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

Yunlong Li[1] Boyuan Chen[1] Chongjie Ye[1,2] Bohan Li[1,3]
Zhaoxi Chen[1,4] Shaocong Xu[5] Hao Tang[6] Hao Zhao[1,5]
[1]AIR, THU [2]CUHK [3]SJTU [4]NTU [5]BAAI [6]PKU

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 DrRobot 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).

1. Introduction

Modeling articulated robots requires a differentiable 3D representation that captures fine-grained geometry and view-dependent appearance for simulation and vision-based control. Gaussian fields are efficient and differentiable, but without explicit geometric constraints they tend to oversmooth details and produce artifacts under novel poses or viewpoints—as observed in prior work such as DrRobot [7]. Jointly supervising depth/normals while learning articulation further entangles geometry with pose, which often destabilizes optimization.

We propose GRADRobot, a compact pipeline that *decouples* geometry learning from articulation and adds an *optional* pose-aware refinement for high-frequency details. First, we learn a *canonical* Gaussian field in a fixed pose with pixel-space RGB, depth, and normal supervision that *anchors* Gaussians on the surface, stabilizing geometry. Second, we introduce a lightweight linear-blend-skinning (LBS) deformation conditioned on joint angles and *fine-tune* appearance using RGB-only losses, preserving canonical geometry while learning pose-dependent effects. Finally, when desired, a single-step structure-conditioned diffusion module (e.g., ControlNet [13]) refines thin structures and specularity without altering the commanded pose; this refiner can be disabled in latency-sensitive loops.

On a MuJoCo benchmark covering nine robot morphologies, GRADRobot reduces Chamfer Distance by up to 37% and improves PSNR by 0.6–1.2 dB over DrRobot under identical protocols. Geometry anchoring accounts for most of the CD gains; the optional refiner contributes the PSNR boost. Our evaluation is synthetic to enable controlled ground truth; real-robot experiments are an important direction for future work.

Our main contributions are:

(i) *Surface-anchored canonical GS:* pixel-space depth/normal losses keep Gaussians near the surface, yielding stable, geometry-aware reconstructions.

(ii) *Pose-conditioned articulation:* an LBS field with RGB-only fine-tuning decouples articulation from geometry while retaining full differentiability.

(iii) *Optional skeleton-conditioned refinement:* a single-step diffusion stage enhances high-frequency details without changing pose.

2. Related Work

2.1. Gaussian-Based Differentiable Rendering

3D Gaussian Splatting (GS) enables real-time, differentiable radiance-field rendering [5]. Extensions adapt GS to humans/sparse views [14], introduce depth/normal cues for inverse rendering [6], align shading with Gaussian normals [4], mitigate aliasing or enable editing via multi-scale

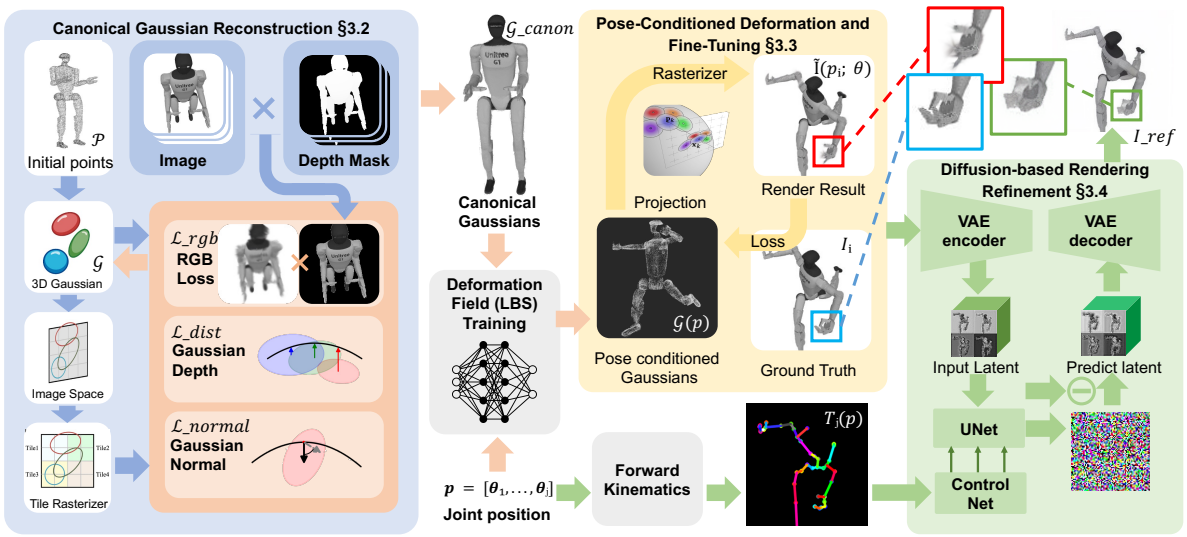

Figure 1. Overview of the GRADRobot Framework

The pipeline consists of canonical Gaussian reconstruction, pose-conditioned deformation, Optimization of Deformable Gaussians, and diffusion-based refinement for photorealistic, pose-aware 3D rendering.

kernels and mesh anchoring [2, 11], and optimize surface-aware opacity [12]. For articulated robots, DrRobot couples GS with kinematics-aware deformation to backpropagate image-space gradients to joint angles [7], but typically learns geometry and articulation jointly without explicit geometric constraints, which can cause over-smoothing and pose-dependent artifacts. In contrast, we first learn a *canonical* field with *pixel-space* depth/normal supervision that *anchors* Gaussians on the surface, then apply LBS articulation with RGB-only fine-tuning—decoupling geometry from appearance and improving stability. Most of the above GS variants are designed for static or human-centric reconstruction and cannot directly handle kinematic articulation. DrRobot is the prior method that explicitly targets articulated robot rendering, which is why we focus on it as the primary baseline in our comparisons.

2.2. Generative Diffusion Models

Diffusion models achieve strong results in image synthesis and restoration [8, 9], with plug-and-play priors improving fidelity/efficiency [1, 15]. ControlNet enables conditioning on external structure [13]. We use an *optional, single-step* skeleton-conditioned refiner: given a coarse GS render and a forward-kinematics skeleton, it recovers thin structures/high-frequency details while preserving the commanded pose. The refiner can be disabled in latency-sensitive, real-time loops.

3. Method

We target articulated rendering with stable geometry and high-frequency detail. GRADRobot decouples learning into three stages: (I) *surface-anchored* canonical reconstruction, (II) pose-conditioned articulation with RGB-only fine-tuning, and (III) an *optional* skeleton-conditioned, single-step diffusion refiner (Fig. 1).

Accurately rendering articulated robots across diverse poses requires a representation that is both geometrically consistent and visually detailed. Jointly learning geometry and articulation often entangles the two and destabilizes optimization. We therefore first learn a canonical Gaussian field under a fixed pose with explicit pixel-space geometry supervision; next, we articulate it using LBS and fine-tune appearance using only RGB; finally, when desired, we apply a pose-aware diffusion refiner that preserves the commanded kinematics.

3.1. Canonical GS with Pixel-Space Geometry Losses

We represent the canonical scene with M anisotropic Gaussians $\mathcal{G} = \{(\mu_i, \Sigma_i, c_i, \alpha_i)\}_{i=1}^{M}$ and render by differentiable splatting [5]. Given a rendered image $\hat{I}$ and ground-truth I, the objective is

$$\begin{aligned} \mathcal{L}_{\text{canon}} &= \mathcal{L}_{\text{rgb}} + \lambda_g \, \mathcal{L}_{\text{geo}}, \\ \mathcal{L}_{\text{rgb}} &= \|I - \hat{I}\|_1 + \lambda_{\text{ssim}} \, \text{DSSIM}(I, \hat{I}). \end{aligned} \tag{1}$$

To *anchor* Gaussians on the surface, we add depth distortion and normal consistency in pixel space. Let Ω be pixel

Table 1. Mean Chamfer Distance ($\times 10^{-3}$ m) over 50 random poses.

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**

indices, $\mathcal{G}(p)$ the Gaussians contributing to pixel p, $\bar{w}_i(p)$ the *stop-gradient* blending weight from the rasterizer, p^z the ground-truth depth, $z_i(p)$ the depth of Gaussian i at p, $n_i(p)$ its local normal, and $\hat{n}(p)$ the normal from depth gradients:

$$\mathcal{L}_{\text{geo}} = \lambda_d \mathcal{L}_{\text{dist}} + \lambda_n \mathcal{L}_{\text{normal}}, \tag{2}$$

$$\mathcal{L}_{\text{dist}} = \sum_{p\in\Omega} \sum_{i\in\mathcal{G}(p)} \bar{w}_i(p)\,|z_i(p) - p^z|, \tag{3}$$

$$\mathcal{L}_{\text{normal}} = \sum_{p\in\Omega} \sum_{i\in\mathcal{G}(p)} \bar{w}_i(p)\,\|n_i(p) - \hat{n}(p)\|_2^2. \tag{4}$$

This surface anchoring reduces interior artifacts and improves articulation; in ablation it yields about 8% CD gains (Table 1).

3.2. Pose-Conditioned Deformation and RGB-Only Fine-Tuning

Let $g_j(\theta) \in \mathrm{SE}(3)$ be the joint transform of joint j and $\mathcal{C}(j)$ its kinematic chain from root to j. For a canonical Gaussian center $x \in \mathbb{R}^3$ and homogeneous $\tilde{x} = [x^\top, 1]^\top$, the posed point is

$$T(\theta, x) = \sum_{j=1}^{J} \underbrace{\frac{\exp(s_j(x))}{\sum_k \exp(s_k(x))}}_{w_j(x)} \Big(\prod_{k\in\mathcal{C}(j)} g_k(\theta) \Big) \tilde{x}, \tag{5}$$

where $s_j(\cdot)$ is the MLP output before softmax. For anistropic Gaussians with covariance $\Sigma \in \mathbb{R}^{3\times3}$, we transform both mean and covariance via the linear part $R_j(\theta)$ of the chain:

$$\mu' = \sum_j w_j \left(R_j \mu + t_j\right), \qquad \Sigma' = \sum_j w_j\, R_j \Sigma R_j^\top, \tag{6}$$

which preserves positive semidefiniteness and aligns covariance axes with link frames. We additionally regularize $\|\Sigma\|_F$ and $\alpha(1-\alpha)$ to discourage oversized kernels and saturated opacity.

3.3. Skeleton-Conditioned Single-Step Diffusion

For thin structures/specularity, we add an *optional* single-step refiner built on Stable Diffusion v1–5 [3, 8] with a ControlNet [13]. Given the posed coarse render I_{coarse} and a color-coded skeleton map $S(\theta)$ from forward kinematics, we encode I_{coarse} with a frozen VAE and apply one denoising step conditioned on $S(\theta)$:

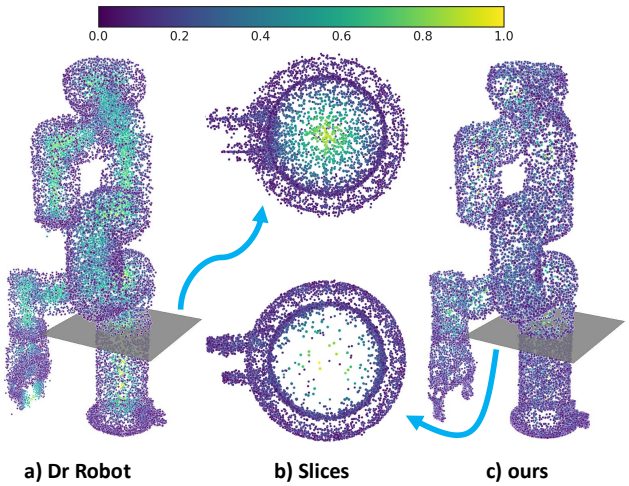

Figure 2. Chamfer-distance visualisation on the *ufactory_xarm7* . (a) DrRobot; (b) Slice; (c) Ours

$$\begin{aligned} \hat{z} &= z(I_{\text{coarse}}) - \varepsilon_\theta\big(z(I_{\text{coarse}}),\, t{=}0,\, c,\, f_\omega(S(\theta))\big), \\ I_{\text{ref}} &= \text{VAE}_{\text{dec}}(\hat{z}). \end{aligned} \tag{7}$$

Training uses $\|I_{\text{ref}} - I\|_1 + \lambda_p \|\phi(I_{\text{ref}}) - \phi(I)\|_2^2$ with VGG features ϕ, freezing the VAE/text encoder and tuning ControlNet/U-Net. The refiner preserves pose and can be disabled for real-time loops.

4. Experiments

4.1. Dataset

We build a synthetic benchmark in MuJoCo [10]. For each robot, we remove fixed joints and standardize lighting. Joint angles are partitioned into: (i) 500 collision-free *canonical* poses, (ii) 10,000 *training* poses (allowing up to 10 self-collisions), and (iii) 500 *test* poses. Each pose is rendered from 12 RGB-D views at 256×256 with azimuth sampled from three uniform bins in $[-180°, 180°]$ (with jitter), elevation in $\{-45°, 45°\}$, and radius in $\{1.0, 2.0\}d$. We back-project each RGB-D frame into a camera-centric point cloud with Open3D, transform it to the world frame, and voxel-downsample to 0.01 m. The 12 per-view clouds are merged and sparsified by progressive voxelization (starting at 0.005 m and increasing by 1% per step) until $\leq$ 10k points, yielding one fused ground-truth cloud per pose.

Table 2. Average PSNR (dB) for each robot model.

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**

4.2. Chamfer Distance Evaluation

Evaluation. We sample 50 poses per robot, deform the learned Gaussians via the pose-conditioned LBS network, and export the resulting 3D points. We then compute the symmetric Chamfer Distance between the deformed cloud $\hat{\mathcal{P}}$ and the ground-truth cloud $\mathcal{P}$:

$$\mathrm{CD}(\hat{\mathcal{P}}, \mathcal{P}) = \frac{1}{|\hat{\mathcal{P}}|} \sum_{\mathbf{x} \in \hat{\mathcal{P}}} \min_{\mathbf{y} \in \mathcal{P}} \|\mathbf{x} - \mathbf{y}\|_2^2 \quad + \frac{1}{|\mathcal{P}|} \sum_{\mathbf{y} \in \mathcal{P}} \min_{\mathbf{x} \in \hat{\mathcal{P}}} \|\mathbf{y} - \mathbf{x}\|_2^2. \tag{8}$$

Baseline. We compare against a re-implementation of "DrRobot", where canonical Gaussians are deformed and fine-tuned by the same LBS network under the same 50-pose protocol, with Chamfer Distance measured identically.

Results. Table 1 reports mean CD ($\times 10^{-3}$ m) over 50 poses. GRADRobot achieves a 37% reduction versus DrRobot, demonstrating tighter surface alignment and far fewer spurious interior points. This surface-proximal distribution is crucial: without it, internal points deform inconsistently, causing geometry collapse or visual artifacts, whereas our method maintains structural integrity and artifact-free deformation.

Figure 2 further illustrates these gains. The global error maps ((a) vs. (b)) show markedly smaller high-error regions, and the horizontal slice in (c) confirms our points lie on the cylindrical link surface, while DrRobot produces a dense ring of internal artifacts—highlighting the coherence of our Gaussian representation before and after articulation.

Ablation Study Row "Ours w/o Geo Loss" in table 1 removes the two surface-anchoring terms introduced in subsection 3.1—the Depth-Distortion (dist) loss and the Normal-Consistency (normal) loss. Re-adding the losses cuts Chamfer Distance by a further 8% on average.

4.3. Image Quality Evaluation

Evaluation Protocols. We evaluate geometric and image-space performance on a held-out set of 500 poses per robot. For each pose, we render several 256×256 views. Image quality is measured by PSNR between diffusion-refined renders and ground-truth RGB. All methods are trained on the same MuJoCo dataset and tested on unseen joint configurations.

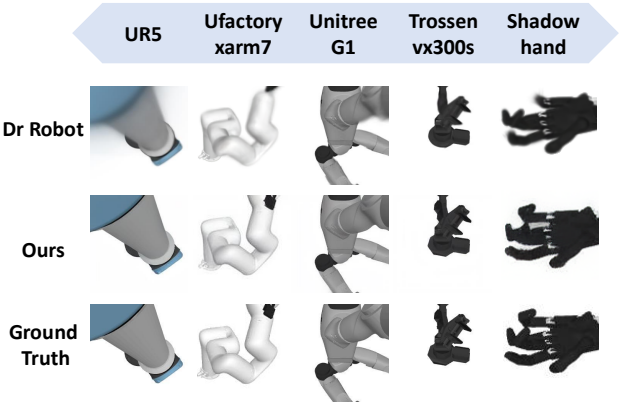

Figure 3. Qualitative results across several robot morphologies.

Baselines. We compare GRADRobot (diffusion-enhanced splatting) to DrRobot's three-stage pipeline (Gaussian splatting + LBS + joint fine-tuning) without diffusion, using identical data, cameras, and backgrounds.

Results. Table 2 shows GRADRobot achieves the highest PSNR across nine robots, with 0.6–1.2,dB gains over the pose-conditioned splat and $>$ 1.0,dB over DrRobot. The diffusion module restores high-frequency details (e.g., thin links, sharp edges), yielding renders closer to ground truth—important for downstream vision-driven control. Qualitative results in Fig. 3 show sharper contours for GRADRobot, while DrRobot appears overly smooth.

Ablation Study. *Ours w/o refine*" disables diffusion and uses the pose-conditioned splat as output, lowering mean PSNR from 32.6 to 31.9 dB (-0.7 dB). *Ours w/o Geo Loss*" reduces PSNR from 32.46 to 30.90,dB, yet still slightly outperforms DrRobot (30.59,dB).

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 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 that geometry losses and diffusion are complementary in locking Gaussians to the true surface and restoring high-frequency detail.

References

[1] 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. 2

[2] 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

[3] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, pages 6840–6851. Curran Associates, Inc., 2020. 3

[4] 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. 1

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

[6] 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. 1

[7] Ruoshi Liu, Alper Canberk, Shuran Song, and Carl Vondrick. Differentiable robot rendering, 2024. 1, 2

[8] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 2, 3

[9] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*. PMLR, 2015. 2

[10] 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. 3

[11] 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

[12] Zehao Yu, Torsten Sattler, and Andreas Geiger. Gaussian opacity fields: Efficient adaptive surface reconstruction in unbounded scenes, 2024. 2

[13] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023. 1, 2, 3

[14] 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. 1

[15] 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. 2