Gaussian-field training.

Despite this progress, explicit geometric constraints are often absent, yielding artefacts under unseen poses or viewpoints. Directly supervising articulated Gaussians with depth and normals is also unstable because pose deformation and appearance learning become entangled, leading to gradient interference and suboptimal convergence. We therefore adopt a two-stage strategy: first, learn a canonical Gaussian field in a fixed pose with dense RGB, depth, and normal supervision for geometry-aware reconstruction; second, fine-tune this field together with a pose-conditioned deformation network using only RGB losses, thereby decoupling appearance from articulation and improving generalisation to novel poses.

2.1. Generative Diffusion Models

Generative Diffusion Models, rooted in nonequilibrium thermodynamics [17, 40], achieve state-of-the-art results in image synthesis [38, 46], 3D reconstruction [15, 46], and image restoration [11, 34, 53]. Discriminator guidance refines score estimates to improve ImageNet sampling quality [22]. DiffPIR embeds diffusion priors for plug-and-play restoration with few NFEs [65], while the Reconstruct-and-Generate (RnG) framework balances fidelity and perceptual quality through reconstructive denoising followed by diffusion-based detail synthesis [46]. For 3D-aware tasks, MagicMan leverages multi-view diffusion with SMPL-X priors and iterative refinement [15]; Diff-Retinex merges Retinex decomposition with diffusion for low-light enhancement [53]; and the Generative Diffusion Prior (GDP) unifies linear and non-linear inverse problems via unsupervised conditional guidance [11]. In our setting, we integrate a lightweight diffusion-based refinement module conditioned on initial Gaussian renderings and robot skeletons: it serves as a learned image prior that suppresses view-dependent artefacts while preserving pose accuracy and maintaining training efficiency.

Recent work at 3DV 2024 learns monocular 3D object localisation using physical laws of motion from only 2D labels, achieving a mean distance error of 6,cm in real-world experiments [7]. Likewise, Li [23] enhances 3D object de-

tection by aligning multi-modal features, improving accuracy in challenging scenarios; together, these advances are complementary to our geometry-aware pipeline and highlight the value of physical priors and cross-modal alignment in downstream perception.

3. Method

3.1. Preliminaries and Notation

This method builds upon 3D Gaussian Splatting (3DGS) [20]. A scene is represented by a set of M anisotropic Gaussians:

$$\mathcal{G} = \{(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, \alpha_i, \mathbf{c}_i)\}_{i=1}^{M} \tag{1}$$

where each Gaussian has a center $\boldsymbol{\mu}_i \in \mathbb{R}^3$, an SPD (Symmetric Positive Definite) covariance matrix $\boldsymbol{\Sigma}_i \in \mathbb{R}^{3\times3}$, opacity $\alpha_i \in [0,1]$, and color coefficients $\mathbf{c}_i$ (spherical harmonics with degree annealed during training). To ensure the covariance is SPD and facilitate unconstrained updates, we parametrize the scale and rotation as follows:

$$\boldsymbol{\Sigma}_i = R_i \,\mathrm{diag}(e^{\mathbf{s}_i})\, R_i^\top, \quad R_i = \exp([\boldsymbol{\omega}_i]_\times), \tag{2}$$

where $\mathbf{s}_i = (s_x, s_y, s_z)$ are the log-scales, and $\boldsymbol{\omega}_i \in \mathbb{R}^3$ represents an axis-angle, with $[\cdot]_\times$ denoting the skew-symmetric matrix. The scale is clamped to the range $[\log s_{\min}, \log s_{\max}]$ to prevent vanishing or excessively blurred splats.

Differentiable Splatting. For a ray $r(t) = \mathbf{o} + t\mathbf{d}$ and camera projection Π, the projected 2D ellipse's Jacobian $J_i = \partial\Pi/\partial\mathbf{x}|_{\boldsymbol{\mu}_i}$ is used to calculate the screen-space covariance:

$$\boldsymbol{\Sigma}_i^{\mathrm{2D}} = J_i \boldsymbol{\Sigma}_i J_i^\top. \tag{3}$$

The visibility-aware compositing computes per-ray opacities $\alpha_i(r)$ and weights as:

$$w_i(r) = T_{i-1}(r)\alpha_i(r), \quad T_{i-1}(r) = \prod_{j<i} (1 - \alpha_j(r)), \tag{4}$$

where $C(r) = \sum_i w_i(r)\mathbf{c}_i$. The projected depth for Gaussian i along ray r is:

$$z_i(r) = \mathbf{d}^\top(\boldsymbol{\mu}_i - \mathbf{o}), \quad \mathbf{x}_i(r) = \mathbf{o} + z_i(r)\mathbf{d}. \tag{5}$$

The ellipsoid normal at $\mathbf{x}_i(r)$ is:

$$\mathbf{n}_i(r) \propto \boldsymbol{\Sigma}_i^{-1}(\mathbf{x}_i(r) - \boldsymbol{\mu}_i), \quad \|\mathbf{n}_i(r)\|_2 = 1. \tag{6}$$

Ground-truth depth $p(r)$ is derived from RGB-D or multi-view fusion, and the supervision for normals $\hat{\mathbf{n}}(r)$ comes from local plane fitting after back-projection.

Decoupling Geometry and Motion. Learning both geometry and motion jointly from RGB(-D) supervision can lead to drift and pose-dependent artifacts due to entangling these two factors. Therefore, we first anchor the Gaussians to the surface in a canonical pose with dense geometry supervision, then introduce a kinematics-aware deformation in a second stage with RGB-only fine-tuning. The final stage uses a diffusion refiner to restore high-frequency details that pure Gaussian fields may miss.

3.2. Approach Overview

The GRADRobot method is organized into three stages, as shown in Fig. 1:

1. **Stage I: Canonical Reconstruction.** In this stage, we learn a surface-anchored Gaussian field in a fixed pose, using RGB, depth, and normal supervision.
2. **Stage II: Pose-Conditioned Deformation and Joint Fine-Tuning.** Here, Gaussians are attached to a kinematic chain using Linear Blend Skinning (LBS), and we jointly fine-tune the field and deformation with RGB-only supervision across multiple poses.
3. **Stage III: Diffusion Refinement.** In the final stage, a ControlNet-augmented latent diffusion model is employed to refine the coarse render, using the robot skeleton as a structural prior. This step recovers high-frequency details, such as thin cables and specularities, while maintaining kinematic consistency.

3.3. Stage I: Canonical Gaussian Reconstruction

Initialization. We initialize M_0 Gaussians using Poisson-disk sampling on a fused canonical point cloud. Each Gaussian seed is initialized at the point, with $R_i = I$, $\mathbf{s}_i \sim \log\mathcal{N}(\log s_0, \sigma^2)$, $\alpha_i \sim \mathcal{U}(0.01, 0.05)$, and spherical harmonics of degree 0. The training alternates between *optimize*, *densify*, and *prune* steps.

Objective. We minimize the following loss function:

$$\mathcal{L}_{\text{can}} = \mathcal{L}_{\text{rgb}} + \lambda_g \mathcal{L}_{\text{geo}}. \tag{7}$$

The photometric loss term, defined on pixels or rays, is:

$$\begin{aligned}\mathcal{L}_{\text{rgb}} = \frac{1}{|\Omega|}\sum_{u\in\Omega} [(1-\lambda_{\text{ssim}})\|C(u) - I_{\text{gt}}(u)\|_1 \\ +\lambda_{\text{ssim}}\text{DSSIM}(C(u), I_{\text{gt}}(u))]\,.\end{aligned} \tag{8}$$

where Ω represents the set of pixels, and DSSIM is the Structural Similarity Index. To anchor the Gaussians to the surface, we align the depths and normals at ray–Gaussian intersections, using compositing weights as confidences while preventing opacity manipulation via the *stopgrad* operation:

$$\mathcal{L}_{\text{dist}} = \sum_{r\in\mathcal{R}}\sum_i \text{stopgrad}(w_i(r)) \cdot |z_i(r) - p(r)|, \tag{9}$$

$$\mathcal{L}_{\text{normal}} = \sum_{r\in\mathcal{R}}\sum_i \text{stopgrad}(w_i(r)) \cdot \|\mathbf{n}_i(r) - \hat{\mathbf{n}}(r)\|_2^2, \tag{10}$$

where $p(r)$ and $\hat{\mathbf{n}}(r)$ are the ground-truth depth and normal, respectively. The geometry loss is:

$$\mathcal{L}_{\text{geo}} = \lambda_d \mathcal{L}_{\text{dist}} + \lambda_n \mathcal{L}_{\text{normal}}. \tag{11}$$

A Huber loss variant is used to handle sensor noise in the depth term.

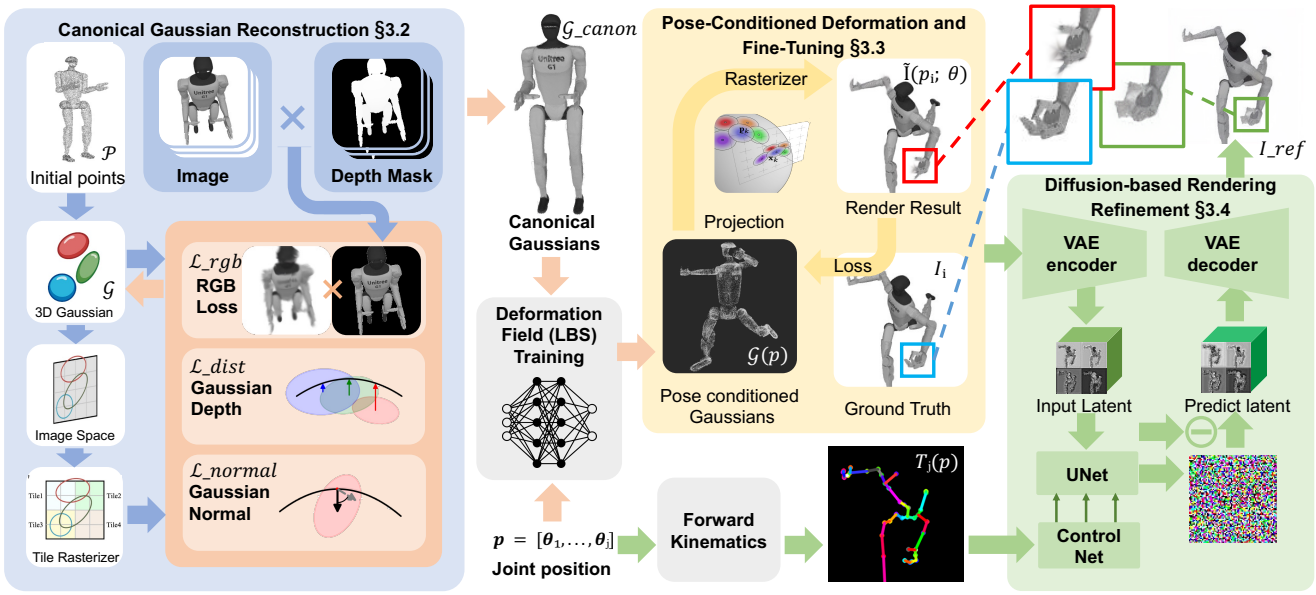


Figure 1. **GRADRobot Overview.** *Stage I* learns a **canonical** surface-anchored Gaussian field using RGB, depth, and normal supervision in a fixed pose. *Stage II* incorporates **pose-conditioned deformation** through Linear Blend Skinning (LBS), fine-tuning both the Gaussians and the deformation using multi-pose RGB supervision. *Stage III* employs a **structure-conditioned diffusion refiner** (ControlNet) guided by the robot skeleton to restore high-frequency details while preserving kinematic constraints.

Densification and Pruning. Every 2k iterations (up to 20k iterations), we split Gaussians that have large 2D radii or exhibit high gradient magnitudes, while pruning those with low opacity or gradient values. The opacity of all Gaussians is reset every 5k iterations to prevent saturation.

Regularization. Mild regularization is applied to the scale and opacity of the Gaussians:

$$R_{\text{scale}} = \sum_i \|\mathbf{s}_i\|_2^2, \quad R_\alpha = \sum_i (\alpha_i - \bar{\alpha})^2. \tag{12}$$

These terms are absorbed into the regularization hyperparameter λ_{reg}.

3.4. Stage II: Pose-Conditioned Deformation and Joint Fine-Tuning

Skinning Weights. Each Gaussian i is assigned skinning weights $\mathbf{w}_i = \{w_{ij}\}_{j=1}^J$, which are predicted by a small MLP function $f_{\text{skin}}(\gamma(\boldsymbol{\mu}_i))$ where $\gamma(\cdot)$ denotes positional encoding. We enforce the constraint $\sum_j w_{ij} = 1$ via softmax and apply an entropy bonus to avoid overly concentrated weight distributions.

Linear Blend Skinning (LBS). The Gaussians are deformed by LBS, which affects both the mean and covariance:

$$\tilde{R}_i(p) = \exp\left(\sum_j w_{ij} \log R_j(p)\right), \qquad \boldsymbol{\Sigma}_i^{(p)} = \tilde{R}_i(p)\, \boldsymbol{\Sigma}_i\, \tilde{R}_i(p)^\top. \tag{13}$$

The mean is transformed using:

$$\boldsymbol{\mu}_i^{(p)} = \sum_j w_{ij} \left(R_j(p)\boldsymbol{\mu}_i + \mathbf{t}_j(p)\right). \tag{14}$$

The condition number of $\boldsymbol{\Sigma}_i^{(p)}$ is clamped by restricting the scales $\mathbf{s}_i$ to a fixed range.

Pose-Dependent Appearance. We use a light adapter f_{app} to model mild pose/view-dependent appearance effects:

$$(\Delta\mathbf{c}_i, \Delta\alpha_i) = f_{\text{app}}\left(\gamma(\boldsymbol{\mu}_i), \gamma(\boldsymbol{\mu}_i^{(p)})\right), \tag{15}$$

and update the color and opacity as:

$$\mathbf{c}_i^{(p)} = \mathbf{c}_i + \Delta\mathbf{c}_i, \quad \alpha_i^{(p)} = \sigma(\alpha_i + \Delta\alpha_i). \tag{16}$$

Deformation Pretraining (Optional). Before fine-tuning with RGB, we optionally pretrain the skinning function f_{skin} using Chamfer distance between deformed canonical samples and observed scans. We avoid directly computing Chamfer distance on Gaussian centers and instead sample uniformly from the ellipsoid iso-surfaces.

Joint RGB Fine-Tuning. In this step, we jointly optimize the parameters $\Theta = \{\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, \alpha_i, \mathbf{c}_i, \phi_{\text{skin}}, \phi_{\text{app}}\}$ via the following objective:

$$\min_\Theta \sum_k \left\| \hat{I}(p_k; \Theta) - I_k \right\|_2^2 + \lambda_{\text{reg}} R(\Theta), \tag{17}$$

where $R(\Theta)$ contains regularization terms for opacity, scale, skinning sum consistency, and ARAP constraints.

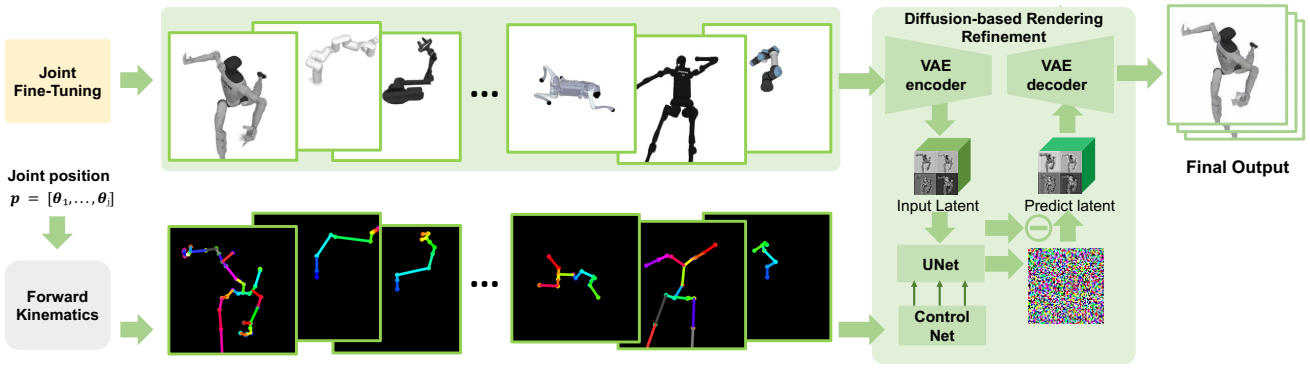


Figure 2. This stage refines the coarse Gaussian renderings generated after joint fine-tuning using a structure-conditioned diffusion module. The pipeline first encodes both the coarse render and the corresponding robot skeleton, obtained through forward kinematics of joint angles, into latent representations using a VAE encoder. The ControlNet module injects structural guidance from the skeleton map into the diffusion process, and the UNet predicts the refined latent that enhances high-frequency details such as metallic reflections, thin links, and cables. The VAE decoder reconstructs the final high-fidelity image, yielding visually sharper and geometrically consistent outputs that respect the robot's kinematic structure.

3.5. Stage III: Diffusion-Based Rendering Refinement

Conditioning Design. In this stage, a coarse RGB image I_{coarse} is rendered from the output of Stage II, and a skeleton map S is generated using Forward Kinematics (FK). These are resized and fed into a ControlNet-augmented Stable Diffusion model to refine the render.

Single-Step Refinement. We use a single step of diffusion to refine the image, with a fixed 256×256 size for the image and skeleton map. The residual $\mathbf{r}$ is predicted by the ControlNet, and the U-Net performs one DDIM step to generate the refined image:

$$\hat{z} = z_{\text{coarse}} - \varepsilon_\theta(z_{\text{coarse}}, 0, c, \mathbf{r}), \quad I_{\text{ref}} = \text{VAE\_dec}(\hat{z}). \quad (18)$$

Loss and Masking. The refinement loss is computed as:

$$\mathcal{L}_{\text{ref}} = \|M \odot (I_{\text{ref}} - I_{\text{gt}})\|_1 + \lambda_p \|\phi(I_{\text{ref}}) - \phi(I_{\text{gt}})\|_2^2, \quad (19)$$

where M is the robot mask to avoid overfitting the background, and ϕ extracts features using a VGG-16 `relu3_3` layer.

3.6. Motivation and Insights

The development of GRADRobot stems from the need to address several key challenges in 3D scene reconstruction and robotic motion capture, particularly for highly dynamic and articulated objects like robots. The primary motivation behind our approach is to leverage the flexibility of Gaussian fields for 3D scene representation while ensuring the preservation of important kinematic and geometric properties in the reconstruction. Several insights guide the design of the GRADRobot method:

1. Geometry and Motion Separation: A common challenge in 3D reconstruction from RGB(-D) supervision is the entanglement of geometry and motion. In traditional methods, geometry and motion are often learned together, leading to pose-dependent artifacts and drift. The key insight here is that by decoupling these two factors, we can improve the accuracy and stability of the reconstruction. Our method anchors the Gaussians to a canonical pose in the first stage, ensuring that the learned geometry is not influenced by pose variations. This separation allows for more precise geometry learning before introducing pose-dependent deformations in the subsequent stages. This decoupling idea has been previously explored in methods such as [2] and [31], where separating geometry and motion is shown to improve reconstruction quality.

2. Gaussian Fields for Continuous Scene Representation: Gaussian splatting has proven to be an effective way to represent 3D scenes with a continuous and smooth surface. The flexibility of Gaussian fields enables us to model complex 3D shapes and deformations, making it particularly suitable for dynamic objects like robots. Additionally, the anisotropic nature of the Gaussians allows for better handling of fine details and complex geometries, such as the intricate wiring and articulation of robot limbs. This continuous representation also facilitates differentiable rendering, allowing for direct optimization from image-based supervision. Gaussian splatting as a continuous representation has been widely used in recent works like [20].

3. High-Frequency Detail Restoration: While Gaussian fields are excellent at capturing the overall structure of an object, they can struggle to preserve high-frequency details such as thin cables, reflections, or other fine surface details. The final stage of GRADRobot employs a diffusion model, augmented by ControlNet, to restore these high-frequency details. This insight leverages the power of generative mod-

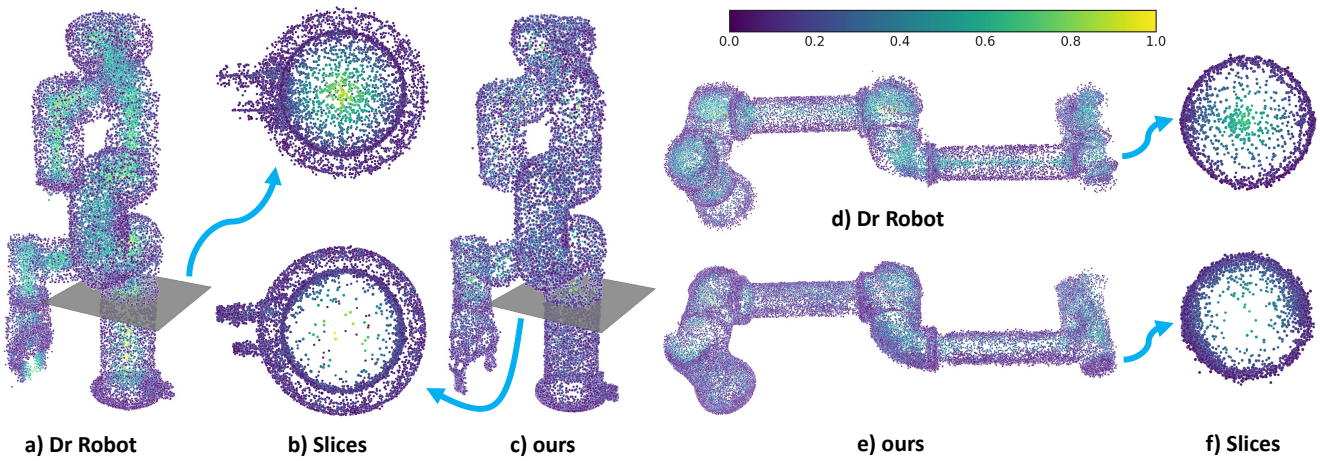


Figure 3. **Chamfer-Distance Visualization** on the *ufactory_xarm7* and *google_robot*. (a,d) DrRobot; (b,e) Ours; (c,f) Planar slice. Our method significantly reduces high-error regions (yellow) and avoids internal artifacts within hollow links.

Table 1. **Mean Chamfer Distance** (lower is better; $\times 10^{-3}$ m) over 50 random test poses per robot.

Method	Shadow Hand	Unitree G1	Unitree H1	Google Robot	ViperX300	xArm7	Unitree GO1	Unitree GO2	UR5	**Average**
DrRobot	0.1372	0.3344	0.5119	0.2956	0.1292	0.4321	0.3103	0.1697	0.3009	0.2913
Ours w/o Geo Loss	0.1273	0.2078	0.3162	0.2937	0.1019	0.2001	0.1885	0.1152	0.2409	0.1991
Ours	**0.1210**	**0.1895**	**0.3083**	**0.2827**	**0.0846**	**0.1936**	**0.1538**	**0.1053**	**0.2205**	**0.1844**

els to refine the output of the earlier stages, ensuring that the reconstructed scene is both geometrically accurate and visually detailed. Recent work in neural rendering and generative modeling, such as [38], has demonstrated the ability of diffusion models to refine and generate realistic textures and fine details.

4. Structural Prior from Skeletons: The use of a robot skeleton as a structural prior in the diffusion refinement stage offers another key insight. By incorporating kinematic constraints into the generative process, we ensure that the refined details are consistent with the robot's physical structure and movement. This approach allows us to capture subtle details that would otherwise be difficult to model, such as the bending of cables or the specular reflections on metallic surfaces, while maintaining realistic motion and structure. Similar ideas have been explored in [25] and [16], where skeleton-based priors help improve the consistency and realism of generated results.

These insights drive the GRADRobot pipeline to achieve both accurate 3D reconstructions and realistic motion modeling, paving the way for more effective robotic design, simulation, and real-time applications.

4. Experiments

4.1. Setup

4.1.1 Robots and Pose Splits

We evaluate our method on a synthetic benchmark generated using MuJoCo [43], which includes nine different robot morphologies: industrial arms, mobile quadrupeds, and dexterous hands. Each robot is standardized in terms of lighting and material properties, with fixed joints removed to ensure controlled comparisons across models.

Pose Splits. To evaluate performance, we divide the poses into three distinct splits:

- **Canonical** poses: 500 collision-free poses are used solely for Stage I supervision and optional deformation pretraining (Sec. 3.4).
- **Training** poses: 10,000 poses are sampled with up to 10 collisions per pose to increase diversity in the training set.
- **Test** poses: 500 held-out poses, sampled in the same manner as the training set, but are not used during optimization.

4.1.2 Cameras and Rendering

Each pose is rendered from 12 distinct RGB-D viewpoints at a resolution of 256×256. Cameras are distributed by sampling the azimuth in three bins with jitter (covering the range $[-180°, 180°]$), elevation from the set $\{-45°, 45°\}$,

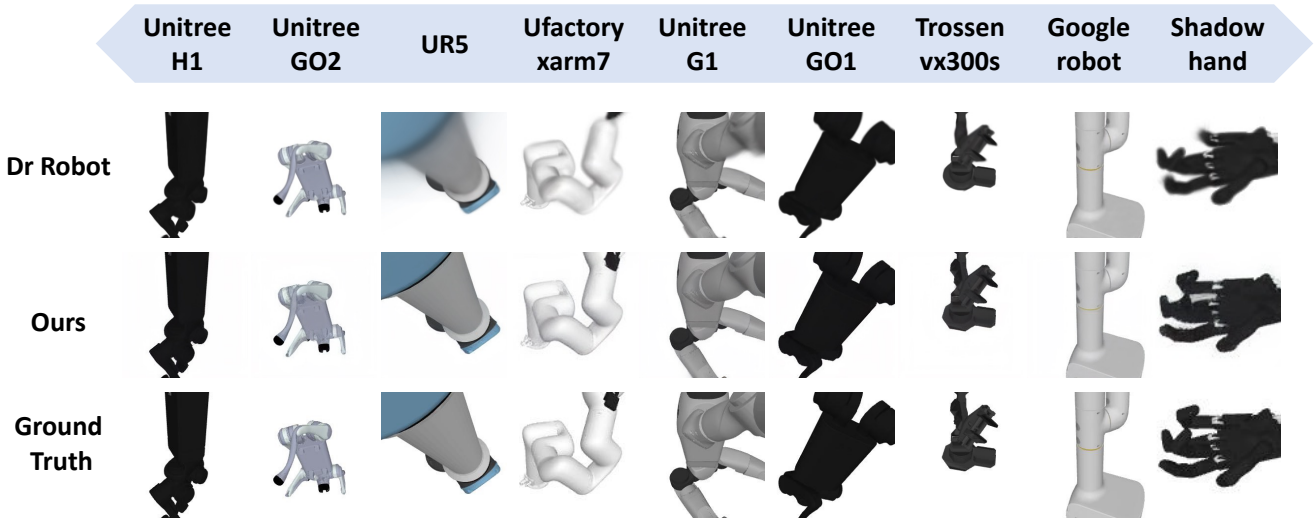


Figure 4. Qualitative comparison across robot morphologies. Columns: Unitree H1, Unitree GO2, UR5, UFactory xArm7, Unitree G1, Unitree GO1, Trossen VX300S, Google Robot, Shadow Hand. Rows: Dr Robot (top), GRADRobot (middle), Ground Truth (bottom). GRADRobot recovers thinner links and sharper boundaries, reduces pose-dependent artifacts (e.g., leakage inside hollow links), and visually aligns with the ground truth under novel poses/views, matching the quantitative gains in higher PSNR and lower Chamfer Distance.

Table 2. **Average PSNR** (dB) on held-out test poses and views.

Method	Shadow Hand	Unitree G1	Unitree H1	Google Robot	ViperX300	xArm7	Unitree GO1	Unitree GO2	UR5	**Average**
K-Plane	9.74	11.23	8.26	13.21	8.45	18.93	11.23	8.41	13.63	11.45
DrRobot w/o Deform	29.75	26.95	26.08	30.89	29.73	31.68	26.95	28.16	26.35	28.50
DrRobot	31.44	28.31	27.77	32.60	30.59	33.25	29.74	29.60	32.01	30.59
Ours w/o Geo Loss	30.85	28.30	25.80	33.11	30.09	35.71	30.12	31.72	32.80	30.94
Ours w/o Refine	31.25	29.49	28.07	34.20	30.84	**37.95**	31.27	31.41	32.92	31.93
Ours	**31.86**	**29.69**	**28.12**	**35.53**	**31.08**	36.51	**32.01**	**33.94**	**33.43**	**32.46**

and radius from the set $\{1.0, 2.0\} \times d$ where d is the scene scaling factor. The RGB-D frames are back-projected using Open3D into world coordinates, followed by voxel-downsampling at a resolution of 0.01 m. These per-view point clouds are fused and adaptively voxelized (starting at 0.005 m and increasing by 1% until the cloud contains no more than 10k points), yielding a single ground-truth fused point cloud per pose.

4.2. Baselines and Evaluation Protocols

4.2.1 Baselines

We compare our method against several baselines:

- **DrRobot**: Our re-implementation of Gaussian splats with kinematics-aware deformation and joint fine-tuning.
- **DrRobot w/o Deform**: A variant using only canonical reconstruction, omitting the deformation stage.
- **K-Plane**: A 2D layered field proxy used as an image-only baseline for comparison.

We also perform ablation studies, such as:

- **Ours w/o Geo Loss**: Removing the depth and normal anchoring in Stage I.
- **Ours w/o Refine**: Disabling the diffusion-based refinement in Stage III.

4.2.2 Training and Evaluation Parity

All methods are trained and evaluated using the same train/test splits, camera setups, and background conditions to ensure a fair comparison. We use fixed random seeds across all dataloaders and report the results on the held-out test poses only.

4.2.3 Metrics

Geometry (Chamfer Distance). We evaluate the symmetric Chamfer Distance (CD) between the predicted model surfaces and the fused ground-truth point clouds. To avoid bias in representation, both our method and DrRobot are evaluated using uniform surface samples. For Gaussians, the points are sampled on ellipsoid iso-surfaces after the Linear Blend Skinning (LBS) transformation, while for meshes, the points are sampled uniformly across triangle areas.

Image Quality. The Peak Signal-to-Noise Ratio (PSNR) is reported for full-frame renders at a resolution of 256×256. Additionally, masked-robot PSNR and Structural Similarity Index (SSIM) scores are included in the supplementary material.

4.3. Geometry Results: Chamfer Distance

Metric. The Chamfer Distance (CD) between a predicted surface $\hat{\mathcal{P}}$ and a ground-truth cloud $\mathcal{P}$ is computed as follows:

$$\begin{aligned}\mathrm{CD}(\hat{\mathcal{P}},\mathcal{P}) = \frac{1}{|\hat{\mathcal{P}}|}\sum_{\mathbf{x}\in\hat{\mathcal{P}}}\min_{\mathbf{y}\in\mathcal{P}}\|\mathbf{x}-\mathbf{y}\|_2^2 \\ + \frac{1}{|\mathcal{P}|}\sum_{\mathbf{y}\in\mathcal{P}}\min_{\mathbf{x}\in\hat{\mathcal{P}}}\|\mathbf{y}-\mathbf{x}\|_2^2.\end{aligned} \tag{20}$$

Protocol. We sample 50 random test poses per robot, apply the learned models via LBS, export approximately 50k surface samples, and report the means over these poses. 95% confidence intervals (CIs) are provided in the supplementary material.

Results. Table 1 shows a 37% average reduction in CD compared to DrRobot. Figure 3 visualizes the per-point errors and a planar slice of the surfaces, illustrating that our method reduces internal artifacts and avoids dense error regions in hollow links, indicating better surface anchoring and deformation coherence.

Ablation: Geometry Anchoring. When we remove the depth/normal anchoring ("Ours w/o Geo Loss"), the CD increases across robots. Re-adding the depth and normal losses results in an additional $\sim$8% reduction in CD on average.

4.4. Image Quality Results: PSNR

Protocol. For each test pose, we render the scene from 12 distinct cameras at a resolution of 256×256 and compute the Peak Signal-to-Noise Ratio (PSNR) relative to the ground-truth RGB. Unless otherwise specified, PSNR is computed on full-frame renders. The masked-robot PSNR and SSIM results are reported in the supplementary material. All methods use identical camera intrinsics and backgrounds.

Results and Analysis. Table 2 shows consistent improvements in PSNR. Our full pipeline yields a PSNR improvement of 0.6–1.2 dB over the pose-conditioned splatting stage and more than 1 dB over DrRobot on average. Qualitative results shown in Figure 4 highlight that our diffusion refinement stage restores thin links and sharp edges, while preserving the intended kinematic pose.

Ablations. Disabling the diffusion stage (*Ours w/o Refine*) results in a decrease in mean PSNR by -0.7 dB, confirming the contribution of the refiner is non-trivial and cannot be replaced by splatting alone. Removing the geometry anchoring also lowers PSNR (from 32.46 to 30.94 dB), which aligns with the increase in CD.

5. Conclusion

GRADRobot fuses surface-anchored 3-D Gaussian splatting with a pose-aware diffusion refiner, producing sharper renders and tighter geometry for articulated robots. Across nine robot models it cuts Chamfer Distance by up to 37 % and boosts PSNR by 0.6–1.2 dB over Dr Robot, all while preserving real-time speed. The study shows geometry losses and diffusion are complementary—one locks Gaussians to the true surface, the other restores high-frequency detail.

References

[1] Chong Bao, Xiyu Zhang, Zehao Yu, Jiale Shi, Guofeng Zhang, Songyou Peng, and Zhaopeng Cui. Free360: Layered gaussian splatting for unbounded 360-degree view synthesis from extremely sparse and unposed views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 2

[2] Ambrish Bhat, Anshul Kadian, and Deva Ramanan. Learning to generate 3d objects with an integrated generative model. *CVPR*, 2020. 6

[3] Xiaoxue Chen, Tianyu Liu, Hao Zhao, Guyue Zhou, and Ya-Qin Zhang. Cerberus transformer: Joint semantic, affordance and attribute parsing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19649–19658, 2022. 3

[4] Ziwen Chen, Hao Tan, Kai Zhang, Sai Bi, Fujun Luan, Yicong Hong, Fuxin Li, and Zexiang Xu. Long-lrm: Long-sequence large reconstruction model for wide-coverage gaussian splats. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025. 2

[5] Gunjan Chhablani, Xiaomeng Ye, Muhammad Zubair Irshad, and Zsolt Kira. Embodiedsplat: Personalized real-to-sim-to-real navigation with gaussian splats from a mobile device. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. 3

[6] Ziteng Cui, Xuangeng Chu, and Tatsuya Harada. Luminance-gs: Adapting 3d gaussian splatting to challenging lighting conditions with view-adaptive curve adjustment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 3

[7] Katja Ludwig Rainer Lienhart Daniel Kienzle, Julian Lorenz. Towards learning monocular 3d object localization using the physical laws of motion. In *2024 International Conference on 3D Vision (3DV)*, 2024. 3

[8] Jixuan Fan, Wanhua Li, Yifei Han, and Yansong Tang. Momentum-gs: Momentum gaussian self-distillation for high-quality large scene reconstruction. In *Proceedings of*

the IEEE/CVF International Conference on Computer Vision (ICCV), 2025. 2

[9] Jiahui Fan, Fujun Luan, Jian Yang, Milos Hasan, and Beibei Wang. Rng: Relightable neural gaussians. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 3

[10] Shuangkang Fang, I-Chao Shen, Takeo Igarashi, Yufeng Wang, ZeSheng Wang, Yi Yang, Wenrui Ding, and Shuchang Zhou. Nerf is a valuable assistant for 3d gaussian splatting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. 3

[11] Ben Fei, Zhaoyang Lyu, Liang Pan, Junzhe Zhang, Weidong Yang, Tianyue Luo, Bo Zhang, and Bo Dai. Generative diffusion prior for unified image restoration and enhancement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9935–9946, 2023. 3

[12] Xiangjun Gao, Xiaoyu Li, Yiyu Zhuang, Qi Zhang, Wenbo Hu, Chaopeng Zhang, Yao Yao, Ying Shan, and Long Quan. Mani-gs: Gaussian splatting manipulation with triangular mesh. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 2

[13] Chun Gu, Xiaofei Wei, Zixuan Zeng, Yuxuan Yao, and Li Zhang. Irgs: Inter-reflective gaussian splatting with 2d gaussian ray tracing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 3

[14] Antoine Guédon and Vincent Lepetit. Sugar: Surface-aligned gaussian splatting for efficient 3d mesh reconstruction and high-quality mesh rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5354–5363, 2024. 2

[15] Xu He, Zhiyong Wu, Xiaoyu Li, Di Kang, Chaopeng Zhang, Jiangnan Ye, Liyang Chen, Xiangjun Gao, Han Zhang, and Haolin Zhuang. Magicman: Generative novel view synthesis of humans with 3d-aware diffusion and iterative refinement. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3437–3445, 2025. 3

[16] Xiaoming Huang, Jin Zhang, and Wei Liu. Skeleton-based action recognition via deep learning. In *IEEE Transactions on Cybernetics*, pages 523–536, 2020. 7

[17] Hao Jiang and Yadong Mu. Conditional diffusion process for inverse halftoning. *NeurIPS*, 35:5498–5509, 2022. 3

[18] Yingwenqi Jiang, Jiadong Tu, Yuan Liu, Xifeng Gao, Xiaoxiao Long, Wenping Wang, and Yuexin Ma. Gaussianshader: 3d gaussian splatting with shading functions for reflective surfaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5322–5332, 2024. 3

[19] 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. 2

[20] Matthias Kerbl and Wojciech Matusik. 3d gaussian splatting for efficient and continuous scene representation. *SIGGRAPH*, 2023. 4, 6

[21] Shakiba Kheradmand, Delio Vicini, George Kopanas, Dmitry Lagun, Kwang Moo Yi, Mark Matthews, and Andrea Tagliasacchi. Stochasticsplats: Stochastic rasterization for sorting-free 3d gaussian splatting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. 2

[22] Dongjun Kim, Yeongmin Kim, Se Jung Kwon, Wanmo Kang, and Il-Chul Moon. Refining generative process with discriminator guidance in score-based diffusion models. *arXiv preprint arXiv:2211.17091*, 2022. 3

[23] Meng Li Yunyang Xiong Raghuraman Krishnamoorthi Qiang Liu Vikas Chandra Lemeng Wu, Dilin Wang. Enhancing 3d detection through feature aligned deep fusion. In *2024 International Conference on 3D Vision (3DV)*, 2024. 3

[24] Yang Li, Jinglu Wang, Lei Chu, Xiao Li, Shiu-Hong Kao, Ying-Cong Chen, and Yan Lu. Streamgs: Online generalizable gaussian splatting reconstruction for unposed image streams. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. 2

[25] Zhe Li, Weiming Yu, Tian Li, Hao Huang, and Tao Xue. Generative modeling of 3d human poses and shapes. In *CVPR*, 2019. 7

[26] Zhihao Liang, Qi Zhang, Ying Feng, Ying Shan, and Kui Jia. Gs-ir: 3d gaussian splatting for inverse rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21644–21653, 2024. 2, 3

[27] Hanxi Liu, Yifang Men, and Zhouhui Lian. Creating your editable 3d photorealistic avatar with tetrahedron-constrained gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 2

[28] Hengyu Liu, Yuehao Wang, Chenxin Li, Ruisi Cai, Kevin Wang, Wuyang Li, Pavlo Molchanov, Peihao Wang, and Zhangyang Wang. Flexgs: Train once, deploy everywhere with many-in-one flexible 3d gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 2

[29] Junchen Liu, Wenbo Hu, Zhuo Yang, Jianteng Chen, Guoliang Wang, Xiaoxue Chen, Yantong Cai, Huan ang Gao, and Hao Zhao. Rip-NeRF: Anti-aliasing radiance fields with ripmap-encoded platonic solids. In *SIGGRAPH '24: ACM SIGGRAPH 2024 Conference Papers*, New York, NY, USA, 2024. Association for Computing Machinery. 3

[30] Ruoshi Liu, Alper Canberk, Shuran Song, and Carl Vondrick. Differentiable robot rendering, 2024. 2, 3

[31] Weiyue Liu, Shuai Li, Yifan Dai, and Hongdong Li. Learning to reconstruct 3d objects with multiple views. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 6

[32] Xinyu Liu, Beiwen Tian, Zhen Wang, Rui Wang, Kehua Sheng, Bo Zhang, Hao Zhao, and Guyue Zhou. Delving into shape-aware zero-shot semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2999–3009, 2023. 3

[33] Yifan Liu, Shengjun Zhang, Chensheng Dai, Yang Chen, Hao Liu, Chen Li, and Yueqi Duan. Learning efficient and generalizable human representation with human gaussian model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. 2

[34] Shitong Luo and Wei Hu. Diffusion probabilistic models for 3d point cloud generation. In *CVPR*, pages 2837–2845, 2021. 3

[35] Juliette Marrie, Romain Menegaux, Michael Arbel, Diane Larlus, and Julien Mairal. Ludvig: Learning-free uplifting of 2d visual features to gaussian splatting scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. 3

[36] Chensheng Peng, Chengwei Zhang, Yixiao Wang, Chenfeng Xu, Yichen Xie, Wenzhao Zheng, Kurt Keutzer, Masayoshi Tomizuka, and Wei Zhan. Desire-gs: 4d street gaussians for static-dynamic decomposition and surface reconstruction for urban driving scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 3

[37] Chengwei Ren, Fan Zhang, Liangchao Xu, Liang Pan, Ziwei Liu, Wenping Wang, Xiao-Ping Zhang, and Yuan Liu. Gaupdate: New object insertion in 3d gaussian fields with consistent global illumination. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. 2

[38] Robin Rombach, Arne Blum, Thomas Wenzel, and et al. High-quality image generation with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 3, 7

[39] Hongyu Shen, Junfeng Ni, Yixin Chen, Weishuo Li, Mingtao Pei, and Siyuan Huang. Trace3d: Consistent segmentation lifting via gaussian instance tracing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. 3

[40] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*. PMLR, 2015. 3

[41] Xiaowei Song, Jv Zheng, Shiran Yuan, Huan ang Gao, Jingwei Zhao, Xiang He, Weihao Gu, and Hao Zhao. SA-GS: Scale-adaptive gaussian splatting for training-free anti-aliasing. 2024. 3

[42] Anh Thai, Songyou Peng, Kyle Genova, Leonidas Guibas, and Thomas Funkhouser. Splattalk: 3d vqa with gaussian splatting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. 3

[43] Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 5026–5033. IEEE, 2012. 7

[44] Cong Wang, Di Kang, Heyi Sun, Shenhan Qian, Zixuan Wang, Linchao Bao, and Song-Hai Zhang. Mega: Hybrid mesh-gaussian head avatar for high-fidelity rendering and head editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 2

[45] Rui Wang, Quentin Lohmeyer, Mirko Meboldt, and Siyu Tang. Degauss: Dynamic-static decomposition with gaussian splatting for distractor-free 3d reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. 3

[46] Yujin Wang, Lingen Li, Tianfan Xue, and Jinwei Gu. Reconstruct-and-generate diffusion model for detail-preserving image denoising. *arXiv preprint arXiv:2309.10714*, 2023. 3

[47] Dongxu Wei, Zhiqi Li, and Peidong Liu. Omni-scene: Omni-gaussian representation for ego-centric sparse-view scene reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 2

[48] Joey Wilson, Marcelino Almeida, Sachit Mahajan, Martin Labrie, Maani Ghaffari, Omid Ghasemalizadeh, Min Sun, Cheng-Hao Kuo, and Arnab Sen. Pop-gs: Next best view in 3d-gaussian splatting with p-optimality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 3

[49] Minye Wu, Haizhao Dai, Kaixin Yao, Tinne Tuytelaars, and Jingyi Yu. Bg-triangle: Bézier gaussian triangle for 3d vectorization and rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 2

[50] Jiale Xu, Shenghua Gao, and Ying Shan. Freesplatter: Pose-free gaussian splatting for sparse-view 3d reconstruction. *arXiv preprint arXiv:2412.09573*, 2024. 2

[51] Zhiwen Yan, Weng Fei Low, Yu Chen, and Gim Hee Lee. Multi-scale 3d gaussian splatting for anti-aliased rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20923–20931, 2024. 2, 3

[52] Runyi Yang, Zhenxin Zhu, Zhou Jiang, Baijun Ye, Xiaoxue Chen, Yifei Zhang, Yuantao Chen, Jian Zhao, and Hao Zhao. Spectrally pruned gaussian fields with neural compensation. 2024. 3

[53] Xunpeng Yi, Han Xu, Hao Zhang, Linfeng Tang, and Jiayi Ma. Diff-retinex: Rethinking low-light image enhancement with a generative diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12302–12311, 2023. 3

[54] Zehao Yu, Torsten Sattler, and Andreas Geiger. Gaussian opacity fields: Efficient and compact surface reconstruction in unbounded scenes. *arXiv preprint arXiv:2404.10772*, 2024. 2

[55] Shiran Yuan and Hao Zhao. SlimmeRF: Slimmable radiance fields. In *2024 International Conference on 3D Vision (3DV)*, pages 64–74. IEEE, 2024. 3

[56] Vladimir Yugay, Theo Gevers, and Martin R. Oswald. Magic-slam: Multi-agent gaussian globally consistent slam. *arXiv preprint arXiv:2411.16785*, 2024. 3

[57] Toshiya Yura, Ashkan Mirzaei, and Igor Gilitschenski. Eventsplat: 3d gaussian splatting from moving event cameras for real-time rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 3

[58] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023. 2

[59] Ruida Zhang, Chengxi Li, Chenyangguang Zhang, Xingyu Liu, Haili Yuan, Yanyan Li, Xiangyang Ji, and Gim Hee Lee.

Street gaussians without 3d object tracker. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. 3

[60] Youjia Zhang, Anpei Chen, Yumin Wan, Zikai Song, Junqing Yu, Yawei Luo, and Wei Yang. Ref-gs: Directional factorization for 2d gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 3

[61] Ziyu Zhang, Binbin Huang, Hanqing Jiang, Liyang Zhou, Xiaojun Xiang, and Shuhan Shen. Quadratic gaussian splatting: High quality surface reconstruction with second-order geometric primitives. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. 2

[62] Shunyuan Zheng, Boyao Zhou, Ruizhi Shao, Boning Liu, Shengping Zhang, Liqiang Nie, and Yebin Liu. Gps-gaussian: Generalizable pixel-wise 3d gaussian splatting for real-time human novel view synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19680–19690, 2024. 2

[63] Chengliang Zhong, Peixing You, Xiaoxue Chen, Hao Zhao, Fuchun Sun, Guyue Zhou, Xiaodong Mu, Chuang Gan, and Wenbing Huang. SNAKE: Shape-aware neural 3d keypoint field. In *Advances in Neural Information Processing Systems*, 2022. 3

[64] Chengliang Zhong, Yuhang Zheng, Yupeng Zheng, Hao Zhao, Li Yi, Xiaodong Mu, Ling Wang, Pengfei Li, Guyue Zhou, Chao Yang, Xinliang Zhang, and Jian Zhao. 3d implicit transporter for temporally consistent keypoint discovery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3869–3880, 2023. 3

[65] Yuanzhi Zhu, Kai Zhang, Jingyun Liang, Jiezhang Cao, Bihan Wen, Radu Timofte, and Luc Van Gool. Denoising diffusion models for plug-and-play image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1219–1229, 2023. 3