MGD:Mesh-guided Gaussians with Diffusion Priors for Dynamic Objects Reconstruction from Monocular RGB-D Video

Weixing Xie[2,4*], Ying Ye[1*], Xian Wu[3], Jintian Li[3], Bingchuan Li[1], Yanchen Lin[5], Junfeng Yao[1,2,3,4†]

[1]School of Informatics, Xiamen University
[2]National Institute for Data Science in Health and Medicine, Xiamen University
[3]School of Film, Institute of Artificial Intelligence, Xiamen University
[4]Key Laboratory of Digital Protection and Intelligent Processing of Intangible Cultural Heritage of Fujian and Taiwan Ministry of Culture and Tourism
[5]College of Computer and Cyber Security, Fujian Normal University
xwxxmu@gmail.com, {xieweixing@stu.,yeying0513@stu.,yao0010@}xmu.edu.cn

Abstract

Reconstructing dynamic objects from monocular RGB-D video is critical for advancing 3D vision applications and enhancing user experience. However, monocular RGB-D video provides limited 3D observations, making the reconstruction of unobserved regions highly under-constrained. Despite recent advances that combine neural implicit surfaces with diffusion models, the inherent limitations of implicit representations and the lack of effective guidance in diffusion priors lead to blurry appearance and inaccurate geometry in dynamic object reconstruction. To address the issue, we present MGD, which leverages scene-adaptive diffusion priors and Mesh-guided Gaussians for realistic rendering and geometrically accurate reconstruction of dynamic objects, including unobserved regions. The dynamic 3D objects reconstructed by MGD are represented using our proposed Mesh-guided Gaussians, which leverage global and local Gaussians to capture large-scale deformations and fine-grained appearance details, respectively. Additionally, in order to utilize depth information, we integrate a depth ControlNet into the diffusion model and conduct scene-adaptive fine-tuning. We design a self-generated image-pair strategy to produce image pairs used for fine-tuning. Extensive experiments demonstrate that MGD achieves state-of-the-art performance in both high-fidelity reconstruction and structural completeness, while maintaining real-time efficiency during training and rendering.

Introduction

Fast and high-fidelity 3D object reconstruction from a set of input images is necessary for many applications such as AR/VR, 3D content production and entertainment. Although RGB-D cameras provide depth information, the limited viewpoints in monocular videos still pose a significant challenge for high-fidelity 3D reconstruction.

Early methods leverage Neural Radiance Fields (NeRF) (Mildenhall et al. 2020) to reconstruct dynamic scenes, focusing on observed regions while exploring extensions to unobserved regions. NDR (Cai et al. 2022) reconstructs dynamic scenes from monocular RGB-D videos, recovering observed geometry but struggling with unobserved regions. MorpheuS (Wang, Wang, and Agapito 2024) employs diffusion models to supervise unobserved regions, achieving full 360° reconstruction. Nevertheless, it exhibits geometric inconsistency and unrealistic textures in unobserved views. Moreover, the high computational cost of training and rendering in NeRF-based methods further limits their practicality. To improve rendering efficiency and visual quality, recent studies have integrated diffusion priors into 3D Gaussian Splatting (3DGS) (Kerbl et al. 2023). Some methods (Wang et al. 2025; Zhang et al. 2025) leverage explicit templates (e.g., SMPL (Loper et al. 2023)) for structural guidance, but their performance heavily relies on template accuracy, limiting generalization to non-human or freely deformable scenes. Template-free approaches like DreamGaussian4D (Ren et al. 2023) enhance texture synthesis via diffusion but often produce floating artifacts and structural distortions due to the lack of stable geometric constraints. From existing studies, although diffusion priors show strong potential in texture generation, they remain limited by weak structural awareness, poor geometric consistency, and insufficient scene adaptability, which hinders the realism of completion. Therefore, the task of dynamic 3D object reconstruction from monocular RGB-D videos currently faces two major challenges: (1) the lack of an effective representation that can reliably establish accurate geometry and efficiently reconstruct high-quality textures from monocular input; (2) the lack of a strategy to fully exploit the generative capabilities of diffusion models while effectively leveraging additional depth information to overcome the limitations imposed by single-view input and complex scenes.

To address the aforementioned challenges, we present MGD for high-quality 3D reconstruction from monocular RGB-D videos. To tackle the first challenge, some existing methods (Guédon and Lepetit 2024; Liu, Su, and Wang 2025; Zheng et al. 2025) combine mesh and Gaussian representations, leveraging the rendering flexibility of 3D Gaussians along with the strong geometric priors provided by

*These authors contributed equally.
†Corresponding author.

mesh structures. However, these approaches often rely on point-wise deformation strategies, which are inefficient and tend to produce low-quality results due to the lack of modeling of deformation similarity among neighboring vertices. Moreover, Gaussians constrained to the mesh surface suffer from limited degrees of freedom during deformation, which further degrades the rendering quality. To this end, we propose a novel Mesh-guided Gaussians hybrid representation, as illustrated in Fig. 1, which consists of two complementary components: a Global Gaussian Deformation Module (GGDM) that drives mesh deformation to model large-scale object motion, and a Local Gaussian Adjustment Module (LGAM) that adjusts the attributes of Local Gaussians attached to the mesh for capturing fine-grained dynamic details. For the second challenge, we incorporate a depth ControlNet into the diffusion prior to enhance texture-geometry alignment using additional depth information. To further adapt the diffusion prior to scene-specific geometry and appearance, we design a two-stage scene-adaptive fine-tuning strategy.

In summary, the contributions of our method are as follows: **First,** we propose a novel Mesh-guided Gaussians representation that integrates global and local Gaussians to capture large-scale deformations and fine-grained appearance details. **Second,** we introduces ControlNet-based geometric conditioning to improve texture-geometry alignment and propose a scene-adaptive fine-tuning strategy to improve the diffusion prior's adaptability. **Third,** the overall framework MGD consistently outperforms current SOTA methods in both high-fidelity reconstruction and structural completeness, while maintaining efficiency during training and real-time rendering.

Related Work

RGB-D based dynamic reconstruction. Neural implicit representations provide strong expressive capabilities for RGB-D based dynamic reconstruction. Some works (Bozic et al. 2020; Lin et al. 2022; Cai et al. 2022) extend NeRF (Mildenhall et al. 2020) to dynamic reconstruction, effectively enhancing the modeling of geometry and appearance. Subsequently, 3DGS (Kerbl et al. 2023) has been applied to dynamic reconstruction due to its highly efficient rendering, but suffers from limited geometric consistency. To address this, some methods (Guédon and Lepetit 2024; Liu, Su, and Wang 2025; Zheng et al. 2025) further incorporate explicit meshes with Gaussians, introducing structural priors to improve the stability and fidelity of dynamic reconstruction. Although existing mesh-Gaussian hybrid representations enhance structural consistency, they remain insufficient in handling large-scale non-rigid motion and representing local fine-grained details. To this end, we propose a novel Mesh-guided Gaussians representation that jointly models deformation and high-frequency details through a coordinated design of global control and local flexibility. Compared to NeRF (Mildenhall et al. 2020)-based representations, our method significantly improves training and rendering efficiency while maintaining high reconstruction quality.

Diffusion Models for dynamic reconstruction. In recent years, diffusion priors have demonstrated powerful capabilities in generative modeling, 3D reconstruction, and other fields. Since DreamFusion (Poole et al. 2022) introduced Score Distillation Sampling (SDS), a series of follow-up works (Chen et al. 2023; Lin et al. 2023; Wang et al. 2023) have extended this approach to enable the generation of dynamic 3D scenes from text prompts. Another category of methods (Tang et al. 2023; Yi et al. 2024) leverage 3D Gaussian representations to achieve efficient 3D content generation under multi-view inputs, balancing rendering quality and generation speed. MorpheuS (Wang, Wang, and Agapito 2024) is a notable work that combines diffusion priors with SDS loss to optimize NeRF representations from monocular RGB-D videos, enabling dynamic 3D reconstruction across 360° viewpoints. Although these methods fully demonstrate the potential of diffusion models in dynamic scene reconstruction, they lack sufficient adaptability to specific scene geometry and appearance, resulting in difficulty generating highly consistent and detail-rich reconstructions. Our method explores scene-adaptive fine-tuning of diffusion models to better align with the geometry and appearance features of input scenes, thereby improving the modeling of object surface geometry and textures.

Method

We introduce MGD, a novel Mesh-guided Gaussians representation equipped with a scene-adaptive diffusion prior enhanced by a depth ControlNet. This design enables high-fidelity reconstruction and structural completeness of dynamic objects from a single viewpoint. As shown in Fig. 1, MGD comprises two components: the Mesh-guided Gaussians representation and a two-stage training strategy, which are detailed in the following sections.

Preliminary

3D Gaussian Splatting. 3DGS is an explicit and differentiable scene representation that models a 3D scene as a collection of anisotropic Gaussian primitives. Each Gaussian primitive $\mathcal{G}_i$ is defined by a center $\mathbf{x}_i \in \mathbb{R}^3$, a color vector $\mathbf{c}_i \in \mathbb{R}^3$, an opacity $\sigma_i \in \mathbb{R}$, and a covariance matrix $\mathbf{\Sigma}_i \in \mathbb{R}^{3\times3}$ that controls its spatial extent. The covariance matrix $\mathbf{\Sigma}_i$ is further decomposed as: $\mathbf{\Sigma}_i = \mathbf{R}_i\mathbf{S}_i\mathbf{S}_i^\top\mathbf{R}_i^\top$, where $\mathbf{S}_i = \mathrm{diag}(\mathbf{s}_i)$ is a diagonal scaling matrix with $\mathbf{s}_i \in \mathbb{R}^3$, and $\mathbf{R}_i$ is the rotation matrix from the quaternion $\mathbf{r}_i \in \mathbb{R}^4$. During rendering, the Gaussians are projected onto the image plane and composited via differentiable alpha blending. The color is further modulated using spherical harmonics to model view-dependent appearance, enabling real-time photo-realistic synthesis.

View-Conditioned Diffusion and ControlNet. Diffusion models generate images by gradually denoising Gaussian noise through a learned reverse process. Latent Diffusion Models (LDM) (Rombach et al. 2022) improve efficiency by operating in a compressed latent space. ControlNet (Zhang, Rao, and Agrawala 2023) builds on LDM by introducing a parallel branch that takes structural conditions, enhancing spatial consistency during generation.

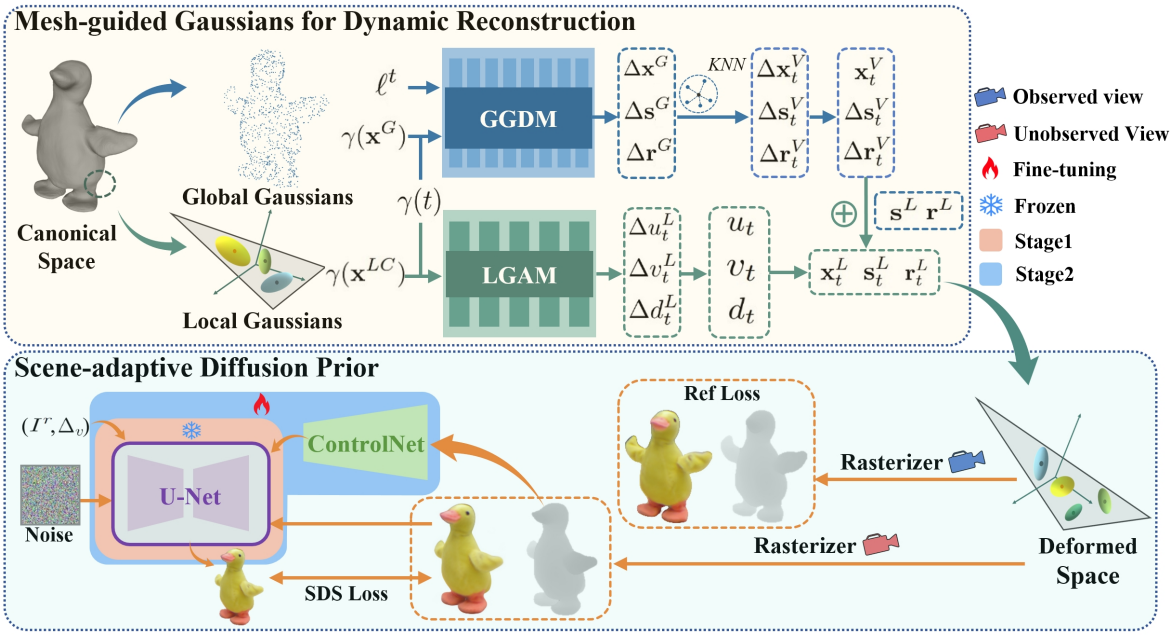

Figure 1: **The framework of MGD. Mesh-guided Gaussians:** We employ a GGDM to drive mesh deformation to model large-scale object motions, and design a LGAM to adjust Local Gaussians attributes attached to the mesh for capturing dynamic details. **Two-Stage Training:** In Stage1, we employ the Mesh-guided Gaussians with View-Conditioned Diffusion to obtain a coarse object representation. In Stage2, a depth ControlNet is integrated to enhance texture-geometry alignment of the diffusion. Then, leveraging the coarse representation, we generate paired training images to fine-tune the diffusion for each scene, adapting it to the scene-specific data distribution. Finally, the coarse representation is refined with a scene-adaptive diffusion prior for high-fidelity reconstruction.

In view-conditioned generation, models such as Zero-1-to-3 (Liu et al. 2023) synthesize novel views based on an input image and a relative camera transformation Δ_v. The camera pose is typically represented in spherical coordinates (α, β, ρ), where α, β, and ρ denote the elevation, azimuth, and radius. The relative transformation is defined as $\Delta_v = (\alpha_t - \alpha_s,\ \beta_t - \beta_s,\ \rho_t - \rho_s)$, enabling view synthesis from monocular input for downstream 3D reconstruction.

Mesh-guided Gaussians for Dynamic Reconstruction

We propose a hybrid Mesh-guided Gaussians representation for dynamic reconstruction that combines global and local Gaussians to capture both large-scale deformations and fine-grained appearance variations.

Mesh-guided Gaussians. We utilize a mesh representation as a structural prior and distribute N Local Gaussians $\mathcal{G}^L$ on each triangular face. Each Local Gaussian is represented using a local coordinate system (u, v, d), where $(u, v) \in [0, 1]^2$ indicates the relative location on the face, and $d \in \mathbb{R}$ denotes the offset from the surface along the normal direction.

Given a triangular face f with vertices $\{\mathbf{V}_1, \mathbf{V}_2, \mathbf{V}_3\}$ and normal vector $\mathbf{n}$, the point $\mathbf{p}$ on the face corresponding to the barycentric coordinates (u, v) is computed as:

$$\mathbf{p} = \mathcal{V}_p(f, u, v) = u \cdot \mathbf{V}_1 + v \cdot \mathbf{V}_2 + (1 - u - v) \cdot \mathbf{V}_3, \quad (1)$$

where $\mathcal{V}(f, u, v)$ interpolates per-point geometric attributes (e.g., position, rotation, scale) on the f-th triangle using barycentric coordinates (u, v) and the attributes of its vertices. The global position $\mathbf{x}^L$ of the Local Gaussian is then obtained by offsetting the interpolated point $\mathbf{p}$ along the face normal $\mathbf{n}$ by a distance d:

$$\mathbf{x}^L = \mathbf{p} + d \cdot \mathbf{n}. \quad (2)$$

Deformation for Mesh-guided Gaussians. We design two modules: Global Gaussian Deformation Module (GGDM) and Local Gaussian Adjustment Module (LGAM). GGDM predicts global deformations, while LGAM adjusts the position, rotation, and scale of Local Gaussians.

Global Gaussian Deformation Module. We simplify the modeling of large-scale non-rigid deformations using a set of Global Gaussians $\mathcal{G}^G$. Each Global Gaussian $\mathcal{G}_i^G$ is represented as a 3D Gaussian distribution, parameterized by a center position $\mathbf{x}_i^G$, scale $\mathbf{s}_i^G$ and rotation $\mathbf{r}_i^G$. We define a transformation function $\mathcal{F}_G$ that transforms Global Gaussians from one space to another by:

$$(\Delta\mathbf{x}_i^G, \Delta\mathbf{s}_i^G, \Delta\mathbf{r}_i^G) = \mathcal{F}_G(\ell_i^t, \gamma(\mathbf{x}_i^G), \gamma(t)), \quad (3)$$

where $\gamma(\cdot)$ is a frequency-based embedding function used to enhance the spatiotemporal modeling capacity. Specifically, to capture complex variations of the global transformation across spatial and temporal dimensions, we introduce a learnable, frame-dependent latent code ℓ_i^t into the function.

To enable the geometric deformation transfer from Global Gaussians to Local Gaussians, we augment each mesh vertex $\mathbf{V}$ with additional deformation attributes: scale $\mathbf{s}^V$ and rotation $\mathbf{r}^V$. Following SC-GS (Huang et al. 2024), we compute the deformation of each mesh vertex $\mathbf{V}$ by interpolating from its K nearest global Gaussians in space, with weights determined by the Mahalanobis distance:

$$(\Delta\mathbf{x}_t^V, \Delta\mathbf{s}_t^V, \Delta\mathbf{r}_t^V) = \sum_{k\in\mathcal{C}} w_k \cdot \mathcal{F}_G(\ell_k^t, \gamma(\mathbf{x}_k^G), \gamma(t)), \quad (4)$$

where $\mathcal{C}$ denotes the set of K Global Gaussians neighboring the vertex $\mathbf{V}$, the weights w_k are computed as follows:

$$\delta_k = (\mathbf{x}^V - \mathbf{x}_k^G)^\top (\mathbf{\Sigma}_k^G)^{-1} (\mathbf{x}^V - \mathbf{x}_k^G), \quad (5)$$

$$w_k = \text{softmax}([-\delta_1, -\delta_2, ..., -\delta_k]), \quad (6)$$

where δ_k is defined as the squared distance.

Local Gaussian Adjustment Module. Despite the GGDM's strength in modeling global geometric dynamics, our experiments show that the strict fixation of Local Gaussians to the surfaces of triangular faces limits the expressiveness of Mesh-guided Gaussians representation in capturing complex and fine-grained local texture variations. To address this limitation, we introduce an adjustment function $\mathcal{F}_L$ that adjusts the attributes of each Local Gaussian and grants it a certain degree of freedom within the local region of its associated face. The adjustment function $\mathcal{F}_L$ takes as input its local coordinates $\mathbf{x}^{LC} = (u^L, v^L, d^L)$ and the time step t:

$$\Delta\mathbf{x}_t^{LC} = (\Delta u_t^L, \Delta v_t^L, \Delta d_t^L) = \mathcal{F}_L(\gamma(\mathbf{x}^{LC}), \gamma(t)). \quad (7)$$

Then, the deformed vertex $\mathbf{V}$ and the local coordinates of the Local Gaussian are given by:

$$\mathbf{x}_t^V = \mathbf{x}^V + \Delta\mathbf{x}_t^V, \quad (u_t, v_t, d_t) = \mathbf{x}^{LC} + \Delta\mathbf{x}_t^{LC}. \quad (8)$$

Here, the final attributes of the Local Gaussian can be expressed as:

$$\mathbf{x}_t^L = \mathcal{V}_p(u_t, v_t) + d_t \cdot \mathbf{n}, \quad (9)$$

$$\mathbf{s}_t^L = \mathbf{s}^L \cdot \mathcal{V}_s(u_t, v_t),\ \mathbf{r}_t^L = \mathbf{r}^L \cdot \mathcal{V}_r(u_t, v_t). \quad (10)$$

Two-Stage Training Strategy

In the first stage, we learn a coarse object representation using Mesh-guided Gaussians from a single RGB-D video, aided by a pre-trained diffusion prior to complete unobserved regions. In the second stage, the representation is refined by a diffusion model enhanced with depth ControlNet and fine-tuned for scene adaptation.

Stage 1: Mesh-guided Gaussians with Diffusion Prior. We leverage pre-trained diffusion priors from unobserved views to optimize the Mesh-guided Gaussians via the SDS loss:

$$\mathcal{L}_{\text{SDS}} = \mathbb{E}_{t,\epsilon}\left[w(t) \cdot \|\epsilon_\phi(\mathbf{z}_t,\, t,\, I^r,\, \Delta_v) - \epsilon\|_2^2\right], \quad (11)$$

where $\epsilon_\phi(\cdot)$ denotes the noise predicted by the diffusion prior ϕ, Δ_v represents the relative camera pose with respect to the reference view r, and w is a time-dependent weighting factor. Inspired by GaussianObject (Yang et al. 2024), we weight the SDS loss from unobserved views according to the Euclidean distance between a novel view q and the reference view r:

$$\lambda(q) = \min\left(1,\, \kappa \cdot \|q - r\|_2\right), \quad (12)$$

where κ is a scaling factor that controls the influence of the confidence weight. The loss function of unobserved regions is defined as:

$$\mathcal{L}_{\text{unobs}} = \lambda_1 \cdot \lambda(q) \cdot \mathcal{L}_{\text{SDS}}, \quad (13)$$

where λ_1 is a global weight for the unobserved-view loss. For the observed regions, to preserve mesh surface smoothness and local rigidity, we adopt the Laplacian regularization following DG-Mesh (Liu, Su, and Wang 2025) and incorporate the As-Rigid-As-Possible (ARAP) loss as introduced in SC-GS (Huang et al. 2024):

$$\mathcal{L}_{\text{lap}} = \frac{1}{n}\sum_{i=1}^{n} \|V_i - \frac{1}{|N_i|}\sum_{k\in N_i} V_k\|_2^2, \quad (14)$$

$$\mathcal{L}_{\text{arap}} = \sum_{k\in N_i} \omega_{ik} \|(\tilde{V}_i - \tilde{V}_k) - R_{V_i}(V_i - V_k)\|_2^2, \quad (15)$$

where N_i denotes the one-ring neighborhood of vertex V_i, and n is the total number of mesh vertices. R_{V_i} is the local rotation matrix associated with V_i, and ω_{ik} is the edge weight between V_i and its neighbor V_k. The loss function of observed regions is defined as:

$$\begin{aligned}\mathcal{L}_{\text{ref}} = \lambda_2\mathcal{L}_{\text{C}} + \lambda_3\mathcal{L}_{\text{M}} + \lambda_4\mathcal{L}_{\text{N}} \\ + \lambda_5\mathcal{L}_{\text{D}} + \lambda_6\mathcal{L}_{\text{lap}} + \lambda_7\mathcal{L}_{\text{arap}}.\end{aligned} \quad (16)$$

Here, λ_2 to λ_7 denote the weighting coefficients balancing the contributions of the respective loss terms. The RGB reconstruction loss $\mathcal{L}_{\text{C}}$, mask loss $\mathcal{L}_{\text{M}}$, and depth loss $\mathcal{L}_{\text{D}}$ are all formulated using the ℓ_2 norm, while the normal consistency loss $\mathcal{L}_{\text{N}}$ is implemented following the PyTorch3D framework. Hence, the overall loss in the first stage is defined as:

$$\mathcal{L}_{\text{Stage1}} = \mathcal{L}_{\text{unobs}} + \mathcal{L}_{\text{ref}}. \quad (17)$$

Stage 2: Mesh-guided Gaussians Refinement with Scene-adaptive Diffusion Prior. We first enhance the diffusion prior by incorporating a ControlNet, followed by a fine-tuning process. Subsequently, the fine-tuned diffusion model is employed to further refine Mesh-guided Gaussians representation obtained from stage 1.

Adding Depth ControlNet. Despite the strong generative capabilities of diffusion models, they often struggle to maintain spatial and geometric consistency. To address this limitation, we incorporate a depth ControlNet to condition the diffusion process on depth maps derived from RGB-D inputs, serving as geometric priors. Following ControlNet (Zhang, Rao, and Agrawala 2023), we design the module to guide the frozen diffusion model via an additional trainable branch, enhancing spatial controllability, structural consistency, and texture fidelity.

Learning Scene-Adaptive Diffusion Prior. Although we introduce a ControlNet to provide geometric priors during the generation process, directly using a ControlNet pre-trained on general datasets still presents certain limitations.

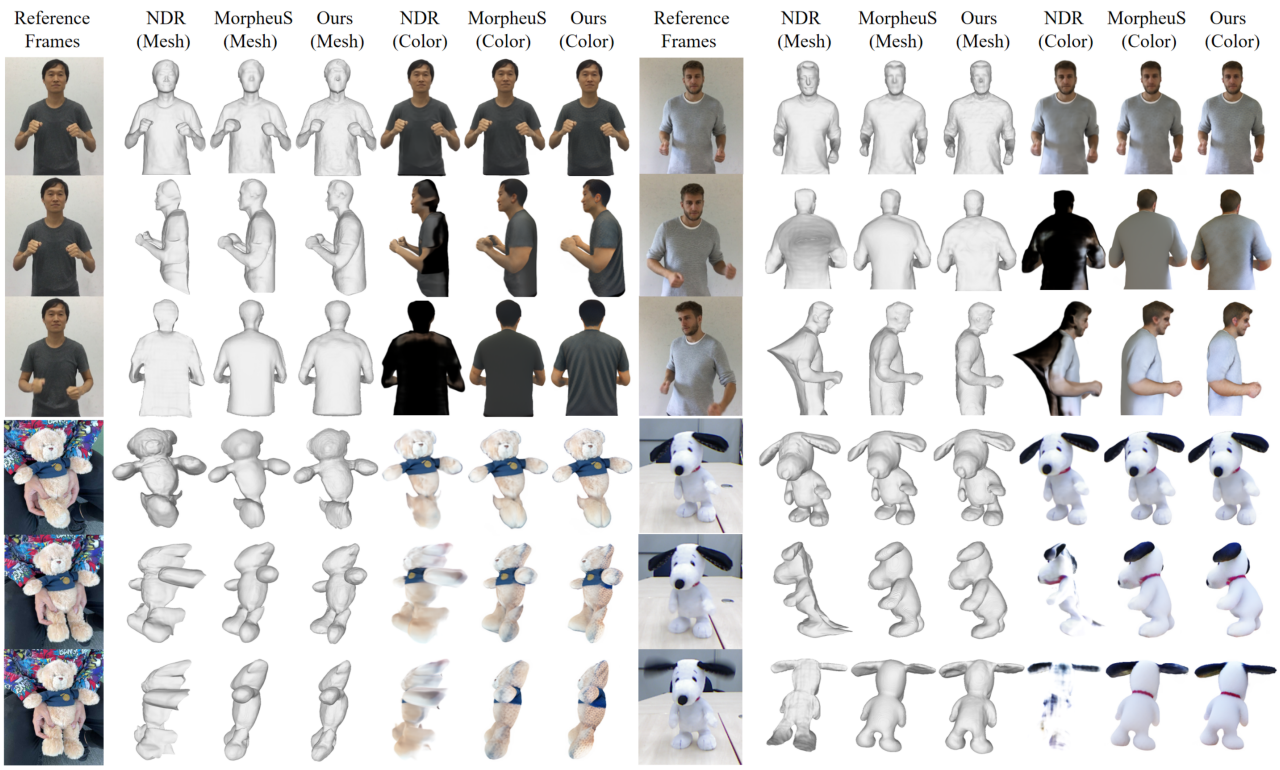

Figure 2: **Qualitative comparison on three datasets.** We compare the novel view synthesis quality with Morphues and NDR.

We observe that the pre-trained model struggles to fully adapt to the unique geometric characteristics and texture distributions of a specific scene. Therefore, to further enhance structural consistency and detail fidelity in the target scene, we perform targeted fine-tuning of the entire diffusion model, enabling it to better align with the depth distribution and visual features of the current scene.

We design two image pair construction strategies based on Mesh-guided Gaussians trained in the first stage: Cross-View Supervision and Gaussian Perturbation Simulation. For the Cross-View Supervision strategy, considering that the input data consist of monocular RGB-D video sequences where each frame contains only a single ground-truth viewpoint, we leverage the rendering capability of the trained Mesh-guided Gaussians across different views. Specifically, at each time step t, we select a viewpoint within Mesh-guided Gaussians that differs from the current real viewpoint r, and render the image I'. The second strategy, Gaussian Perturbation Simulation, involves statistically analyzing the temporal variations of Local Gaussians' attributes across adjacent frames to extract their mean and standard deviation. Based on these statistics, controlled random perturbations Δ are applied to the current Local Gaussians $\mathcal{G}_t^L$. The perturbed Local Gaussians $\mathcal{G}_t^{L\prime} = \mathcal{G}_t^L + \Delta$ are then rendered from the real viewpoint r to generate a rendered image I'. Thus, we construct image pair $(I', (I^r, D^r))$.

We adopt the Low-Rank Adaptation (LoRA) (Hu et al. 2022) strategy to efficiently fine-tune the overall diffusion model. LoRA layers are injected into both the ControlNet and the U-Net, and only the parameters of these inserted layers are updated during training, which preserves the stability and generalization ability of the pre-trained model. During training, the rendered image is used as the input to the diffusion model, the ground-truth depth map serves as the geometric condition, and the ground-truth image provides the supervision target. The loss function used for LoRA fine-tuning is defined as:

$$\mathcal{L}(\theta') = \mathbb{E}_{I',\, D^r,\, t,\, \epsilon} \left\| \epsilon - \epsilon_\phi(z_t, t, I', D^r) \right\|_2^2, \tag{18}$$

where θ' denotes the trainable parameters introduced by LoRA in both the ControlNet and the U-Net. Fig. 5 presents the ablation study results, demonstrating the effectiveness of our ControlNet and scene-adaptive strategy.

Refining Mesh-guided Gaussians. The effectiveness of ControlNet relies on accurate alignment between input depth maps and target views. In early training, incomplete 3D reconstruction leads to unreliable depth maps, which can destabilize training if used for condition. Therefore, we exclude ControlNet in stage 1. After obtaining a structurally coherent Mesh-guided Gaussians representation, we introduce ControlNet and jointly fine-tune it with the diffusion prior, enabling more effective integration of geometric priors and enhancing scene-specific adaptability.

In the second stage, we further refine Mesh-guided Gaussians representation obtained from the first stage. The primary difference lies in the introduction of a scene-adaptive diffusion prior and a depth ControlNet. It is worth noting that the loss function used in this stage remains consistent

Method	Metric	KillingFusion	DeepDeform	iPhone	Avg.	Train time	FPS
NDR	Acc. [cm] ↓	1.27	0.76	3.83	1.97	∼22h	0.06
	Comp. [cm] ↓	1.03	0.65	1.72	1.13		
	Clip sim. ↑	81.82	81.66	75.65	79.71		
	PSNR [dB] ↑	26.69	26.38	25.09	26.05		
MorpheuS	Acc. [cm] ↓	1.00	0.71	1.17	0.96	∼12h	1.01
	Comp. [cm] ↓	0.84	0.57	1.02	0.81		
	Clip sim. ↑	91.57	85.58	82.65	86.60		
	PSNR [dB] ↑	27.09	26.87	25.37	26.44		
Ours	Acc. [cm] ↓	**0.89**	**0.63**	**1.06**	**0.86**	**∼1h**	**35.26**
	Comp. [cm] ↓	**0.77**	**0.48**	**0.97**	**0.74**		
	Clip sim. ↑	**92.12**	**87.08**	**84.69**	**87.96**		
	PSNR [dB] ↑	**32.34**	**27.58**	**28.63**	**29.52**		

Table 1: **Quantitative results on real world datasets.** Our method achieves superior results across all evaluation metrics.

with that of the first stage:

$$\mathcal{L}_{\text{Stage2}} = \mathcal{L}_{\text{Stage1}}. \tag{19}$$

Figure 3: **Qualitative comparison with 4DTAM on the Haru scene from iPhone dataset.** Our method generates high-quality appearance and smoother, more complete surfaces.

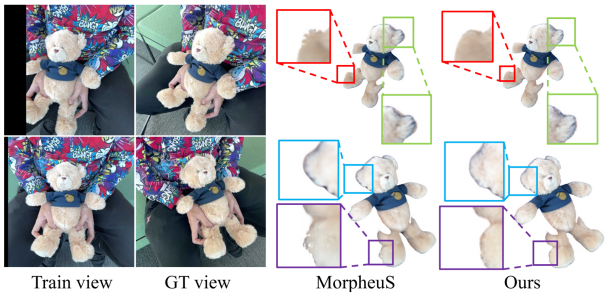

Figure 4: **Qualitative Comparison with MorpheuS on the Teddy scene from iPhone dataset for Novel View Synthesis.** Our method generates images that are visually consistent with the captured GT images.

Experiments

Experimental Setup

Dataset. To comprehensively evaluate MGD, we conduct experiments on three real-world datasets: KillingFusion (Slavcheva et al. 2017), DeepDeform (Bozic et al. 2020), and iPhone (Gao et al. 2022) Datasets. For each dataset, we select three representative scenes following the MorpheuS (Wang, Wang, and Agapito 2024). The videos in these datasets are captured using consumer-grade RGB-D cameras. The training data for our ControlNet model is primarily derived from Objaverse (Deitke et al. 2023), an open-source dataset containing a large number of high-quality and diverse 3D models. In accordance with the data preparation procedure of Zero-1-to-3 (Liu et al. 2023), we randomly sample 12 viewpoints for each 3D model in the dataset and employ a renderer to generate the corresponding images, which are used as training samples.

Metrics. To comprehensively evaluate our method, we assess it from two perspectives: appearance quality and geometric accuracy. For appearance evaluation, we compare the rendered images with the ground-truth images using three standard metrics: PSNR, SSIM, and LPIPS. For geometric evaluation, we follow the evaluation protocol of MorpheuS and adopt two commonly used metrics, Accuracy (Acc.) and Completion (Comp.), which measure the average distance between the reconstructed surface and the ground-truth surface. We also introduce CLIP similarity (Radford et al. 2021) to evaluate the realism of completion.

Implementation details. For the diffusion model, we initialize the U-Net using the pretrained Zero-1-to-3 (Liu et al. 2023). During fine-tuning, we adopt the LoRA approach (Hu et al. 2022) by inserting trainable LoRA modules into the ControlNet and the U-Net. We adopt the minLoRA approach, where the LoRA rank is set to 64 and the learning rate is 10^{-3}. The Mesh-guided Gaussians representation model is trained using a two-stage strategy. We first perform 30K iterations for base training, followed by an additional 20K refinement iterations, which completes in roughly one hour. All experiments are conducted using NVIDIA A100 GPUs.

Quantitative and Qualitative Results

Table 1 reports quantitative comparison results on nine real-world scenes, where our method consistently outperforms MorpheuS and NDR across all evaluation metrics.

Method	Metric	KillingFusion	DeepDeform	iPhone
4DTAM	PSNR ↑	31.13	24.15	27.54
	SSIM ↑	0.934	0.902	0.799
	LPIPS ↓	0.132	0.274	0.265
Ours	PSNR ↑	**32.34**	**27.58**	**28.63**
	SSIM ↑	**0.957**	**0.921**	**0.820**
	LPIPS ↓	**0.114**	**0.198**	**0.201**

Table 2: **Quantitative comparison with 4DTAM.**

Method	Acc. ↓	Clip sim. ↑	PSNR ↑
w/o diffusion prior	1.03	83.07	29.14
w/o depth ControlNet	0.95	86.93	29.25
w/o fine-tuning	0.89	87.29	29.34
w/o GGDM	1.10	87.21	28.96
w/o LGAM	0.92	87.33	28.01
Full model	**0.86**	**87.96**	**29.52**

Table 3: **Ablation studies on three datasets.**

In particular, we achieve significantly higher CLIP similarity scores, indicating stronger realism in the completion results. As shown in Fig. 2, our reconstructions also surpass those of MorpheuS and NDR in terms of geometric completeness and local detail quality. These improvements stem from the proposed Mesh-guided Gaussians representation, the integration of a depth ControlNet into the diffusion prior, and a scene-adaptive fine-tuning strategy. Together, these components enable more realistic and structurally consistent 3D reconstructions while demonstrates superior training and rendering efficiency. In contrast, MorpheuS and NDR employ neural implicit surface modeling, which often fails to recover fine details in complex structures, requires extensive training time, and cannot support real-time rendering. As shown in Fig. 2, although MorpheuS is able to hallucinate unobserved regions, the results exhibit inferior texture-geometry consistency and overly smoothed meshes. Furthermore, since the Teddy scene contains video frames from additional viewpoints, we present a novel-view completion comparison in Fig. 4. The results show that our method maintains high geometric consistency and visual realism even under novel viewpoints, with the generated images closely resembling real captured ones. In contrast, MorpheuS still suffers from noticeable deficiencies in both structural consistency and fine-detail reconstruction.

To further validate the effectiveness of our Mesh-guided Gaussians representation, we compare with 4DTAM (Matsuki, Bae, and Davison 2025) under observed viewpoints, as shown in Table 2. Our method achieves higher PSNR, SSIM, and LPIPS scores. While 4DTAM also adopts Gaussian representations, it lacks global geometric priors and relies on TSDF fusion to extract the mesh, resulting in noisy surfaces and poor structural consistency, as shown in Fig. 3. In contrast, our mesh-guided approach provides a stable global scaffold for Local Gaussians, enabling more coherent and high-fidelity reconstructions.

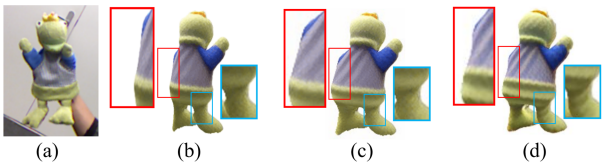

Figure 5: **Ablation study about ControlNet and Fine-turing.** (a) conditioning image for SDS. (b) w/ diffusion prior only. (c) w/o fine-tuning. (d) full model.

Ablation Study

We examine the effectiveness of each individual module. The results, presented in Table 3, indicate that each element significantly contributes to performance, with their absence leading to a decline in results.

Ablation on Mesh-guided Gaussians We first analyze the key modules of Mesh-guided Gaussians representation. As shown in Table 3, ablating GGDM results in a noticeable drop in Acc, indicating that Global Gaussians play a crucial role in driving mesh deformation and improving geometric alignment. In contrast, disabling LGAM leads to a significant decline in PSNR, suggesting that Local Gaussians are essential for capturing fine-grained details and enhancing texture fidelity.

Ablation on Scene-adaptive Diffusion Prior We further investigate the contribution of modules related to the diffusion prior. When the diffusion prior is entirely removed, the CLIP similarity drops significantly, indicating that the model loses its ability to semantically complete unobserved regions. Moreover, the ControlNet and scene-adaptive fine-tuning play key roles in structural recovery and more realistic texture synthesis for unobserved regions. As shown in Fig. 5, the comparison between (b) and (c) demonstrates that ControlNet substantially enhances geometric consistency in unobserved regions. Furthermore, (d) shows that fine-tuning not only improve texture-geometry alignment but also enhances detail expression. It is worth noting that these modules also exhibit positive effects in observed regions, as shown in Table 3. The reconstructed mesh becomes more accurate, and rendering quality is significantly improved. This suggests that the design of the diffusion modules not only improves the generative quality in completed regions but also enhances global consistency and expressiveness.

Conclusion

In this paper, we present MGD, a novel framework for dynamic 3D object reconstruction from monocular RGB-D videos. By introducing a Mesh-guided Gaussians representation and integrating a depth ControlNet into the diffusion prior with scene-adaptive fine-tuning, MGD effectively addresses the inconsistency between multi-view geometry and texture. Extensive experiments demonstrate that our method achieves state-of-the-art performance in both texture fidelity and structural completeness, while significantly improving training and rendering efficiency.

Acknowledgments

The paper is supported by the National Natural Science Foundation of China (No. 62072388), Fujian Provincial Science and Technology Major Project (No. 2024HZ022003), Jiangxi Provincial Natural Science Foundation Key Project (No. 20244BAB28039), Xiamen Public Technology Service Platform (No. 3502Z20231043), and Fujian Sunshine Charity Public Welfare Foundation.

References

Bozic, A.; Zollhofer, M.; Theobalt, C.; and Nießner, M. 2020. Deepdeform: Learning non-rigid rgb-d reconstruction with semi-supervised data.

Cai, H.; Feng, W.; Feng, X.; Wang, Y.; and Zhang, J. 2022. Neural Surface Reconstruction of Dynamic Scenes with Monocular RGB-D Camera.

Chen, R.; Chen, Y.; Jiao, N.; and Jia, K. 2023. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. *arXiv preprint arXiv:2303.13873*.

Deitke, M.; Liu, R.; Wallingford, M.; Ngo, H.; Michel, O.; Kusupati, A.; Fan, A.; Laforte, C.; Voleti, V.; Gadre, S. Y.; et al. 2023. Objaverse-xl: A universe of 10m+ 3d objects. *Advances in Neural Information Processing Systems*, 36: 35799–35813.

Gao, H.; Li, R.; Tulsiani, S.; Russell, B.; and Kanazawa, A. 2022. Monocular dynamic view synthesis: A reality check.

Guédon, A.; and Lepetit, V. 2024. 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*, 5354–5363.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2): 3.

Huang, Y.-H.; Sun, Y.-T.; Yang, Z.; Lyu, X.; Cao, Y.-P.; and Qi, X. 2024. Sc-gs: Sparse-controlled gaussian splatting for editable dynamic scenes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4220–4230.

Kerbl, B.; Kopanas, G.; Leimkühler, T.; and Drettakis, G. 2023. 3D Gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4): 139–1.

Lin, C.-H.; Gao, J.; Tang, L.; Takikawa, T.; Zeng, X.; Huang, X.; Kreis, K.; Fidler, S.; Liu, M.-Y.; and Lin, T.-Y. 2023. Magic3d: High-resolution text-to-3d content creation.

Lin, W.; Zheng, C.; Yong, J.-H.; and Xu, F. 2022. Occlusionfusion: Occlusion-aware motion estimation for real-time dynamic 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1736–1745.

Liu, I.; Su, H.; and Wang, X. 2025. Dynamic gaussians mesh: Consistent mesh reconstruction from dynamic scenes. *ICLR*, 5: 6.

Liu, R.; Wu, R.; Hoorick, B. V.; Tokmakov, P.; Zakharov, S.; and Vondrick, C. 2023. Zero-1-to-3: Zero-shot One Image to 3D Object.

Loper, M.; Mahmood, N.; Romero, J.; Pons-Moll, G.; and Black, M. J. 2023. SMPL: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, 851–866.

Matsuki, H.; Bae, G.; and Davison, A. 2025. 4DTAM: Non-Rigid Tracking and Mapping via Dynamic Surface Gaussians. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Mildenhall, B.; Srinivasan, P. P.; Tancik, M.; Barron, J. T.; Ramamoorthi, R.; and Ng, R. 2020. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis.

Poole, B.; Jain, A.; Barron, J. T.; and Mildenhall, B. 2022. DreamFusion: Text-to-3D using 2D Diffusion.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision.

Ren, J.; Pan, L.; Tang, J.; Zhang, C.; Cao, A.; Zeng, G.; and Liu, Z. 2023. Dreamgaussian4d: Generative 4d gaussian splatting. *arXiv preprint arXiv:2312.17142*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models.

Slavcheva, M.; Baust, M.; Cremers, D.; and Ilic, S. 2017. Killingfusion: Non-rigid 3d reconstruction without correspondences.

Tang, J.; Ren, J.; Zhou, H.; Liu, Z.; and Zeng, G. 2023. Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. *arXiv preprint arXiv:2309.16653*.

Wang, H.; Wang, J.; and Agapito, L. 2024. Morpheus: Neural dynamic 360deg surface reconstruction from monocular rgb-d video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20965–20976.

Wang, Z.; Dou, Z.; Liu, Y.; Lin, C.; Dong, X.; Guo, Y.; Zhang, C.; Li, X.; Wang, W.; and Guo, X. 2025. Wonderhuman: Hallucinating unseen parts in dynamic 3d human reconstruction. *arXiv preprint arXiv:2502.01045*.

Wang, Z.; Lu, C.; Wang, Y.; Bao, F.; Li, C.; Su, H.; and Zhu, J. 2023. ProlificDreamer: High-Fidelity and Diverse Text-to-3D Generation with Variational Score Distillation. *arXiv preprint arXiv:2305.16213*.

Yang, C.; Li, S.; Fang, J.; Liang, R.; Xie, L.; Zhang, X.; Shen, W.; and Tian, Q. 2024. GaussianObject: High-Quality 3D Object Reconstruction from Four Views with Gaussian Splatting. *ACM Transactions on Graphics*.

Yi, T.; Fang, J.; Wang, J.; Wu, G.; Xie, L.; Zhang, X.; Liu, W.; Tian, Q.; and Wang, X. 2024. Gaussiandreamer: Fast generation from text to 3d gaussians by bridging 2d and 3d diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6796–6807.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, 3836–3847.

Zhang, Z.; Cheng, Y.; Pérez-Pellitero, E.; Zhou, Y.; Deng, J.; Chang, H. J.; and Song, J. 2025. Single-view Image to Novel-view Generation for Hand-Object Interactions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 10394–10402.

Zheng, C.; Xue, L.; Zarate, J.; and Song, J. 2025. GauSTAR: Gaussian Surface Tracking and Reconstruction. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 16543–16553.