Single-view Image to Novel-view Generation for Hand-Object Interactions

Zhongqun Zhang[1], Yihua Cheng[1], Eduardo Pérez-Pellitero[2], Yiren Zhou[2], Jiankang Deng[2], Hyung Jin Chang[1], Jifei Song[2*]

[1]University of Birmingham, UK
[2]Huawei, Noah's Ark Lab, UK
zxz064@student.bham.ac.uk, y.cheng.2,h.j.chang@bham.ac.uk,
e.perez.pellitero,zhouyiren,jiankangdeng,jifeisong@huawei.com

Abstract

Hand-object interaction modeling from a single RGB image is a significantly challenging task. Previous works typically reconstruct hand-object interactions as texture-less meshes, ignoring photo-realistic image generation. In this work, we introduce the HO123, a novel method to synthesize novel-view hand-object interaction images from a single image. To this end, we first train a 2D diffusion prior. Given the camera pose in novel views, our approach transfers the camera information into explicit hand representations, including hand depth and skeleton images. We propose a global hand embedding to control the diffusion model based on these hand representations. We then learn a 3D Gaussian splatting for novel-view rendering using the diffusion prior. However, occluded objects present a persistent challenge. To address this issue, we further introduce local hand embedding, where a contact field is defined in the 3D Gaussian Splatting. We leverage contact information to guide the rendering in the contact field. Extensive experiments on the HO3D and DexYCB datasets demonstrate that our method significantly outperforms state-of-the-art novel-view synthesis for hand-object interactions.

Introduction

Humans have the capability to imagine the 3D shapes and appearances of hand-object interactions from single-view images, even if the object is partially occluded by the hand. This ability is increasingly valuable in various practical scenarios, including virtual/augmented reality (Wu et al. 2022; Tendulkar, Surís, and Vondrick 2023), 3D content creation (Tang et al. 2023b,a; Fei et al. 2024), digital humans (Moreau et al. 2023; Zhang et al. 2023), and robotic grasping (Jiang, Hsu, and Zhu 2022; Zhang et al. 2021; Qin et al. 2022). However, reconstructing hand-object interactions in real-world scenarios is challenging and inherently ambiguous due to mutual and self-occlusion. One key insight to solve the ill-posed nature of single-view reconstruction is that the holding hand shape implies the unobserved handheld object shape. Following this idea, current methods (Ye, Gupta, and Tulsiani 2022; Chen et al. 2022) typically use Signed Distance Fields (SDFs) as the geometric representation and conditionally reconstruct the object based on the articulation and the visual input.

*Corresponding author.

Although previous works (Zhang et al. 2024a,b) have advanced the reconstruction of fine hand-object meshes from a single RGB image, they often yield texture-less meshes. To generate photo-realistic renderings in novel view, some methods utilize Neural Radiance Fields (NeRF) (Mildenhall et al. 2020) and 3D Gaussian Splatting (Kerbl et al. 2023) to reconstruct highly detailed shapes and textures of hand-object interactions (Zhang et al. 2024c; Pokhariya and Ishaan 2024) through novel view synthesis. However, these methods rely on multi-view videos, leaving the problem of generating novel-view hand-object interaction renderings from a single-view image unsolved.

Recently, the combination of generative models (Rombach et al. 2022) with NeRF (Mildenhall et al. 2020; Long et al. 2022) has proven to be remarkably effective in novel-view generation from single-view images. Firstly, diffusion-based image generation, trained on large-scale datasets, has enabled approaches to "imagine" unobserved parts of objects by learning to control camera viewpoints (Liu et al. 2024a, 2023, 2024b). Then, Score Distillation Sampling (SDS)(Poole et al. 2022) trains a neural radiance field (Mildenhall et al. 2020) by distilling multi-view priors from pre-trained diffusion models to create 3D representations from a single image. However, existing methods (Liu et al. 2023; Tang et al. 2023a) fail in producing accurate hand shapes and appearances due to the complex articulated structures when applied to hand-object interactions.

To address the aforementioned problems, we propose a novel pipeline, , that leverages 3D Gaussian Splitting with contact and diffusion priors to generate novel-view renderings for hand-object interactions from a single-view image, exploiting the generative capability of pre-trained diffusion models to guide hand-object novel-view image generation, as depicted in Figure 1. Specifically, given an input-view hand-object interaction image, in the first stage utilize the generative capacities of diffusion models to produce novel-view images conditioned on the camera view. In this stage, despite the remarkable generative capabilities demonstrated by diffusion models when conditioned on images, they encounter constraints in generating desired hand visuals. To overcome this, we introduce a global hand embedding module incorporating an additional trainable ControlNet (Zhang, Rao, and Agrawala 2023) to provide depth and skeleton for controlling. We also fine-tune the whole diffusion model us-

ing multi-view hand-object interaction images. In the second stage, we refine our approach by introducing the local hand embedding module, aimed at reconstructing the occluded segments of the object. We posit that these occluded areas should mirror the shape of the hand grasping the object. The proposed local hand embedding module deforms the localization of 3D Gaussians of the object, guiding them toward the corresponding contact points on the hand. After the training, can generate novel-view rendering from a single image in the inference process.

In summary, our main contributions are as follows: **First,** we introduce , the first pipeline using the generative model to synthesize novel-view renderings and reconstruct 3D models for hand-object interactions from a single image. **Second,** we propose the global hand embedding to offer hand structure and object scale guidance during novel-view generation, and the local hand embedding to predict the occluded shapes of the object from the holding hand. **Third,** extensive quantitative and qualitative experiments on real-world datasets HO3D (Hampali et al. 2020) and DexYCB (Chao et al. 2021) demonstrate that our method achieves the best performance on novel-view synthesis and single-view 3D shape reconstruction.

Related Works

Modeling Hand-Object Interactions. Hand-object interaction modeling from single RGB input initially regressing hand MANO (Romero, Tzionas, and Black 2017) parameters and object poses based on predefined object templates (Liu et al. 2021; Hasson et al. 2021; Qi et al. 2024), and subsequently, they proceed to reconstruct object meshes. To optimize the initial pose, some works estimate the hand-object contact (Tse et al. 2022b; Grady et al. 2021) with contact labels and avoid penetration by the physical constraints. To build photo-realistic hand-object modeling, NCRF (Zhang et al. 2024c) propose a dynamic hand-object neural radiance field achieving high-quality hand-object reconstructions and photo-realistic novel view rendering. Our method adopts the MANO model (Romero, Tzionas, and Black 2017) to provide hand prior, but we reconstruct agnostic objects without relying on any category assumption or 3D template.

Hand-held Object Reconstruction. Hand-held object reconstruction from single-view images does not rely on any prior assumptions. Tse et al. (Tse et al. 2022a) developed a collaborative learning network that predicts object mesh vertices and hand MANO parameters. Leng et al. (Leng et al. 2023) explore learning object mesh and hand parameters in hyperbolic space. Additionally, iHOI (Ye, Gupta, and Tulsiani 2022) introduces an approach for conditionally reconstructing objects based on hand articulation and the input image. Building on iHOI, DDF-HO (Zhang et al. 2024a) employs the Directed Distance Field (DDF) as the shape representation, excelling in modeling hand-object interactions. Recently, MOHO (Zhang et al. 2024b) presented a synthetic-to-real framework for hand-held object reconstruction, demonstrating that 2D multi-view supervision outperforms 3D supervision. Our method also achieves high-quality reconstruction of hand-held object meshes.

Diffusion-based Image-to-3D Generation. The goal of image-to-3D generation is to create 3D models from a single-view image. Recently, Zero-1-to-3 (Liu et al. 2023) trained a large-scale 2D diffusion model conditioned on camera views, enabling zero-shot image-conditioned novel view synthesis through learned diffusion priors. To reconstruct 3D models from diffusion priors, most approaches (Liu et al. 2023, 2024a; Tang et al. 2023a; Liu et al. 2024b) follow an optimization pipeline that updates 3D representations (e.g., NeRF (Mildenhall et al. 2020), Gaussian Splatting (Tang et al. 2023a)) via neural rendering and a score distillation sampling (SDS) (Poole et al. 2022) loss. DreamGaussian (Tang et al. 2023a) employs Gaussian Splatting to reduce optimization time to 2 minutes with minimal quality loss. However, these methods struggle to produce high-quality 3D models of human hands due to their complex architecture. Our method follows the diffusion-based image-to-3D generation pipeline but introduces unique modules that enhance robustness in handling hand-object interactions.

Methodology

Preliminaries

View-Conditioned Diffusion. Given an input-view image $I \in \mathbb{R}^{H\times W\times 3}$ of an object, our goal is to synthesize an image $\hat{I} \in \mathbb{R}^{H\times W\times 3}$ of the object from a target camera viewpoint. Zero123 (Liu et al. 2023) leverage a latent diffusion model (LDM) (Rombach et al. 2022) to learn distribution conditioned on the relative camera rotation and translation ΔV of the desired viewpoint:

$$\hat{I}_{\Delta V} = f(I, \Delta V). \tag{1}$$

Zero123 uses a spherical camera and assumes the object is at the center of the origin of the coordinate. The camera pose is represented as $V = (\theta, \phi, r)$, where denote the polar angle, azimuth angle, and radius respectively. Thus the relative camera transformation between the input view and the target view is parameterized as $\Delta V = (\theta_2 - \theta_1, \phi_2 - \phi_1, r_2 - r_1)$.

3D Gaussian Splatting. 3DGS (Tang et al. 2023a) represents a static object as a set of 3D Gaussians (3D anisotropic balls) primitives G, which can be rendered in real-time via differentiable rasterization. Each 3D Gaussian is defined by its mean x, covariance Σ, opacity α, and spherical harmonics coefficients sh. During 2D rendering, the projected covariance matrix is calculated as:

$$\Sigma_i^{\mathrm{2D}} = JW\Sigma_i W^{\mathrm{T}} J^{\mathrm{T}}, \tag{2}$$

where J is the Jacobian of the affine approximation of the projective transformation, and W is the world-to-camera matrix. To simplify learning, $\mathbf{\Sigma}$ is decomposed into a quaternion r for rotation and a 3D-vector s for scaling, yielding rotation matrix R and scaling matrix S. Hence, the final i^{th} 3D Gaussian is represented as $\{x_i, r_i, s_i, o_i, sh_i\}$.

HO123: Overview

Given a single RGB image I containing hand-object interaction, HO123 aims at generating a novel-view rendering $\hat{I}$

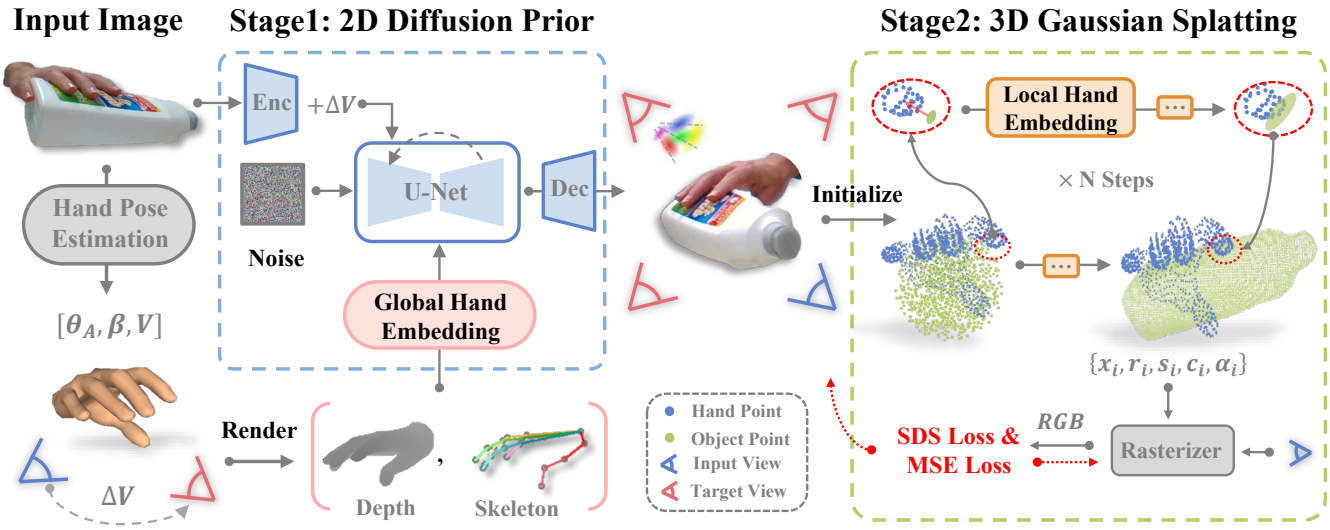

Figure 1: **Method Overview.** HO123 is a two-stage pipeline composed of 2D diffusion prior and 3D Gaussian Splatting. In the first stage, we train a 2D diffusion prior to generate the target-view image from the input-view image, conditioned on the hand depth and skeleton map. We leverage the generative power of a pre-trained diffusion model and fine-tune it with hand-object interaction data (Section). In the second stage, we train 3D Gaussian Splatting with SDS loss to lift the generated 2D images to 3D representation (Section). We introduce the Global Hand Embedding module and Local Hand Embedding modules to intricately incorporate hand priors into the pipeline.

conditioned on the relative camera pose ΔV and reconstructing 3D shape O of the object from the generated multi-view images. We perform data pre-processing in the input image I to acquire hand articulation $(\theta_A, \boldsymbol{\beta})$, and camera pose of the input image V.

In detail, we follow the approach of (Ye, Gupta, and Tulsiani 2022) by using an off-the-shelf hand pose estimator (Rong, Shiratori, and Joo 2020) to estimate the hand articulation $(\theta_A, \boldsymbol{\beta})$ and camera pose V from an input image I, where θ_A is defined in the parametric MANO model (Romero, Tzionas, and Black 2017) with 45D hand pose parameters and β consists of 10D hand shape parameters, V denotes the 6D camera pose, consist of rotation $R \in \mathbb{SO}(3)$ and translation $t \in \mathbb{R}^3$. The camera pose is relative to the hand coordinate system. To fine-tune the pre-trained Zero123 model, we convert R and t to the spherical coordinate (θ, ϕ, r).

Our pipeline comprises two modules and is summarized in Figure 1. In the first stage, we train a 2D diffusion model to generate the target-view rendering from the input-view image, conditional on the hand depth and skeleton map, where we leverage the generative power of a pre-trained diffusion model and fine-tune with hand-object interaction data. In the second stage, we train 3D Gaussian Splatting with SDS loss to lift the generated 2D images to 3D representation.

2D Diffusion Prior: Multi-view Generation

Multi-view Diffusion Model. Given an input image I with the hand pose $(\boldsymbol{\theta}_A, \boldsymbol{\beta})$, the objective of Stage 1 is to generate the image under the target view ΔV. As shown in Figure 1, we start with a pre-trained multi-view diffusion model, *e.g.*, Zero123, and fine-tune using hand-object interaction data. Firstly, a latent code z is extracted by a VAE autoencoder $\mathcal{E}$ from the target-view GT image. During the training, we randomly add sampled noise to the latent code z. Secondly, we utilize the pre-trained CLIP (Radford et al. 2021) image encoder to extract feature embedding from the input hand-object image. The image feature is concatenated with target view $\Delta \boldsymbol{V}$ to form a View-CLIP embedding. We apply cross-attention to enforce this embedding to the denoising U-Net as a condition. A latent code $\tilde{z}$ is obtained by iterative denoising procedure. In the end, the final image is reconstructed through the VAE decoder $\tilde{I} = \mathcal{D}(\tilde{z})$.

Global Hand Embedding Module. Our goal is to learn additional hand-conditioned distribution to enhance the performance of the multi-view diffusion model for our task. To retain the structures and parameters from the multi-view diffusion model as much as possible, as shown in Figure 2, we propose an attention-based hand pose embedding module and additional ControlNet-based network for hand shape embedding. For the hand shape embedding module, we propose to use a hand-depth map $I_{dp}^{H} \in \mathbb{R}^{H \times W \times 3}$ from the target view as an additional condition. The goal of this module is to provide extra information to guide the output image, ensuring it generates a hand shape similar to the one in the depth image. Following ControlNet (Zhang, Rao, and Agrawala 2023), the local hand embedding module replicates the encoder of the multi-view diffusion model (Rombach et al. 2022) as a trainable side path and incorporates additional zero convolution layers. The extra conditions outputted from the zero convolution layers are then added to the skip connections of the multi-view diffusion UNets. As to the hand pose embedding module, we propose to use a hand skeleton map

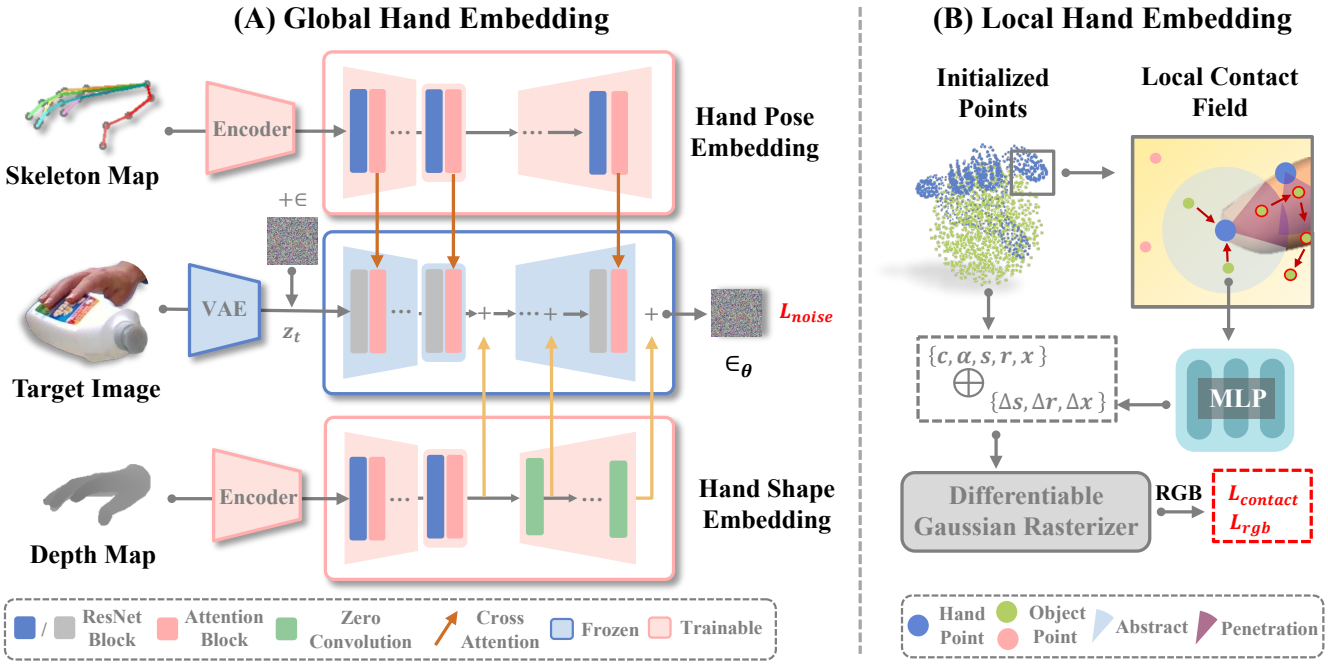

Figure 2: **Training of Hand Embedding Modules. (A) Global hand embedding:** This module enhances the performance of the multi-view diffusion model for our task by learning additional hand-conditional distribution. We employ a pre-trained diffusion model and ControlNet-like architecture to enable hand pose and shape conditioning. To train our model, we render training pairs of hand depth maps and hand skeleton maps. We train the hand pose and shape embedding model and attention block of the diffusion model while keeping other parameters frozen. **(B) Local hand embedding:** We separately initialize the 3D Gaussians for hand and object. The position of the hand Gaussians is fixed. To reconstruct the shape of the parts of the object occluded by the hand, we designed a contact-guided 3D Gaussian deformation strategy. It penalized the penetration region's 3D Gaussians to move far from the hand, while the attraction region will be encouraged to move closer to the hand.

$I^H_{skl} \in \mathbb{R}^{H\times W\times 3}$ to provide scale information to the object since the scale of the hand-held object is constrained by the hand. We observed that using the same network for hand shape embedding could cause the object generation to be affected by the hand shape. To extract scale information, we opt to replicate the UNet of the multi-view diffusion model to serve as the network for the hand pose embedding module. The outputs from the attention block are then integrated into the attention block of the multi-view diffusion UNets through cross-attention operations.

Training. We first pre-train the hand pose embedding module. The objective function of the pre-train is:

$$\mathcal{L}_{noise} = \min_\theta \mathbb{E}_{\mathcal{E}(I),P_\theta\left(I^H_{skl}\right),t,\epsilon\sim\mathcal{N}(0,\mathbf{I})} \left\|\epsilon - \epsilon_\theta\left(z_t, t, P_\theta\left(I^H_{skl}\right)\right)\right\|_2^2, \quad (3)$$

where P_θ is the hand pose embedding module, ϵ_θ is the predicted noise. We then train the hand shape embedding module by utilizing the fine-tuning strategy (Zhang, Rao, and Agrawala 2023) to optimize cross-attention layers of the multi-view diffusion and ControlNet parameters while keeping most of the other parameters frozen. The objective function can be represented as:

$$\mathcal{L}_{noise} = \min_\theta \mathbb{E}_{\mathcal{E}(I),t,\epsilon\sim\mathcal{N}(0,\mathbf{I})} \left\|\epsilon - \epsilon_\theta\left(z_t, t, I, I^H_{dp}\right)\right\|_2^2. \quad (4)$$

3D Gaussian Splatting: Novel-view Rendering

We represent 3D Gaussian as $\{\mathbf{x}_i, \mathbf{s}_i, \mathbf{r}_i, \alpha_i, \mathbf{c}_i\}$, where $\mathbf{x}_i$ is the location of each Gaussian, $\mathbf{s}_i$ is a scaling factor, and $\mathbf{r}_i$ describes the rotation quaternion. Following DreamGaussian (Tang et al. 2023a), we also store an opacity value α and a color feature c for volumetric rendering. Spherical harmonics are disabled since we only want to model diffuse colors for simplicity. We use a differentiable Gaussian rasterizer to splat a set of 3D Gaussians into the image plane. Volumetric rendering is then performed for each pixel in front-to-back depth order to evaluate the final color and alpha.

We separately initialize the 3D Gaussians for hand and object. For hand, we initialize 3D Gaussians using the estimated MANO point cloud, while for the object, we initialize the 3D Gaussians with random positions sampled inside a sphere, as shown in Figure 2. We only initialize the positions with unit scaling and no rotation. The hand 3D Gaussians positions are fixed, while the object 3D Gaussians are dynamic and periodically densified during optimization.

Local hand embedding Module. To reconstruct the shape of the parts of the object occluded by the hand, we designed a contact-guided 3D Gaussian deformation strategy. Specifically, we define a local contact field with a radius of 2mm for each Gaussian point of the hand. If the object's 3D Gaussian

falls within this contact field, we consider it to be in contact with the hand. As shown in Figure 2, we further divide the contact field into a penetration region (purple) and an attraction region (light blue). The 3D Gaussians in the penetration region will be penalized to move away from this area, while the 3D Gaussians in the attraction region will be encouraged to move closer to the hand Gaussians. To achieve this, we compute per-Gaussian residual outputs learned by an MLP that can translate and rotate the 3D object Gaussians.

Training. To lift 2D priors to 3D representation, we use SDS (Poole et al. 2022) loss to optimize the 3D Gaussians. The SDS loss can be formulated as:

$$\nabla_{\Theta}\mathcal{L}_{\text{SDS}} = \mathbb{E}_{t,v,\epsilon}\left[\left(\epsilon_{\phi}\left(I^{v}_{\text{RGB}}; t, \tilde{I}^{r}_{\text{RGB}}, \Delta v\right) - \epsilon\right)\frac{\partial I^{v}_{\text{RGB}}}{\partial \Theta}\right] \tag{5}$$

where $\epsilon_{\phi}(\cdot)$ is the predicted noise by the 2D diffusion prior ϕ, and Δv is the relative camera pose compared to the reference camera r. Additionally, we optimize the reference view image I^{r}_{RGB} with the input using photo-metric loss. The local hand embedding module is also regularized with the contact guidance, implied by the combination of penetration and abstract losses. The penetration loss is defined as below:

$$\mathcal{L}_{\text{pen}} = \sum_{x \in O} \max\left(0, \left(\mathbf{x}^{O}_{i} - \mathbf{x}^{H}_{j}\right) \cdot \mathbf{n}^{H}_{i} - c_{pen}\right) \tag{6}$$

, where x is the position of 3D Gaussian and $\mathbf{n}^{H}_{i}$ is hand normal, and c_{pen}=2mm. The abstract loss is $\mathcal{L}_{\text{abs}} = \sum_{x \in O} \min\left(\mathbf{x}^{O}_{i} - \mathbf{x}^{H}_{j}\right)$. We only apply the contact loss to the 3D Gaussians detected to be in contact. The final loss is the weighted sum of all the above losses.

Experiments

Experimental Setup

Datasets. We conduct experiments on two real-world datasets: HO3D (Hampali et al. 2020) and DexYCB (Chao et al. 2021). Both datasets capture dynamic hand-object interactions from multiple views and provide comprehensive annotations. HO3D is captured with 5 cameras, while DexYCB is captured with 8 cameras. HO3D consists of 77,558 images from 68 sequences, featuring 10 different individuals manipulating 10 different objects. One sequence per object is recorded as a single-view video, while others are multi-view videos. We follow iHOI (Ye, Gupta, and Tulsiani 2022) to split training and testing sets. DexYCB is one of the largest real-world hand-object video datasets, and we focus on right-hand samples using the official s0 split. Following gSDF (Chen et al. 2023), we use 29,656 training samples and 5,928 testing samples. Notably, our method is only trained on the HO3D dataset and utilizes only the RGB images, segmentations, and poses for training, without requiring 3D ground-truth meshes.

Baseline. For novel-view synthesis baselines, we include HandNeRF (Choi et al. 2024) and MOHO (Zhang et al. 2024b), which are designed for hand-object interactions, and pixelNeRF (Yu et al. 2021), which is for general scenes from a single image. Zero123 (Liu et al. 2023) and Zero123++(Shi et al. 2023) represent the state-of-the-art in image-to-3D generation methods. We use the Zero123*, which is a fine-tuned model on the HO3D dataset. For shape reconstruction baselines, we consider HO(Hasson et al. 2019), iHOI (Ye, Gupta, and Tulsiani 2022), DDF-HO (Zhang et al. 2024a), and MOHO (Zhang et al. 2024b) from the hand-object interaction field, along with DreamGaussian (Tang et al. 2023a), which employs generative Gaussian Splatting with Zero123. We primarily compare the reconstructed object meshes with these methods to demonstrate our method's single-view reconstruction capability.

Evaluation Metrics. To evaluate the novel view synthesis quality, we use three metrics by comparing with the ground truth images: we report PSNR, SSIM (Wang et al. 2004), and LPIPS* (Zhang et al. 2018) (LPIPS* = LPIPS $\times 10^{3}$). For the shape reconstruction metrics, we follow (Ye, Gupta, and Tulsiani 2022) to uniformly sample 30,000 points on the reconstructed mesh, and report mean Chamfer Distance (CD, mm) and F-score at thresholds of 5mm (F-5) and 10mm (F-10).

Implementation Details. For the diffusion part, we initialize the diffusion U-Net's weights using the Zero123XL (Liu et al. 2023) model. We utilize a batch size of 30 images and an AdamW (Kingma and Ba 2015) optimizer with a learning rate of 10^{-4}, incorporating a constant warmup scheduling. We finetune our model on the HO3D (Hampali et al. 2020) dataset selecting 110,00 image pairs. Training the diffusion model takes about 10 hours on 8 NVIDIA A100 GPUs for 30k steps. We use the same offline systems (Rong, Shiratori, and Joo 2020) as (Ye, Gupta, and Tulsiani 2022) to estimate the hand and camera poses. During testing, our method takes 2 seconds on a single A100 GPU to generate one image. For the 3D Gaussian Splatting part, we train 500 steps for SDS loss. The 3D Gaussians are initialized with an opacity of 0.1 and a grey color inside a sphere with a radius of 0.2. We sample random camera poses at a fixed radius of 1, with a y-axis field of view (FOV) of 49 degrees. The azimuth ranges from -180 to 180 degrees and the elevation ranges from -30 to 30 degrees. The Gaussian Splatting takes 1 min to create a 3D model. We extract mesh from 3D Gaussians following (Tang et al. 2023a).

Evaluation on Novel View Synthesis

Quantitative Results. Given a single-view image for hand-object interaction, we generate novel views and make comparisons on rendering quality. From Quantitative results shown in Table 1, our method significantly outperforms related works. HandNeRF (Choi et al. 2024)'s motivation is similar to ours, which uses the hand shape to constrain the object geometry. Our method leads by 10.72 on PSNR, 0.23 on SSIM, and 0.15 on LPIPS against the handNeRF, demonstrating the generative model has a more realistic image quality than neural rendering. Furthermore, HO123 and HO123++, like our method, are based on generative models, but they perform poorly in hand-object interactions. Notely, we report the result for HO123 using our fine-tuning model on HO3D dataset. However, its performance still shows no significant improvement, because directly fine-tuning parameters of diffusion

Method	HO3D Dataset			DexYCB Dataset		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
iHOINeRF (Ye, Gupta, and Tulsiani 2022)	19.40	0.68	0.210	19.82	0.64	0.270
MonoNHR (Choi et al. 2022)	19.34	0.70	0.220	19.66	0.66	0.290
HandNeRF (Choi et al. 2024)	20.54	0.74	0.180	21.66	0.70	0.240
MOHO (Zhang et al. 2024b)	26.01	0.96	0.049	35.80	**0.99**	0.013
Zero123++ (Shi et al. 2023)	17.20	0.78	0.130	22.06	0.75	0.210
Zero123* (Liu et al. 2023)	20.15	0.81	0.100	23.93	0.79	0.190
HO123 (Ours)	**31.26**	**0.97**	**0.037**	**36.40**	**0.99**	**0.011**

Table 1: Quantitative results of novel view synthesis on HO3D and DexYCB dataset.

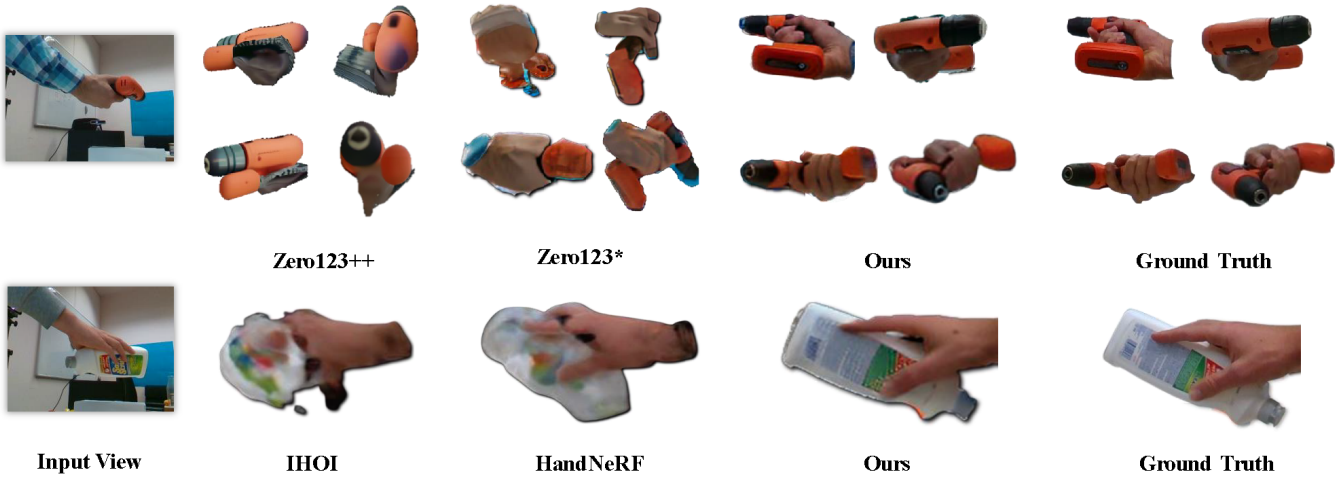

Figure 3: **Qualitative comparison of novel-view image synthesis.** Zero123* (Liu et al. 2023) is a fine-tuned Zero123 model on HO3D dataset. The visualizations of IHOI and HandNeRF is from the original paper (Choi et al. 2024).

model requires very large data. In contrast, our method trains an additional control network while preserving the multi-view generation capability of the original model as much as possible. Experimental results demonstrate that our approach significantly improves performance in hand-object interactions. MOHO (Zhang et al. 2024b) only focus on object's novel view rendering. It cannot render appearance information for hand, so we only report their object render quality. Despite the higher difficulty of rendering hands compared to rigid objects, our method still outperforms MOHO across various metrics.

Qualitative Results. We also present qualitative comparisons in Figure 3. Zero123++ can generate coarse multi-view images for the visible part of the object. However, it struggles to produce accurate hand structures, and the occluded part of the object is often missing. This highlights the challenges posed by severe occlusion by hands in the multi-view diffusion model. By fine-tuning Zero123 on our task using hand-object interaction data, we have significantly improved performance. We observed that the original Zero123 assumes the object is centrally located and static. In our scenario, where the hand and object move freely in 6 degrees of freedom (DoF), using spherical coordinates alone is insufficient to control the viewpoint accurately, as depicted in Figure 3. In our approach, we incorporate hand depth as a condition, which provides a strong prior for the hand position. By combining spherical coordinates and hand depth, we achieve precise viewpoint control for hand-object interactions, leading to the most satisfying results. More visualizations are provided in the Supplementary.

Evaluation on Shape Reconstruction

We compared our method with state-of-the-art single-view hand-held object reconstruction methods, including HO (Hasson et al. 2019), IHOI (Ye, Gupta, and Tulsiani 2022), DDF-HO (Zhang et al. 2024a), and MOHO (Zhang et al. 2024b). Additionally, we compared DreamGaussian (Tang et al. 2023a), which also utilizes Zero123 and Gaussian Splatting to generate 3D assets from a single image. The quantitative results in Table 2 demonstrate that methods incorporating a hand prior outperform DreamGaussian. DreamGaussian utilizes multi-view diffusion as a 2D prior, which has previously shown poor performance in hand-object interaction scenes. Our method significantly outperforms traditional reconstruction methods. For instance, IHOI relies on Signed Distance Function (SDF) to learn the representation of the object. Our method outperforms IHOI by 1.08 on Chamfer Distance, 0.09 on F-5, and 0.11 on F-10. MOHO trains its model with multi-view videos and A-model masks, achieving the cur-

Method	HO3D Dataset			DexYCB Dataset		
	CD↓	F-5↑	F-10↑	CD↓	F-5↑	F-10↑
HO (Hasson et al. 2019)	4.19	0.11	0.22	0.42	0.38	0.64
IHOI (Ye, Gupta, and Tulsiani 2022)	1.53	0.28	0.50	-	-	-
DDF-HO (Zhang et al. 2024a)	0.55	0.28	0.42	-	-	-
MOHO (Zhang et al. 2024b)	0.91	0.31	0.50	0.15	0.60	0.81
DreamGaussian (Tang et al. 2023a)	7.36	0.05	0.18	0.63	0.19	0.48
HO123 (Ours)	**0.45**	**0.37**	**0.61**	**0.12**	**0.69**	**0.87**

Table 2: Quantitative results of shape reconstruction on HO3D and DexYCB dataset.

GPE	GSE	LHE	Novel View Synthesis			Reconstruction		
			PSNR↑	SSIM↑	LPIPS↓	CD↓	F-5↑	F-10↑
×	×	×	18.23	0.69	0.190	7.56	0.05	0.18
✓	×	×	27.35	0.95	0.043	0.53	0.24	0.45
×	✓	×	28.36	0.95	0.041	0.56	0.27	0.48
✓	✓	×	-	-	-	2.46	0.18	0.25
✓	✓	✓	**31.26**	**0.97**	**0.037**	**0.45**	**0.37**	**0.61**

Table 3: Ablation studies of each component of our method on the HO3D dataset.

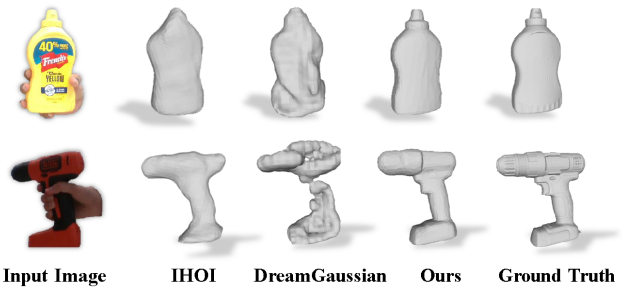

Figure 4: **Qualitative comparison of object shape reconstruction.** Compared with other methods, we can produce more complete and detailed reconstruction results. The visualization of IHOI is from (Jiang et al. 2024).

rent state-of-the-art. Compared with MOHO, our method relies on pre-trained 2D diffusion priors, resulting in better generalization.

We visualize the reconstructed objects in Figure 4. SDF-based method IHOI (Ye, Gupta, and Tulsiani 2022) can predict the coarse shape of the object, but it typically loses the finer details of the object surface. Hand-induced occlusion significantly decreases the reconstruction quality of DreamGaussian, which leads to incomplete surface reconstructions.

Ablation Study

We conduct ablation studies on the HO3D datasets to evaluate the impact of three key assets of our method: the global hand pose embedding module (GPE), the global hand shape embedding module (GSE), and the local hand embedding module (LHE). The results of the ablation studies are presented in Table 3.

We first ablate the global shape embedding, resulting in a degraded performance with a 3.91 decrease in the PSNR metric compared to the full model. Similarly, ablating the global pose embedding leads to a PSNR decrease of around 2.90. This indicates that local hand embedding is more critical for novel view synthesis. When we ablate the global hand embedding, the reconstruction performance also decreases, with a 0.13 drop in the F-10 metric. This illustrates that the quality of the reconstruction depends on the effectiveness of the 2D diffusion prior.

Next, we ablate the local hand embedding, resulting in a 2.01 increase in Chamfer Distance and a 0.56 decrease in the F-5 metric. This indicates that considering the local contact information of the hand can improve the accuracy of hand-held object reconstruction. This module is crucial for recovering the occluded parts of the object. It is specifically related to 3D reconstruction and does not affect novel view synthesis.

Conclusion

In this paper, we propose HO123, a novel pipeline that leverages a large-scale diffusion model as well as 3D Gaussian Splatting to achieve novel-view image generation for hand-object interactions. To this end, we propose a global hand-embedding module to guide prior learning within the diffusion model and a local hand-embedding module to guide free-viewpoint rendering within gaussian splatting. Our approach supports novel arbitrary view synthesis for hand-object interactions as well as object mesh reconstruction. Extensive experiment results on the HO3D and DexYCB datasets demonstrate that our method significantly outperforms state-of-the-art models for novel-view synthesis of hand-object interactions. We expect our work will advance the application of generative AI in 3D content creation.

Limitations. Our method focuses on static hand-object interaction modeling. In the future, we will consider learning an animatable model driven by the hand-object poses.

Acknowledgments

This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(IITP-2024-RS-2024-00437102) supervised by the IITP(Institute for Information & Communications Technology Planning & Evaluation), China Scholarship Council (CSC) Grant No. 202208060266.

References

Chao, Y.-W.; Yang, W.; Xiang, Y.; Molchanov, P.; Handa, A.; Tremblay, J.; Narang, Y. S.; Van Wyk, K.; Iqbal, U.; Birchfield, S.; et al. 2021. DexYCB: A Benchmark for Capturing Hand Grasping of Objects. In *CVPR*.

Chen, Z.; Chen, S.; Schmid, C.; and Laptev, I. 2023. gsdf: Geometry-driven signed distance functions for 3d hand-object reconstruction. In *CVPR*.

Chen, Z.; Hasson, Y.; Schmid, C.; and Laptev, I. 2022. AlignSDF: Pose-Aligned Signed Distance Fields for Hand-Object Reconstruction. In *ECCV*.

Choi, H.; Chavan-Dafle, N.; Yuan, J.; Isler, V.; and Park, H. 2024. HandNeRF: Learning to Reconstruct Hand-Object Interaction Scene from a Single RGB Image. *ICRA*.

Choi, H.; Moon, G.; Armando, M.; Leroy, V.; Lee, K. M.; and Rogez, G. 2022. Mononhr: Monocular neural human renderer. In *3DV*.

Fei, B.; Xu, J.; Zhang, R.; Zhou, Q.; Yang, W.; and He, Y. 2024. 3D Gaussian as a New Vision Era: A Survey. *arXiv*.

Grady, P.; Tang, C.; Twigg, C. D.; Vo, M.; Brahmbhatt, S.; and Kemp, C. C. 2021. ContactOpt: Optimizing Contact to Improve Grasps. In *CVPR*.

Hampali, S.; Rad, M.; Oberweger, M.; and Lepetit, V. 2020. Honnotate: A method for 3D annotation of hand and object poses. In *CVPR*.

Hasson, Y.; Varol, G.; Laptev, I.; and Schmid, C. 2021. Towards unconstrained joint hand-object reconstruction from RGB videos. In *3DV*.

Hasson, Y.; Varol, G.; Tzionas, D.; Kalevatykh, I.; Black, M. J.; Laptev, I.; and Schmid, C. 2019. Learning joint reconstruction of hands and manipulated objects. In *CVPR*.

Jiang, S.; Ye, Q.; Xie, R.; Huo, Y.; Li, X.; Zhou, Y.; and Chen, J. 2024. In-Hand 3D Object Reconstruction from a Monocular RGB Video. In *AAAI*.

Jiang, Z.; Hsu, C.-C.; and Zhu, Y. 2022. Ditto: Building digital twins of articulated objects from interaction. In *CVPR*.

Kerbl, B.; Kopanas, G.; Leimkühler, T.; and Drettakis, G. 2023. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*.

Kingma, D. P.; and Ba, J. 2015. Adam: A method for stochastic optimization. In *ICLR*.

Leng, Z.; Wu, S.-C.; Saleh, M.; Montanaro, A.; Yu, H.; Wang, Y.; Navab, N.; Liang, X.; and Tombari, F. 2023. Dynamic hyperbolic attention network for fine hand-object reconstruction. In *ICCV*.

Liu, M.; Xu, C.; Jin, H.; Chen, L.; Varma T, M.; Xu, Z.; and Su, H. 2024a. One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization. *NeurIPS*.

Liu, R.; Wu, R.; Van Hoorick, B.; Tokmakov, P.; Zakharov, S.; and Vondrick, C. 2023. Zero-1-to-3: Zero-shot one image to 3d object. In *ICCV*.

Liu, S.; Jiang, H.; Xu, J.; Liu, S.; and Wang, X. 2021. Semi-supervised 3D hand-object poses estimation with interactions in time. In *CVPR*.

Liu, Y.; Lin, C.; Zeng, Z.; Long, X.; Liu, L.; Komura, T.; and Wang, W. 2024b. Syncdreamer: Generating multiview-consistent images from a single-view image. *ICLR*.

Long, X.; Lin, C.; Wang, P.; Komura, T.; and Wang, W. 2022. Sparseneus: Fast generalizable neural surface reconstruction from sparse views. In *ECCV*.

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. In *ECCV*.

Moreau, A.; Song, J.; Dhamo, H.; Shaw, R.; Zhou, Y.; and Pérez-Pellitero, E. 2023. Human gaussian splatting: Real-time rendering of animatable avatars. *CVPR*.

Pokhariya, C.; and Ishaan, N. 2024. Manus: Markerless hand-object grasp capture using articulated 3d gaussians. In *CVPR*.

Poole, B.; Jain, A.; Barron, J. T.; and Mildenhall, B. 2022. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv*.

Qi, H.; Zhao, C.; Salzmann, M.; and Mathis, A. 2024. HOISDF: Constraining 3D Hand-Object Pose Estimation with Global Signed Distance Fields. *CVPR*.

Qin, Y.; Wu, Y.-H.; Liu, S.; Jiang, H.; Yang, R.; Fu, Y.; and Wang, X. 2022. Dexmv: Imitation learning for dexterous manipulation from human videos. In *ECCV*, 570–587.

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. In *ICML*.

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

Romero, J.; Tzionas, D.; and Black, M. J. 2017. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics (ToG)*.

Rong, Y.; Shiratori, T.; and Joo, H. 2020. Frankmocap: Fast monocular 3d hand and body motion capture by regression and integration. *ICCVW*.

Shi, R.; Chen, H.; Zhang, Z.; Liu, M.; Xu, C.; Wei, X.; Chen, L.; Zeng, C.; and Su, H. 2023. Zero123++: a single image to consistent multi-view diffusion base model. *arXiv*.

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

Tang, J.; Wang, T.; Zhang, B.; Zhang, T.; Yi, R.; Ma, L.; and Chen, D. 2023b. Make-it-3d: High-fidelity 3d creation from a single image with diffusion prior. In *Proceedings of the*

IEEE/CVF International Conference on Computer Vision, 22819–22829.

Tendulkar, P.; Surís, D.; and Vondrick, C. 2023. FLEX: Full-Body Grasping Without Full-Body Grasps. In *CVPR*.

Tse, T. H. E.; Kim, K. I.; Leonardis, A.; and Chang, H. J. 2022a. Collaborative learning for hand and object reconstruction with attention-guided graph convolution. In *CVPR*.

Tse, T. H. E.; Zhang, Z.; Kim, K. I.; Leonardis, A.; Zheng, F.; and Chang, H. J. 2022b. S2Contact: Graph-Based Network for 3D Hand-Object Contact Estimation with Semi-supervised Learning. In *ECCV*.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. In *TIP*.

Wu, Y.; Wang, J.; Zhang, Y.; Zhang, S.; Hilliges, O.; Yu, F.; and Tang, S. 2022. Saga: Stochastic whole-body grasping with contact. In *ECCV*.

Ye, Y.; Gupta, A.; and Tulsiani, S. 2022. What's in your hands? 3d reconstruction of generic objects in hands. In *CVPR*.

Yu, A.; Ye, V.; Tancik, M.; and Kanazawa, A. 2021. pixelnerf: Neural radiance fields from one or few images. In *CVPR*.

Zhang, C.; Di, Y.; Zhang, R.; Zhai, G.; Manhardt, F.; Tombari, F.; and Ji, X. 2024a. DDF-HO: Hand-Held Object Reconstruction via Conditional Directed Distance Field. *NeurIPS*.

Zhang, C.; Jiao, G.; Di, Y.; Huang, Z.; Wang, G.; Zhang, R.; Fu, B.; Tombari, F.; and Ji, X. 2024b. MOHO: Learning Single-view Hand-held Object Reconstruction with Multi-view Occlusion-Aware Supervision. *CVPR*.

Zhang, H.; Ye, Y.; Shiratori, T.; and Komura, T. 2021. Manipnet: neural manipulation synthesis with a hand-object spatial representation. *ACM Transactions on Graphics (ToG)*.

Zhang, J.; Luo, H.; Yang, H.; Xu, X.; Wu, Q.; Shi, Y.; Yu, J.; Xu, L.; and Wang, J. 2023. NeuralDome: A neural modeling pipeline on multi-view human-object interactions. In *CVPR*.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. In *ICCV*.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*.

Zhang, Z.; Song, J.; Pérez-Pellitero, E.; Zhou, Y.; Chang, H. J.; and Leonardis, A. 2024c. NCRF: Neural Contact Radiance Fields for Free-Viewpoint Rendering of Hand-Object Interaction. *3DV*.