Human 3Diffusion: Realistic Avatar Creation via Explicit 3D Consistent Diffusion Models

Yuxuan Xue[1,2] **Xianghui Xie**[1,2,3] **Riccardo Marin**[1,2] **Gerard Pons-Moll**[1,2,3]

[1]University of Tübingen [2] Tübingen AI Center
[3]Max Planck Institute for Informatics, Saarland Informatics Campus
https://yuxuan-xue.com/human-3diffusion/

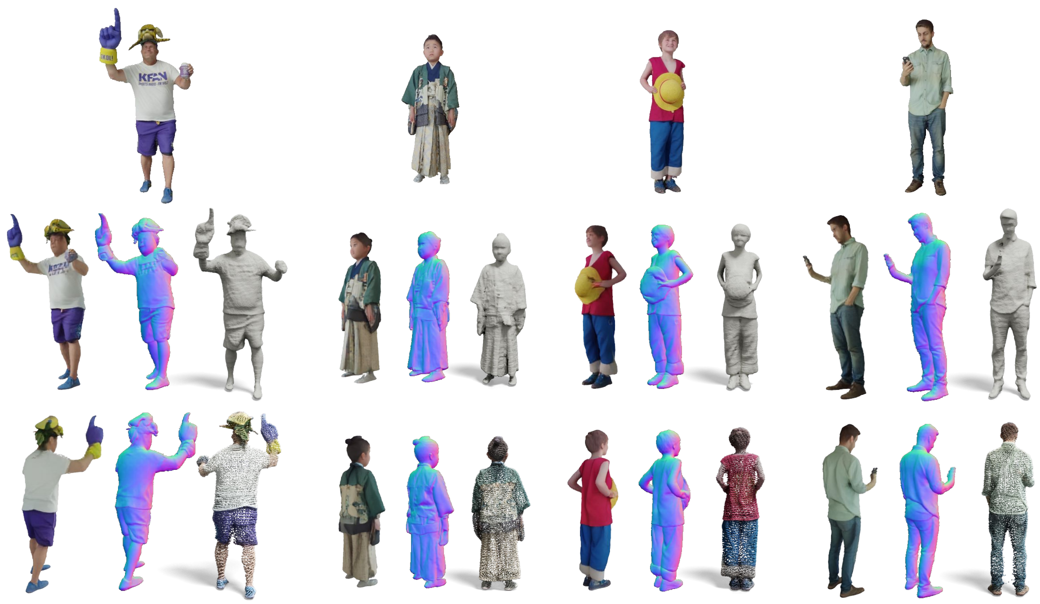


Figure 1: Given a single image of a person (top), our method **Human 3Diffusion** creates 3D Gaussian Splats of realistic avatars with high-fidelity geometry and texture.

Abstract

Creating realistic avatars from a single RGB image is an attractive yet challenging problem. Due to its ill-posed nature, recent works leverage powerful prior from 2D diffusion models pretrained on large datasets. Although 2D diffusion models demonstrate strong generalization capability, they cannot provide multi-view shape priors with guaranteed 3D consistency. We propose **Human 3Diffusion**: Realistic Avatar Creation via Explicit **3D** Consistent **Diffusion**. Our key insight is that 2D multi-view diffusion and 3D reconstruction models provide complementary information for each other, and by coupling them in a tight manner, we can fully leverage the potential of both models. We introduce a novel image-conditioned generative 3D Gaussian Splats reconstruction model that leverages the priors from 2D multi-view diffusion models, and provides an explicit 3D representation, which further guides the 2D reverse sampling process to have better 3D consistency. Experiments show that our proposed framework outperforms state-of-the-art methods and enables the creation of realistic avatars from a single RGB image, achieving high-fidelity in both geometry and appearance. Extensive ablations also validate the efficacy of our design, (1) multi-view 2D priors conditioning in generative

Preprint. Under review.

3D reconstruction and (2) consistency refinement of sampling trajectory via the explicit 3D representation. Our code and models will be released here.

1 Introduction

Realistic human avatar creation is crucial for various applications such as AR/VR, as well as the movie and gaming industry. Methods for creating a 3D avatar from a single RGB image are especially important to scale up avatar creation and make it more consumer-friendly compared to traditional studio-based capture methods. This task is, however, very challenging due to the vast diversity of human bodies and poses, further complicated by the wide variety of clothing and accessories. These challenges are exacerbated by the lack of large-scale 3D human data and ambiguities inherent in a monocular 2D view setting.

Recent image-to-3D approaches can be categorized into reconstruction-based and multi-view diffusion-based methods. Reconstruction-based approaches directly predict a 3D representation that can be rendered from any viewpoint. Due to the explicit 3D representation, these methods produce an arbitrary number of consistent viewpoint renderings. They obtain the 3D reconstruction either based on common template [21, 85, 86, 106] which utilize the SMPL [44] body model as the shape prior, or a flexible implicit function to represent loose clothing [5, 54, 55]. These methods, either template-based [21, 85, 86, 106] or template-free [5, 24, 54, 55, 72, 108], are typically deterministic which produce blurry textures and geometry in the occluded regions. More importantly, they are trained on small-scale datasets due to the limited amount of high-quality 3D data, which further restricts their ability to generalize to diverse shapes and textures.

Multi-view diffusion methods [40, 58, 75] distill the inherent 3D structure present in 2D diffusion models [53]. Typically, they fine-tune a large-scale 2D foundation model [23, 62] on a large 3D dataset of objects [12, 80, 97], to produce a *fixed* number of viewpoints. However, since these models diffuse images purely in 2D without explicit 3D constraints or representation, the resulting multi-views often lack 3D consistency [52, 39], which restricts downstream applications [65].

To address these challenges, we propose **3Diffusion**: realistic avatar creation via **3D** consistent **Diffusion** models. We design our method based on two key insights: 1) 2D multi-view diffusion models provide large-scale shape priors that can help 3D reconstruction; 2) A reconstructed 3D representation ensures 3D consistency across multi-views in 2D diffusion. Specifically, we propose a novel diffusion method, which bridges 3D Gaussian Splatting (3D-GS) [34] generation with a 2D multi-view diffusion model. At every iteration, multi-view images are denoised and reconstructed to 3D-GS to be re-rendered to continue the diffusion process. This 3D lifting during iterative sampling ensures the 3D consistency of the 2D diffusion model while leveraging a large-scale foundation model trained on billions of images. Our framework elegantly combines reconstruction methods with multi-view diffusion models. In summary, our contributions are:

- We propose a novel image-conditioned 3D-GS generation model for 3D reconstruction that bridges large-scale priors from 2D multi-view diffusion models and the efficient and explicit 3D-GS representation.
- A sophisticated diffusion process that incorporates reconstructed 3D-GS to improve the 3D consistency of 2D diffusion models by refining the reverse sampling trajectory.
- Our proposed formulation enables us to jointly train 2D diffusion and our 3D model on $\sim$ 6000 high-quality human scans and our method shows superior performance and generalization capability than prior works. Our code and pretrained models will be publicly released on our project page.

2 Related Work

Image to 3D. Creating realistic human avatar from consumer grade sensors [31, 94, 71, 90–92] is essential for downstream tasks such as human behaviour understanding [8, 49, 82, 84, 83] and gaming application [36, 42, 18, 103, 104]. Researchers have explored avatar creation from monocular RGB [29, 78], Depth [15, 91] video or single image [54, 55, 57, 85, 86]. Avatar from single image is particularly interesting and existing methods can be roughly categorized as template-based [21, 85, 86, 106] and template-free [54, 55, 57, 93]. Despite the impressive performance, template-based approaches rely on the naked body model [44, 48] and fail to reason extremely loose clothing, while template-free methods produce blurry back side textures. Instead of SMPL shape prior, our method is template-free and leverages strong 2D image priors to create high-quality avatars. Orthogonal to humans, object reconstruction methods typically adopt template-free paradigms and early works [6, 69, 79, 81, 105] focus mainly on geometry. With the advance of 2D diffusion models [53] and efficient 3D representation [11], recent works can reconstruct 3D objects with detailed textures [24, 39, 43, 58, 65, 72, 87, 88, 108]. One popular paradigm is first using strong 2D models [40, 59, 75] to produce multi-view images and then train another model to reconstruct 3D from multi-view images [39, 38, 43, 65, 89]. In practice, their performance is limited by the accuracy of the multi-view images generated by 2D diffusion modes. Our method tightly couples 2D and 3D models and yields better performance by guiding 2D sampling with 3D reconstruction.

Shape Prior from 2D Diffusion Model. Being trained on billions of images [56], 2D image diffusion models [53] have been shown to have 3D awareness and some works tried to use score distillation sampling [51, 77] to distil 3D knowledge of 2D models [37, 46, 107]. Other works propose to further enhance the 3D reasoning ability by fine-tuning the model on large-scale datasets [12, 80, 97] to generate multi-view images [32, 35, 40, 41, 58, 59, 66, 74, 75]. Dense self-attention [73, 75], depth-aware attention [26] or epipolar attention [28, 64] are introduced to enhance the 3D consistency of multi-views. However, these methods do not have explicit 3D while our method incorporates explicit 3D consistency into the reverse sampling process and obtains better results.

3 Preliminaries

Denoising Diffusion Probabilistic Models. DDPM [23] is a generative model which learns a data distribution by iteratively adding (forward process) and removing (reverse process) the noise. Formally, the forward process iteratively adds noise to a sample $\mathbf{x}_0$ drawn from a distribution $p_{\text{data}}(\mathbf{x})$:

$$\mathbf{x}_t \sim \mathcal{N}(\mathbf{x}_t; \sqrt{\alpha_t}\mathbf{x}_{t-1}, (1-\alpha_t)\mathbf{I}) := \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, \text{ where } \epsilon \sim \mathcal{N}(0, \mathbf{I}), \tag{1}$$

where $\alpha_t, \bar{\alpha}_t$ schedules the amount of noise added at each step t [23]. To sample data from the learned distribution, the reverse process starts from $\mathbf{x}_T \sim \mathcal{N}(0, \mathbf{I})$ and iteratively denoises it until $t = 0$:

$$\mathbf{x}_{t-1} \sim \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \tilde{\beta}_{t-1}\mathbf{I}), \text{ where } \tilde{\beta}_{t-1} = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}(1-\alpha_t) \tag{2}$$

A network parametrized by θ is trained to estimate the posterior mean μ_θ at each step t. One can also model conditional distribution with DDPM by adding the condition to the network input [14, 22].

2D Multi-View Diffusion Models. Many recent works [40, 41, 43, 58, 66, 75] propose to leverage strong 2D image diffusion prior [53] pre-trained on billions images [56] to generate multi-view images from a single image. Among them, ImageDream [75] demonstrated a superior generalization capability to unseen objects [65]. Given a single condition image $\mathbf{x}^{\text{c}}$ and an optional text description y, ImageDream generate 4 orthogonal target views $\mathbf{x}^{\text{tgt}}$ with a model ϵ_θ, which is trained to estimate the noise added at each step t. With the estimated noise ϵ_θ, one can compute the "clear" target views $\tilde{\mathbf{x}}_0^{\text{tgt}}$ with close-form solution in Eq. (1):

$$\tilde{\mathbf{x}}_0^{\text{tgt}} = \frac{1}{\sqrt{\bar{\alpha}_t}}(\mathbf{x}_t^{\text{tgt}} - \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_\theta(\mathbf{x}_t^{\text{tgt}}, \mathbf{x}^{\text{c}}, y, t)). \tag{3}$$

This *one-step* estimation of $\tilde{\mathbf{x}}_0^{\text{tgt}}$ can be noisy, especially when t is large and $\mathbf{x}_t^{\text{tgt}}$ is extremely noisy. Thus, the iterative sampling of $\mathbf{x}_t^{\text{tgt}}$ is required until $t = 0$. To sample next step $\mathbf{x}_{t-1}^{\text{tgt}}$, standard DDPM [23] computes the posterior mean μ_θ from current $\mathbf{x}_t^{\text{tgt}}$ and estimated $\tilde{\mathbf{x}}_0^{\text{tgt}}$ at step t with:

$$\mu_\theta(\mathbf{x}_t^{\text{tgt}}, t) := \mu_{t-1}(\mathbf{x}_t^{\text{tgt}}, \tilde{\mathbf{x}}_0^{\text{tgt}}) = \frac{\sqrt{\alpha_t}\,(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{x}_t^{\text{tgt}} + \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\tilde{\mathbf{x}}_0^{\text{tgt}}, \text{ where } \beta_t = 1-\alpha_t. \tag{4}$$

Afterwards, $\mathbf{x}_{t-1}^{\text{tgt}}$ can be sampled from Gaussian distribution with mean μ_{t-1} and variance $\tilde{\beta}_{t-1}\mathbf{I}$ (Eq. (2)) and used as the input for the next iteration. The reverse sampling is repeated until $t = 0$ where 4 clear target views are generated.

Although multi-view diffusion models [41, 58, 75] generate multiple views together, the 3D consistency across these views is not guaranteed due to the lack of an explicit 3D representation. Thus, we propose a novel 3D consistent diffusion model, which ensures the multi-view consistency at each step of the reverse process by diffusing 2D images using reconstructed 3D Gaussian Splats [34].

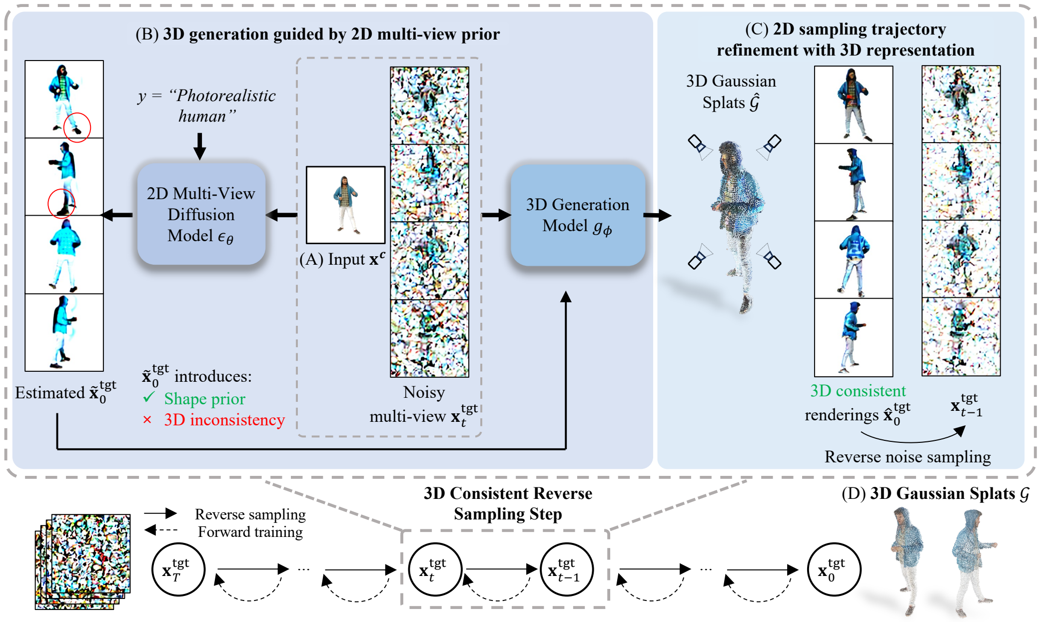


Figure 2: **Method Overview.** Given a single RGB image (A), we sample a realistic 3D avatar represented as 3D Gaussian Splats (D). At each reverse step, our 3D generation model g_ϕ leverages 2D multi-view diffusion prior from ϵ_θ which provides a strong shape prior but is not 3D consistent (B, cf. Sec. 4.1). We then refine the 2D reverse sampling trajectory with generated 3D renderings that are guaranteed to be 3D consistent (C, cf. Sec. 4.2). Our tight coupling ensures 3D consistency at each sampling step and obtains a high-quality 3D avatar (D).

4 3Diffusion

Overview. Given a single RGB image, we aim to create a realistic 3D avatar consistent with the input. We adopt an image-conditioned 3D generation paradigm due to inherent ambiguities in the monocular view. We introduce a novel 3D Gaussian Splatting (3D-GS [34]) generative model that combines shape priors from 2D multi-view diffusion models with the explicit 3D-GS representation. This allows us to jointly train our 3D generative model and a 2D multi-view diffusion model end-to-end and improves the 3D consistency of 2D multi-view generation at inference time.

In this section, we first introduce our novel generative 3D-GS reconstruction model in Sec. 4.1. We then describe how we leverage the 3D reconstruction to generate 3D consistent multi-view results by refining the reverse sampling trajectory (Sec. 4.2). An overview of our method can be found in Fig. 2.

4.1 Generative 3D-GS Reconstruction with Diffusion Priors

Given a context image $\mathbf{x}^c$, we use a conditional diffusion model to learn and sample from a plausible 3D distribution. Previous works demonstrated that 3D generation can be done implicitly via diffusing rendered images of a differentiable 3D represetation [7, 33, 68] such as NeRF [47, 95]. In this work, we introduce a novel generative model for 3D Gaussian Splatings [34], which diffuses rendered images of 3D-GS and enables sampling of 3D-GS at inference time. Single image to 3D generation is however very challenging, we hence propose to leverage 2D multi-view diffusion models in a tightly coupled manner which allows us to train it end-to-end with our novel 3D generative model.

Generative 3D-GS Reconstruction. In this work, we propose a 3D-GS generative model g_ϕ, which is conditioned on input context image $\mathbf{x}^c$ to perform reconstruction of 3D Gaussian Splats $\mathcal{G}$. Diffusing directly in the space of $\mathcal{G}$ parameters requires pre-computing Gaussian Splats from scans, which is exorbitant. Instead, we diffuse the multi-view renderings of $\mathcal{G}$ using a differentiable rendering function `renderer`.
We denote $\mathbf{x}_0^{\text{tgt}}$ as the ground truth images at target views to be diffused and $\mathbf{x}_0^{\text{novel}}$ as the additional

novel views for supervision. At training time, we uniformly sample a timestep $t \sim \mathcal{U}(0, T)$ and add noise to $\mathbf{x}_0^{\text{tgt}}$ using Eq. (1) to obtain noisy target views $\mathbf{x}_t^{\text{tgt}}$. Our generative model g_ϕ takes $\mathbf{x}_t^{\text{tgt}}$, diffusion timestep t, and the conditional image $\mathbf{x}^c$ as input, and estimates 3D Gaussians $\hat{\mathcal{G}}$:

$$\hat{\mathcal{G}} = g_\phi(\mathbf{x}_t^{\text{tgt}}, t, \mathbf{x}^c), \text{ where } \mathbf{x}_t^{\text{tgt}} = \sqrt{\bar{\alpha}_t}\mathbf{x}_0^{\text{tgt}} + \sqrt{1-\bar{\alpha}_t}\epsilon, \text{ and } \epsilon \sim \mathcal{N}(0, \mathbf{I}) \tag{5}$$

We adopt an asymmetric U-Net Transformer proposed by [65] for g_ϕ to directly predict 3D-GS parameters from per-pixel features of the last U-Net layer. To supervise the generative model g_ϕ, we use a differentiable rendering function $\texttt{renderer} : \{\mathcal{G}, \pi^{\text{p}}\} \mapsto \mathbf{x}^{\text{p}}$ to render images at target views π^{tgt} and additional novel views π^{novel}. Denoting $\mathbf{x}_0 := \{\mathbf{x}_0^{\text{tgt}}, \mathbf{x}_0^{\text{novel}}\}$ as ground truth and $\hat{\mathbf{x}}_0 := \{\hat{\mathbf{x}}_0^{\text{tgt}}, \hat{\mathbf{x}}_0^{\text{novel}}\}$ as rendered images, we compute the loss on images and generated 3D-GS:

$$\begin{aligned} \mathcal{L}_{gs} = \lambda_1 \cdot \mathcal{L}_{\text{MSE}}\big(\mathbf{x}_0, \hat{\mathbf{x}}_0\big) + \lambda_2 \cdot \mathcal{L}_{\text{Percep}}\big(\mathbf{x}_0, \hat{\mathbf{x}}_0\big) + \lambda_3 \cdot \mathcal{L}_{\text{reg}}(g_\phi(\mathbf{x}_t^{\text{tgt}}, t, \mathbf{x}^c)), \\ \text{where } \hat{\mathbf{x}}_0 := \{\hat{\mathbf{x}}_0^{\text{tgt}}, \hat{\mathbf{x}}_0^{\text{novel}}\} = \texttt{renderer}(g_\phi(\mathbf{x}_t^{\text{tgt}}, t, \mathbf{x}^c), \{\pi^{\text{tgt}}, \pi^{\text{novel}}\}), \end{aligned} \tag{6}$$

here $\mathcal{L}_{\text{MSE}}$ denotes the Mean Square Error (MSE) and $\mathcal{L}_{\text{Percep}}$ is the perceptual loss based on VGG-19 [60]. We also apply $\mathcal{L}_{\text{reg}}$, a geometry regularizer [27, 98] to stabilize the generation of $\hat{\mathcal{G}}$.

With this, we can train a generative model that diffuses 3D-GS *implicitly* by diffusing 2D images $\mathbf{x}_t^{\text{tgt}}$. At inference time, we can generate 3D-GS given the input image by denoising 2D multi-views sampled from Gaussian distribution. We initialize $\mathbf{x}_T^{\text{tgt}}$ from $\mathcal{N}(0, \mathbf{I})$, and iteratively denoise the rendered images of predicted $\hat{\mathcal{G}}$ from our model g_ϕ. At each reverse step, our model g_ϕ estimates a clean state $\hat{\mathcal{G}}$ and render target images $\hat{\mathbf{x}}_0^{\text{tgt}}$. We then calculate target images $\mathbf{x}_{t-1}^{\text{tgt}}$ for the next step via Eq. (4) and repeat the process until $t = 0$. For more details, please refer to Appendix A.3

Our generative 3D-GS reconstruction model archives superior performance on in-distribution human reconstruction yet generalizes poorly to unseen categories such as general objects (Sec. 5.3 Fig. 5). Our key insight for better generalization is leveraging strong priors from pretrained 2D multi-view diffusion models for 3D-GS generation.

3D-GS Generation with 2D Multi-view Diffusion. Pretrained 2D multi-view diffusion models (MVD) [41, 59, 75] have seen billions of real images [56] and millions of 3D data [12], which provide strong prior information and can generalize to unseen objects [65, 87]. Here, we propose a simple yet elegant idea for incorporating this multi-view prior into our generative 3D-GS model g_ϕ. We can also leverage generated 3D-GS to guide 2D MVD sampling process which we discuss in Sec. 4.2.
Our key observation is that both 2D MVD and our proposed 3D-GS generative model are diffusion-based and share the same sampling state $\mathbf{x}_t^{\text{tgt}}$ at timestep t. Thus, they are *synchronized*. This enables us to couple and facilitate information exchange between 2D MVD $\boldsymbol{\epsilon}_\theta$ and 3D-GS generative model g_ϕ at the same diffusion timestep t. To inject the 2D diffusion priors into 3D generation, we first compute *one-step* estimation of $\tilde{\mathbf{x}}_0^{\text{tgt}}$ (Eq. (3)) using 2D MVD ϵ_θ, and condition our 3D-GS generative mode g_ϕ additionally on it. Formally, our 3D-GS generative model enhanced with 2D multi-view diffusion priors is written as:

$$\hat{\mathcal{G}} = g_\phi(\mathbf{x}_t^{\text{tgt}}, t, \mathbf{x}^c, \tilde{\mathbf{x}}_0^{\text{tgt}}), \text{ where } \tilde{\mathbf{x}}_0^{\text{tgt}} = \frac{1}{\sqrt{\bar{\alpha}_t}}(\mathbf{x}_t^{\text{tgt}} - \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_\theta(\mathbf{x}_t^{\text{tgt}}, \mathbf{x}^{\text{c}}, y, t)) \tag{7}$$

The visualization of $\tilde{\mathbf{x}}_0^{\text{tgt}}$ along the whole sampling trajectory in Fig. 7 shows that the pretrained 2D diffusion model $\boldsymbol{\epsilon}_\theta$ can already provide useful multi-view shape prior even in large timestep $t = 1000$. This is further validated in our experiments where the additional 2D diffusion prior $\tilde{\mathbf{x}}_0^{\text{tgt}}$ leads to better avatar reconstruction (Tab. 4) as well as more robust generalization to general objects (Fig. 5). By utilizing the timewise iterative manner of 2D and 3D diffusion models, we can not only leverage 2D priors for 3D-GS generation but also train both models jointly end to end, which we discuss next.

Joint Training with 2D Model. We adopt pretrained ImageDream [75] as our 2D multi-view diffusion model ϵ_θ and jointly train it with our 3D-GS generative model g_ϕ. We observe that our joint training is important for coherent 3D generation, as opposed to prior works that frozen pretrained 2D multi-view models [65, 72]. We summarize our training algorithm in Algorithm 1. We combine the loss of 2D diffusion and our 3D-GS generation loss $\mathcal{L}_{gs}$(Eq. (6)):

$$\mathcal{L}_{total} = \mathcal{L}_{\text{MSE}}(\boldsymbol{\epsilon}, \boldsymbol{\epsilon}_\theta) + \mathcal{L}_{gs} \tag{8}$$

Algorithm 1 Training

Input: Dataset of posed multi-view images $\mathbf{x}_0^{\text{tgt}}$, π^{tgt}, $\mathbf{x}_0^{\text{novel}}$, π^{novel}, a context image $\mathbf{x}^{\text{c}}$, text description y

Output: Optimized 2D multi-view diffusion model ϵ_θ and 3D-GS generative model g_ϕ

repeat
  $\{\mathbf{x}_0^{\text{tgt}}, \mathbf{x}_0^{\text{novel}}, \mathbf{x}^{\text{c}}, y\} \sim q(\{\mathbf{x}_0^{\text{tgt}}, \mathbf{x}_0^{\text{novel}}, \mathbf{x}^{\text{c}}, y\})$
  $t \sim \text{Uniform}(\{1, \ldots, T\})$; $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
  $\mathbf{x}_t^{\text{tgt}} = \sqrt{\bar{\alpha}_t}\mathbf{x}_0^{\text{tgt}} + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}$
  $\tilde{\mathbf{x}}_0^{\text{tgt}} = \frac{1}{\sqrt{\bar{\alpha}_t}}(\mathbf{x}_t^{\text{tgt}} - \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_\theta(\mathbf{x}_t^{\text{tgt}}, \mathbf{x}^{\text{c}}, y, t))$
  $\hat{\mathcal{G}} = g_\phi\left(\mathbf{x}_t^{\text{tgt}}, t, \mathbf{x}^{\text{c}}, \tilde{\mathbf{x}}_0^{\text{tgt}}\right)$ // Enhance conditional 3D generation with 2D diffusion prior $\tilde{\mathbf{x}}_0^{\text{tgt}}$ from ϵ_θ
  $\{\hat{\mathbf{x}}_0^{\text{tgt}}, \hat{\mathbf{x}}_0^{\text{novel}}\} = \texttt{renderer}\left(\hat{\mathcal{G}}, \{\pi^{\text{tgt}}, \pi^{\text{novel}}\}\right)$
  Compute loss $\mathcal{L}_{total}$ (Eq. (8))
  Gradient step to update ϵ_θ, g_ϕ
until converged

Algorithm 2 3D Consistent Sampling

Input: A context image $\mathbf{x}^c$ and text y; Converged 2D diffusion model ϵ_θ and 3D generative model g_ϕ

Output: A 3D Gaussian Avatar $\mathcal{G}$ of the 2D image $\mathbf{x}^c$

$\mathbf{x}_T^{\text{tgt}} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
for $t = T, \ldots, 1$ **do**
  $\tilde{\mathbf{x}}_0^{\text{tgt}} = \frac{1}{\sqrt{\bar{\alpha}_t}}(\mathbf{x}_t^{\text{tgt}} - \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_\theta(\mathbf{x}_t^{\text{tgt}}, \mathbf{x}^{\text{c}}, y, t))$
  $\hat{\mathcal{G}} = g_\phi\left(\mathbf{x}_t^{\text{tgt}}, t, \mathbf{x}^{\text{c}}, \tilde{\mathbf{x}}_0^{\text{tgt}}\right)$
  $\hat{\mathbf{x}}_0^{\text{tgt}} = \texttt{renderer}\left(\hat{\mathcal{G}}, \pi^{\text{tgt}}\right)$
  $\mu_{t-1}(\mathbf{x}_t^{\text{tgt}}, \hat{\mathbf{x}}_0^{\text{tgt}}) = \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t\text{-}1})}{1-\bar{\alpha}_t}\mathbf{x}_t^{\text{tgt}} + \frac{\sqrt{\bar{\alpha}_{t\text{-}1}}\beta_t}{1-\bar{\alpha}_t}\hat{\mathbf{x}}_0^{\text{tgt}}$ // Guide 2D sampling with 3D consistent renderings
  $\mathbf{x}_{t-1}^{\text{tgt}} \sim \mathcal{N}\left(\mathbf{x}_{t-1}^{\text{tgt}}; \tilde{\boldsymbol{\mu}}_t\left(\mathbf{x}_t^{\text{tgt}}, \hat{\mathbf{x}}_0^{\text{tgt}}\right), \tilde{\beta}_{t-1}\mathbf{I}\right)$
end for
return $\mathcal{G} = g_\phi\left(\mathbf{x}_0^{\text{tgt}}, \tilde{\mathbf{x}}_0^{\text{tgt}}, \mathbf{x}^{\text{c}}, t = 0\right)$

Once trained, one can sample a plausible 3D-GS avatar $\mathcal{G}$ conditioned on the input image from the learned 3D distributions. However, we observe that the multi-view diffusion model $\boldsymbol{\epsilon}_\theta$ can still output inconsistent multi-views along the sampling trajectory (see Fig. 2). On the other hand, our 3D generator produces explicit 3D-GS which can be rendered as 3D consistent multi-views. Our second key idea is to use the 3D consistent renderings to guide 2D sampling process for more 3D consistent mulit-view generation. We discuss this in Sec. 4.2.

4.2 Guide 2D Multi-view Sampling with Reconstructed 3D-GS

With the shared and *synchronized* sampling state $\mathbf{x}_t^{\text{tgt}}$ of 2D multi-view diffusion model $\boldsymbol{\epsilon}_\theta$ and 3D-GS reconstruction model g_ϕ, we couple both models at arbitrary t during training. Similarly, they are also connected by both using estimated clean multi-views $\mathbf{x}_0^{\text{tgt}}$ at sampling time. To leverage the full potential of both models, we carefully design a joint sampling process that utilizes the reconstructed 3D-GS $\hat{\mathcal{G}}$ at each timestep t to guide 2D multi-view sampling, which is summarized in Algorithm 2. We observe that the key difference between the clean multi-views estimated $\mathbf{x}_0^{\text{tgt}}$ from 2D diffusion model and our 3D-GS generation lies in 3D consistency: 2D MVD computes multi-view $\tilde{\mathbf{x}}_0^{\text{tgt}}$ from 2D network prediction which can be 3D inconsistent while our $\hat{\mathbf{x}}_0^{\text{tgt}}$ are rendered from explicit 3D-GS representation which are guaranteed to be 3D consistent. Our idea is to guide the 2D multi-view reverse sampling process with our 3D consistent renderings $\hat{\mathbf{x}}_0^{\text{tgt}}$ such that the 2D sampling trajectory is more 3D consistent. Specifically, we leverage 3D consistent multi-view renderings $\hat{\mathbf{x}}_0^{\text{tgt}}$ to refine the posterior mean $\mu_\theta(\mathbf{x}_t^{\text{tgt}}, t)$ at each reverse step:

$$\text{Original: } \mu_\theta(\mathbf{x}_t^{\text{tgt}}, t) := \mu_{t-1}(\mathbf{x}_t^{\text{tgt}}, \tilde{\mathbf{x}}_0^{\text{tgt}}) \quad \rightarrow \quad \text{Ours: } \mu_\theta(\mathbf{x}_t^{\text{tgt}}, t) := \mu_{t-1}(\mathbf{x}_t^{\text{tgt}}, \hat{\mathbf{x}}_0^{\text{tgt}}),$$
$$\text{where } \hat{\mathbf{x}}_0^{\text{tgt}} = \texttt{renderer}(\hat{\mathcal{G}}, \pi^{\text{tgt}}), \text{ and } \mu_{t-1}(\mathbf{x}_t^{\text{tgt}}, \hat{\mathbf{x}}_0^{\text{tgt}}) = \frac{\sqrt{\alpha_t}\,(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{x}_t^{\text{tgt}} + \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\hat{\mathbf{x}}_0^{\text{tgt}} \tag{9}$$

With this refinement, we guarantee the 3D consistency at each reverse step t and avoid 3D inconsistency accumulation in original multi-view sampling [75]. In Fig. 7, we visualize the evolution of originally generated multi-views $\tilde{\mathbf{x}}_0^{\text{tgt}}$ and multi-views rendering $\hat{\mathbf{x}}_0^{\text{tgt}}$ from generated 3D-GS $\hat{\mathcal{G}}$ along the whole reverse sampling process. It intuitively shows how effective the sampling trajectory refinement is. We perform extensive ablation in Sec. 5.3 showing the importance of the consistent refinement for sampling trajectory.

5 Experiments

In this section, we first compare against baseline methods for human reconstruction in Sec. 5.2 and then ablate our design choices in Sec. 5.3.

Figure 3: **Qualitative comparison with baselines** on SIZER [70] and IIIT [30] dataset. Template-based methods (SiTH [21] and SIFU [106]) cannot reconstruct loose clothing coherently and template free methods (PIFu [54] and TripoSR [72]) tend to generate blurry texture in unseen regions due to their deterministic formulation. LGM [65] and InstantMesh [87] cannot correct 3D inconsistency from 2D multi-views hence the texture is also blurry. In contract, our method is able to reconstruct avatars with realistic textures and plausible 3D geometry in both seen and unseen region.

5.1 Experimental Setup

Datasets. We train our model on a combined 3D human dataset [1, 3, 4, 2, 19, 25, 63, 96] compromising $\sim$ 6000 high quality scans. Please refer to Appendix D.1. We evaluate qualitatively and quantitatively on Sizer [70] and more challenging IIIT [30] dataset due to extremely loose clothing of traditional custom suits. Please refer to Appendix D.2 for examples in evaluation dataset.

Implementation Details. We trained our model on 8 NVIDIA A100 GPUs over approximately 5 days. Each GPU was configured with a batch size 2 and gradient accumulations of 16 steps to achieve an effective batch size of 256. For more training details regarding hyperparameters, diffusion schedulers, etc., please refer to Appendix A.1.

Evaluation Metrics. We evaluate the geometry quality using Chamfer Distance (CD), F-score [67] (w/ threshold of $0.01m$), and Normal Consistency (NC) between the extracted mesh (Appendix A.4) and the groundtruth scan. Appearance quality is assessed by rendering the reconstructed avatar from 32 novel views with uniform azimuth and 0 elevation angle. The metrics for appearance reported include multi-scale Structure Similarity (SSIM) [76], Learned Perceptual Image Patch Similarity (LPIPS) [102], and Peak Signal to Noise Ratio (PSNR) between rendered and ground-truth views. Moreover, we report the Fréchet inception distance (FID) [20] between synthesized views and ground truth renderings, which reflects the quality and realism of the unseen regions.

5.2 Realistic Avatar from Image

We compare our approach against prior methods for image-to-avatar reconstruction. including template-based [21, 54, 106] and template-free [54] human reconstruction methods, as well as general image-to-3D methods [65, 72]. To further assess performance, we also fine-tuned the state-of-the-art object reconstruction method LGM [65] and its deployed multi-view diffusion model [75] on our training data, denoted as $\text{LGM}_{\text{human}}$. Quantitative evaluations reported in Tab. 1 demonstrate

Method	$CD_{(cm)}$ ↓	F-score ↑	NC ↑	SSIM ↑	LPIPS ↓	PSNR ↑	FID ↓
TripoSR [72]	2.59	0.360	0.771	0.743	0.095	20.05	26.13
InstantMesh [87]	2.47	0.338	0.787	0.900	0.084	20.54	26.38
LGM [65]	3.29	0.275	0.562	0.892	0.088	20.11	24.21
PIFu [54]	2.83	0.333	0.769	0.907	0.078	20.66	32.73
SiTH [21]	3.92	0.250	0.735	0.907	0.074	20.88	24.76
SIFU [106]	3.34	0.235	0.739	0.896	0.085	20.39	42.63
LGM_{human}	2.08	0.300	0.661	0.904	0.075	20.61	15.01
Ours	**1.35**	**0.550**	**0.798**	**0.918**	**0.060**	**21.49**	**9.57**

Table 1: **Quantitative evaluation** on SIZER [70] and IIIT [30] dataset. Our method reconstructs more accurate geometry (CD, F-score, NC) and realistic textures (SSIM, LPIPS, PSNR, FID).

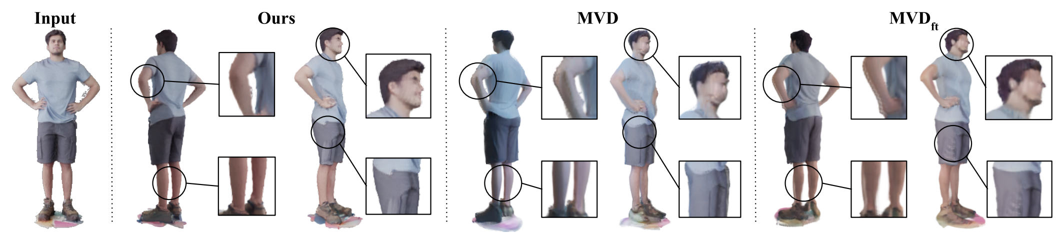


Figure 4: **3D reconstruction conditioned on different multi-view priors.** Without our 3D-consistent sampling, the 2D diffusion model cannot generate 3D consistent multi-views (MVD, MVD_{ft}), leading to artifacts like floating 3D Gaussians splats. Our method obtains more consistent multi-views hence better 3D-GS and rendering.

that our proposed method excels in reconstructing realistic avatars with accurate geometry (CD, NC, F-score) and realistic texture (SSIM, LPIPS, PSNR, FID) from a single RGB image.

We present qualitative comparison examples in Fig. 3 and Appendix B.1, highlighting the strengths and weaknesses of competing methods. Template-based methods such as SiTH [21] and SIFU [106] struggle to accurately reconstruct the geometry of loose clothing (as shown in row 4) due to their reliance on the naked SMPL body shape. In contrast, template-free methods like PIFu [54] and TripoSR [72] offer greater flexibility and better performance on loose clothing. However, they are not generative models and their deterministic formulations lead to blurry textures in unseen regions, as they tend to produce average textures rather than distinct details. Similar to our approach, LGM [65] and InstantMesh [87] utilize 2D diffusion models to generate multi-view images and perform sparse-view 3D reconstruction. Nonetheless, their separation of 2D and 3D models cannot correct the 3D inconsistencies that may arise from the 2D models. Even further fine-tuning of LGM on human scans (Fig. 4) does not adequately address these challenges due to the complex and sensitive nature of human geometry and textures. In contrast, our conditional generative formulation and inherent 3D consistency by tightly coupling 2D-3D models allow us to obtain accurate reconstruction in front view and realistic generation in unseen regions. We also show the generative power of our method in Appendix C.5: by sampling with different seed, we obtain diverse yet plausible reconstruction.

Please also refer to Fig. 6, Appendix C and our project page for additional reconstruction results on challenging subjects not previously observed, encompassing a diverse range of appearances such as loose skirts and custom suits, as well as accessories like bags and gloves.

5.3 Ablation Studies

Importance of Trajectory Refinement. One of our key ideas is leveraging our explicit 3D model to refine the 2D multi-view reverse sampling trajectory, ensuring 3D consistency in Multi-View Diffusion (MVD) generation (see Sec. 4.2 and Eq. (9)). To evaluate this, we compare the multi-view images generated by pretrained MVD, fine-tuned MVD on our data (MVD_{ft}) and MVD with our 3D consistent sampling

Method	LPIPS ↓	SSIM ↑	PSNR ↑
MVD	0.078	0.911	22.32
MVD_{ft}	0.061	0.926	24.14
Ours	**0.048**	**0.934**	**24.69**

Table 2: **Evaluating trajectory refinement** for 2D multi-view diffusion. Our proposed refinement improves multi-view image quality.

(ours), as shown in Tab. 2. The results demonstrate that our proposed method effectively enhances the quality of generated multi-view images by leveraging the explicit 3D model to refine sampling trajectory. Additionally. we analyze the 3D reconstruction results with the multi-view images generated by these models in Fig. 4. MVD and MVD_{ft} produce inconsistent multi-view images, which typically lead to floating Gaussian and hence blurry boundaries. In contrast, our method can generate more consistent multi-views, result in better 3D Gaussians Splats and sharper renderings.

Method	$\text{CD}_{(\text{cm})}\downarrow$	F-score↑	NC ↑	LPIPS↓	SSIM↑	PSNR↑
Our w/o Traj. Ref.	1.57	0.498	0.794	0.064	0.908	21.09
Ours	**1.35**	**0.550**	**0.798**	**0.060**	**0.918**	**21.49**

Table 3: **Evaluating trajectory refinement** for final 3D reconstruction. Our sampling trajectory refinement ensures multi-view consistency and hence yields better 3D results.

We further quantitatively evaluate the impact of our proposed sampling trajectory refinement on final 3D reconstruction in Tab. 3. We compare the reconstruction results of methods with and without our trajectory refinement while using the same 2D MVD and 3D reconstruction models. It can be clearly seen that our trajectory refinement improves the quality of 3D reconstruction.

Importance of 2D Multi-view Prior. Another key idea of our work is the use of multi-view priors $\tilde{\mathbf{x}}_0^{\text{tgt}}$ from 2D diffusion model pretrained on massive data [12, 53, 56] to enhance our 3D generative model. This additional prior information is pivotal for ensuring accurate reconstruction of both in-distribution human dataset and generalizing to out-of-distribution objects.

We evaluate the performance of our 3D model g_ϕ by comparing generation results with and without the 2D diffusion prior $\tilde{\mathbf{x}}_0^{\text{tgt}}$ (refer to Eq. (7) and Eq. (5)). Notably, without the 2D multi-view conditioning, the alignment of the generated 3D model in the front view is not guaranteed due to the relative camera pose settings in our 3D generative model g_ϕ. Therefore, we evaluate the overall quality solely through the Fréchet Inception Distance (FID).

Method	FID ↓
Ours w/o $\tilde{\mathbf{x}}_0^{\text{tgt}}$	12.75
Our full model	**9.57**

Table 4: **2D multi-view priors $\tilde{\mathbf{x}}_0^{\text{tgt}}$** improve human reconstruction quality.

For avatars reconstruction, our powerful 3D reconstruction model can already achieve state-of-the-art performance. Moreover, our full model with multi-view prior $\tilde{\mathbf{x}}_0^{\text{tgt}}$ generates avatars with higher quality as demonstrated in Tab. 4. We further evaluate it on the GSO [16] dataset which consists of unseen general objects to our model. The improvements are even more pronounced in this setting, highlighting the challenges of generating coherent 3D structures from a single 2D image, particularly with unseen objects. For additional examples, please see Fig. 15 in Supp..

Method	FID ↓
w/o $\tilde{\mathbf{x}}_0^{\text{tgt}}$	107.14
Ours	**83.89**

Figure 5: 2D multi-view priors $\tilde{\mathbf{x}}_0^{\text{tgt}}$ enhances generalization to general objects in GSO [16] dataset.

6 Limitations and Future Work

Currently, our method is constrained by the 256×256 resolution of the multi-view diffusion model, which restricts the sharpness of texture details (see Appendix E). Upgrading to more powerful high-resolution (512×512) multi-view diffusion models [17, 66] could potentially resolve these issues. Moreover, our approach may struggle in reconstructing subjects with challenging poses, as we further discussed in Appendix E. Synthesizing training data with challenging poses [9, 84] could be a potential solution.

Our method is a general framework for image-to-3D reconstruction, which is applicable to various objects and compositional shapes like human-object interactions. We leave these to future works.

7 Conclusion

In this paper, we introduce **Human 3Diffusion**, a 3D consistent diffusion model for creating realistic avatars from single RGB images. Our key ideas are two folds: 1) Leveraging strong multi-view priors

Figure 6: Visualization of reconstructed mesh and synthesized novel view of generated 3D-GS on subjects from (A) Sizer [70], (B) IIIT [30], (C) Twindom [4], (D) UBC Fashion [99], (E) GSO [16] and (F) online image "the Rock". More results are presented in Appendix C and our project page.

from pretrained 2D diffusion models to generate 3D Gaussian Splats, and 2) Using the reconstructed explicit 3D Gaussian Splats to refine the sampling trajectory of the 2D diffusion model which enhances 3D consistency. We carefully designed a diffusion process that synergistically combines the strengths of both 2D and 3D models. Our experiments show that our approach outperforms all previous reconstruction works in both appearance and geometry. We also extensively ablate our method which proves the effectiveness of our proposed ideas. Our code and pretrained models will be released on our Project Page to foster future research.

Acknowledgements We appreciate G.Tiwari, Y.He, Y. Xiu, Z.Liu, Z.Qiu, S.Li and others for their feedback to improve the work. This work is made possible by funding from the Carl Zeiss Foundation. This work is also funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - 409792180 (EmmyNoether Programme, project: Real Virtual Humans) and the German Federal Ministry of Education and Research (BMBF): Tübingen AI Center, FKZ: 01IS18039A. The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Y.Xue. G. Pons-Moll is a member of the Machine Learning Cluster of Excellence, EXC number 2064/1 – Project number 390727645. R. Marin has been supported by innovation program under the Marie Skłodowska-Curie grant agreement No 101109330.

References

[1] Axyz, Nov 2023. URL https://secure.axyz-design.com. 7, 34, 36

[2] Renderpeople, Nov 2023. URL https://renderpeople.com/. 7, 34

[3] Treedy, Nov 2023. URL https://treedys.com/. 7, 34

[4] Twindom, Nov 2023. URL https://web.twindom.com/. 7, 10, 34, 36

[5] Thiemo Alldieck, Mihai Zanfir, and Cristian Sminchisescu. Photorealistic Monocular 3D Reconstruction of Humans Wearing Clothing. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1496–1505, New Orleans, LA, USA, June 2022. IEEE. ISBN 978-1-66546-946-3. doi: 10.1109/CVPR52688.2022.00156. URL https://ieeexplore.ieee.org/document/9878998/. 2

[6] Kalyan Vasudev Alwala, Abhinav Gupta, and Shubham Tulsiani. Pre-train, self-train, distill: A simple recipe for supersizing 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3773–3782, June 2022. 3

[7] Titas Anciukevicius, Zexiang Xu, Matthew Fisher, Paul Henderson, Hakan Bilen, Niloy J. Mitra, and Paul Guerrero. Renderdiffusion: Image diffusion for 3d reconstruction, inpainting and generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 12608–12618. IEEE, 2023. doi: 10.1109/CVPR52729.2023.01213. URL https://doi.org/10.1109/CVPR52729.2023.01213. 4, 21

[8] Bharat Lal Bhatnagar, Xianghui Xie, Ilya Petrov, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Behave: Dataset and method for tracking human object interactions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2022. 2

[9] Michael J. Black, Priyanka Patel, Joachim Tesch, and Jinlong Yang. BEDLAM: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In *Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 8726–8737, June 2023. 9

[10] Xu Cao, Hiroaki Santo, Boxin Shi, Fumio Okura, and Yasuyuki Matsushita. Bilateral normal integration. 2022. 23

[11] Eric R. Chan, Connor Z. Lin, Matthew A. Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J. Guibas, Jonathan Tremblay, Sameh Khamis, Tero Karras, and Gordon Wetzstein. Efficient geometry-aware 3d generative adversarial networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 16102–16112. IEEE, 2022. doi: 10.1109/CVPR52688.2022.01565. URL https://doi.org/10.1109/CVPR52688.2022.01565. 3

[12] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 13142–13153. IEEE, 2023. doi: 10.1109/CVPR52729.2023.01263. URL https://doi.org/10.1109/CVPR52729.2023.01263. 2, 3, 5, 9, 34

[13] Maximilian Denninger, Dominik Winkelbauer, Martin Sundermeyer, Wout Boerdijk, Markus Knauer, Klaus H. Strobl, Matthias Humt, and Rudolph Triebel. Blenderproc2: A procedural pipeline for photorealistic rendering. *Journal of Open Source Software*, 8(82):4901, 2023. doi: 10.21105/joss.04901. URL https://doi.org/10.21105/joss.04901. 34

[14] Prafulla Dhariwal and Alexander Quinn Nichol. Diffusion models beat GANs on image synthesis. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=AAWuCvzaVt. 3

[15] Zijian Dong, Chen Guo, Jie Song, Xu Chen, Andreas Geiger, and Otmar Hilliges. Pina: Learning a personalized implicit neural avatar from a single rgb-d video sequence. *arXiv*, 2022. 2

[16] Laura Downs, Anthony Francis, Nate Koenig, Brandon Kinman, Ryan Hickman, Krista Reymann, Thomas Barlow McHugh, and Vincent Vanhoucke. Google scanned objects: A high-quality dataset of 3d scanned household items. In *2022 International Conference on Robotics and Automation, ICRA 2022, Philadelphia, PA, USA, May 23-27, 2022*, pages 2553–2560. IEEE, 2022. doi: 10.1109/ICRA46639.2022.9811809. URL https://doi.org/10.1109/ICRA46639.2022.9811809. 9, 10, 26, 31

[17] Ruiqi Gao*, Aleksander Holynski*, Philipp Henzler, Arthur Brussee, Ricardo Martin-Brualla, Pratul P. Srinivasan, Jonathan T. Barron, and Ben Poole*. Cat3d: Create anything in 3d with multi-view diffusion models. *arXiv*, 2024. 9, 38

[18] Vladimir Guzov, Aymen Mir, Torsten Sattler, and Gerard Pons-Moll. Human poseitioning system (HPS): 3d human pose estimation and self-localization in large scenes from body-mounted sensors. *CoRR*, abs/2103.17265, 2021. URL https://arxiv.org/abs/2103.17265. 2

[19] Sang-Hun Han, Min-Gyu Park, Ju Hong Yoon, Ju-Mi Kang, Young-Jae Park, and Hae-Gon Jeon. High-fidelity 3d human digitization from single 2k resolution images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR2023)*, June 2023. 7, 34, 36

[20] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 6626–6637, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/8a1d694707eb0fefe65871369074926d-Abstract.html. 7

[21] Hsuan-I Ho, Jie Song, and Otmar Hilliges. Sith: Single-view textured human reconstruction with image-conditioned diffusion. *CoRR*, abs/2311.15855, 2023. doi: 10.48550/ARXIV.2311.15855. URL https://doi.org/10.48550/arXiv.2311.15855. 2, 7, 8, 34, 35

[22] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. URL https://openreview.net/forum?id=qw8AKxfYbI. 3

[23] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html. 2, 3, 21

[24] Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. LRM: large reconstruction model for single image to 3d. *CoRR*, abs/2311.04400, 2023. doi: 10.48550/ARXIV.2311.04400. URL https://doi.org/10.48550/arXiv.2311.04400. 2, 3

[25] Ho Hsuan-I, Xue Lixin, Song Jie, and Hilliges Otmar. Learning locally editable virtual humans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 7, 34, 36

[26] Hanzhe Hu, Zhizhuo Zhou, Varun Jampani, and Shubham Tulsiani. MVD-Fusion: Single-view 3D via Depth-consistent Multi-view Generation, April 2024. 3

[27] Binbin Huang, Zehao Yu, Anpei Chen, Andreas Geiger, and Shenghua Gao. 2d gaussian splatting for geometrically accurate radiance fields. *CoRR*, abs/2403.17888, 2024. doi: 10.48550/ARXIV.2403.17888. URL https://doi.org/10.48550/arXiv.2403.17888. 5

[28] Zehuan Huang, Hao Wen, Junting Dong, Yaohui Wang, Yangguang Li, Xinyuan Chen, Yan-Pei Cao, Ding Liang, Yu Qiao, Bo Dai, and Lu Sheng. EpiDiff: Enhancing Multi-View Synthesis via Localized Epipolar-Constrained Diffusion, April 2024. 3

[29] Tianjian Jiang, Xu Chen, Jie Song, and Otmar Hilliges. Instantavatar: Learning avatars from monocular video in 60 seconds. *arXiv*, 2022. 2

[30] Sai Sagar Jinka, Astitva Srivastava, Chandradeep Pokhariya, Avinash Sharma, and P. J. Narayanan. SHARP: shape-aware reconstruction of people in loose clothing. *Int. J. Comput. Vis.*, 131(4):918–937, 2023. doi: 10.1007/S11263-022-01736-Z. URL https://doi.org/10.1007/s11263-022-01736-z. 7, 8, 10, 24, 25, 34, 35, 37

[31] Berna Kabadayi, Wojciech Zielonka, Bharat Lal Bhatnagar, Gerard Pons-Moll, and Justus Thies. Gan-avatar: Controllable personalized gan-based human head avatar. In *International Conference on 3D Vision (3DV)*, March 2024. 2

[32] Yash Kant, Ziyi Wu, Michael Vasilkovsky, Guocheng Qian, Jian Ren, Riza Alp Güler, Bernard Ghanem, Sergey Tulyakov, Igor Gilitschenski, and Aliaksandr Siarohin. SPAD : Spatially aware multiview diffusers. *CoRR*, abs/2402.05235, 2024. doi: 10.48550/ARXIV.2402.05235. URL https://doi.org/10.48550/arXiv.2402.05235. 3

[33] Animesh Karnewar, Andrea Vedaldi, David Novotny, and Niloy Mitra. Holodiffusion: Training a 3D diffusion model using 2D images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023. 4, 21

[34] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139:1–139:14, 2023. doi: 10.1145/3592433. URL https://doi.org/10.1145/3592433. 2, 3, 4, 21

[35] Xin Kong, Shikun Liu, Xiaoyang Lyu, Marwan Taher, Xiaojuan Qi, and Andrew J. Davison. Eschernet: A generative model for scalable view synthesis. *CoRR*, abs/2402.03908, 2024. doi: 10.48550/ARXIV.2402.03908. URL https://doi.org/10.48550/arXiv.2402.03908. 3

[36] Yifei Li, Hsiao-yu Chen, Egor Larionov, Nikolaos Sarafianos, Wojciech Matusik, and Tuur Stuyck. DiffAvatar: Simulation-ready garment optimization with differentiable simulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024. 2

[37] Tingting Liao, Hongwei Yi, Yuliang Xiu, Jiaxiang Tang, Yangyi Huang, Justus Thies, and Michael J. Black. Tada! text to animatable digital avatars. *CoRR*, abs/2308.10899, 2023. doi: 10.48550/ARXIV.2308.10899. URL https://doi.org/10.48550/arXiv.2308.10899. 3

[38] Minghua Liu, Ruoxi Shi, Linghao Chen, Zhuoyang Zhang, Chao Xu, Xinyue Wei, Hansheng Chen, Chong Zeng, Jiayuan Gu, and Hao Su. One-2-3-45++: Fast single image to 3d objects with consistent multi-view generation and 3d diffusion. *CoRR*, abs/2311.07885, 2023. doi: 10.48550/ARXIV.2311.07885. URL https://doi.org/10.48550/arXiv.2311.07885. 3

[39] Minghua Liu, Chao Xu, Haian Jin, Linghao Chen, Mukund Varma T, Zexiang Xu, and Hao Su. One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/4683beb6bab325650db13afd05d1a14a-Abstract-Conference.html. 2, 3

[40] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 9264–9275. IEEE, 2023. doi: 10.1109/ICCV51070.2023.00853. URL https://doi.org/10.1109/ICCV51070.2023.00853. 2, 3

[41] Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. Syncdreamer: Generating multiview-consistent images from a single-view image. *CoRR*, abs/2309.03453, 2023. doi: 10.48550/ARXIV.2309.03453. URL https://doi.org/10.48550/arXiv.2309.03453. 3, 5

[42] Zhen Liu, Yao Feng, Yuliang Xiu, Weiyang Liu, Liam Paull, Michael J. Black, and Bernhard Schölkopf. Ghost on the shell: An expressive representation of general 3d shapes. *arXiv preprint arXiv:2310.15168*, 2023. 2

[43] Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, and Wenping Wang. Wonder3d: Single image to 3d using cross-domain diffusion. *CoRR*, abs/2310.15008, 2023. doi: 10.48550/ARXIV.2310.15008. URL https://doi.org/10.48550/arXiv.2310.15008. 3

[44] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: a skinned multi-person linear model. *ACM Trans. Graph.*, 34(6):248:1–248:16, 2015. doi: 10.1145/2816795.2818013. URL https://doi.org/10.1145/2816795.2818013. 2, 3

[45] Qianli Ma, Siyu Tang, Sergi Pujades, Gerard Pons-Moll, Anurag Ranjan, and Michael J. Black. Dressing 3d humans using a conditional mesh-vae-gan. *CoRR*, abs/1907.13615, 2019. URL http://arxiv.org/abs/1907.13615. 34, 35, 37

[46] Luke Melas-Kyriazi, Christian Rupprecht, Iro Laina, and Andrea Vedaldi. Realfusion: 360° reconstruction of any object from a single image. In *Arxiv*, 2023. 3

[47] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: representing scenes as neural radiance fields for view synthesis. *Commun. ACM*, 65(1):99–106, 2022. doi: 10.1145/3503250. URL https://doi.org/10.1145/3503250. 4, 21

[48] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image. *CoRR*, abs/1904.05866, 2019. URL http://arxiv.org/abs/1904.05866. 3

[49] Ilya A. Petrov, Riccardo Marin, Julian Chibane, and Gerard Pons-Moll. Object pop-up: Can we infer 3d objects and their poses from human interactions alone? In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 4726–4736. IEEE, 2023. doi: 10.1109/CVPR52729.2023.00458. URL https://doi.org/10.1109/CVPR52729.2023.00458. 2

[50] Gerard Pons-Moll, Sergi Pujades, Sonny Hu, and Michael J. Black. Clothcap: seamless 4d clothing capture and retargeting. *ACM Trans. Graph.*, 36(4):73:1–73:15, 2017. doi: 10.1145/3072959.3073711. URL https://doi.org/10.1145/3072959.3073711. 34, 35, 37

[51] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv*, 2022. 3

[52] Guocheng Qian, Jinjie Mai, Abdullah Hamdi, Jian Ren, Aliaksandr Siarohin, Bing Li, Hsin-Ying Lee, Ivan Skorokhodov, Peter Wonka, Sergey Tulyakov, and Bernard Ghanem. Magic123: One image to high-quality 3d object generation using both 2d and 3d diffusion priors. *CoRR*, abs/2306.17843, 2023. doi: 10.48550/ARXIV.2306.17843. URL https://doi.org/10.48550/arXiv.2306.17843. 2

[53] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 10674–10685. IEEE, 2022. doi: 10.1109/CVPR52688.2022.01042. URL https://doi.org/10.1109/CVPR52688.2022.01042. 2, 3, 9, 21

[54] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Hao Li, and Angjoo Kanazawa. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 2304–2314. IEEE, 2019. doi: 10.1109/ICCV.2019.00239. URL https://doi.org/10.1109/ICCV.2019.00239. 2, 7, 8, 33

[55] Shunsuke Saito, Tomas Simon, Jason M. Saragih, and Hanbyul Joo. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 81–90. Computer Vision Foundation / IEEE, 2020. doi: 10.1109/CVPR42600.2020.00016. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Saito_PIFuHD_Multi-Level_Pixel-Aligned_Implicit_Function_for_High-Resolution_3D_Human_Digitization_CVPR_2020_paper.html. 2, 23, 33

[56] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5B: an open large-scale dataset for training next generation image-text models. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/a1859debfb3b59d094f3504d5ebb6c25-Abstract-Datasets_and_Benchmarks.html. 3, 5, 9

[57] Akash Sengupta, Thiemo Alldieck, Nikos Kolotouros, Enric Corona, Andrei Zanfir, and Cristian Sminchisescu. DiffHuman: Probabilistic Photorealistic 3D Reconstruction of Humans, March 2024. URL http://arxiv.org/abs/2404.00485. arXiv:2404.00485 [cs]. 2

[58] Ruoxi Shi, Hansheng Chen, Zhuoyang Zhang, Minghua Liu, Chao Xu, Xinyue Wei, Linghao Chen, Chong Zeng, and Hao Su. Zero123++: a single image to consistent multi-view diffusion base model. *CoRR*, abs/2310.15110, 2023. doi: 10.48550/ARXIV.2310.15110. URL https://doi.org/10.48550/arXiv.2310.15110. 2, 3

[59] Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. *CoRR*, abs/2308.16512, 2023. doi: 10.48550/ARXIV.2308.16512. URL https://doi.org/10.48550/arXiv.2308.16512. 3, 5

[60] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1409.1556. 5

[61] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=St1giarCHLP. 21

[62] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-Based Generative Modeling through Stochastic Differential Equations, February 2021. URL http://arxiv.org/abs/2011.13456. arXiv:2011.13456 [cs, stat]. 2

[63] Zhaoqi Su, Tao Yu, Yangang Wang, and Yebin Liu. Deepcloth: Neural garment representation for shape and style editing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):1581–1593, 2023. doi: 10.1109/TPAMI.2022.3168569. 7, 34, 36

[64] Mohammed Suhail, Carlos Esteves, Leonid Sigal, and Ameesh Makadia. Generalizable patch-based neural rendering. In Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXXII*, volume 13692 of *Lecture Notes*

in Computer Science, pages 156–174. Springer, 2022. doi: 10.1007/978-3-031-19824-3_10. URL https://doi.org/10.1007/978-3-031-19824-3_10. 3

[65] Jiaxiang Tang, Zhaoxi Chen, Xiaokang Chen, Tengfei Wang, Gang Zeng, and Ziwei Liu. LGM: large multi-view gaussian model for high-resolution 3d content creation. *CoRR*, abs/2402.05054, 2024. doi: 10.48550/ARXIV.2402.05054. URL https://doi.org/10.48550/arXiv.2402.05054. 2, 3, 5, 7, 8, 21

[66] Shitao Tang, Jiacheng Chen, Dilin Wang, Chengzhou Tang, Fuyang Zhang, Yuchen Fan, Vikas Chandra, Yasutaka Furukawa, and Rakesh Ranjan. Mvdiffusion++: A dense high-resolution multi-view diffusion model for single or sparse-view 3d object reconstruction. *CoRR*, abs/2402.12712, 2024. doi: 10.48550/ARXIV.2402.12712. URL https://doi.org/10.48550/arXiv.2402.12712. 3, 9, 38

[67] Maxim Tatarchenko, Stephan R. Richter, René Ranftl, Zhuwen Li, Vladlen Koltun, and Thomas Brox. What do single-view 3d reconstruction networks learn? In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 3405–3414. Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPR.2019.00352. URL http://openaccess.thecvf.com/content_CVPR_2019/html/Tatarchenko_What_Do_Single-View_3D_Reconstruction_Networks_Learn_CVPR_2019_paper.html. 7

[68] Ayush Tewari, Tianwei Yin, George Cazenavette, Semon Rezchikov, Josh Tenenbaum, Frédo Durand, Bill Freeman, and Vincent Sitzmann. Diffusion with forward models: Solving stochastic inverse problems without direct supervision. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/28e4ee96c94e31b2d040b4521d2b299e-Abstract-Conference.html. 4, 21

[69] Anh Thai, Stefan Stojanov, Vijay Upadhya, and James M. Rehg. 3d reconstruction of novel object shapes from single images, 2020. 3

[70] Garvita Tiwari, Bharat Lal Bhatnagar, Tony Tung, and Gerard Pons-Moll. SIZER: A dataset and model for parsing 3d clothing and learning size sensitive 3d clothing. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part III*, volume 12348 of *Lecture Notes in Computer Science*, pages 1–18. Springer, 2020. doi: 10.1007/978-3-030-58580-8_1. URL https://doi.org/10.1007/978-3-030-58580-8_1. 7, 8, 10, 24, 34, 35, 37

[71] Garvita Tiwari, Nikolaos Sarafianos, Tony Tung, and Gerard Pons-Moll. Neural-gif: Neural generalized implicit functions for animating people in clothing. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 11688–11698. IEEE, 2021. doi: 10.1109/ICCV48922.2021.01150. URL https://doi.org/10.1109/ICCV48922.2021.01150. 2

[72] Dmitry Tochilkin, David Pankratz, ZeXiang Liu, Zixuan Huang, Adam Letts, Yangguang Li, Ding Liang, Christian Laforte, Varun Jampani, and Yan-Pei Cao. Triposr: Fast 3d object reconstruction from a single image. *CoRR*, abs/2403.02151, 2024. doi: 10.48550/ARXIV.2403.02151. URL https://doi.org/10.48550/arXiv.2403.02151. 2, 3, 5, 7, 8, 33

[73] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html. 3

[74] Vikram Voleti, Chun-Han Yao, Mark Boss, Adam Letts, David Pankratz, Dmitry Tochilkin, Christian Laforte, Robin Rombach, and Varun Jampani. SV3D: Novel Multi-view Synthesis and 3D Generation from a Single Image using Latent Video Diffusion, March 2024. 3

[75] Peng Wang and Yichun Shi. Imagedream: Image-prompt multi-view diffusion for 3d generation. *CoRR*, abs/2312.02201, 2023. doi: 10.48550/ARXIV.2312.02201. URL https://doi.org/10.48550/arXiv.2312.02201. 2, 3, 5, 6, 7, 21, 38

[76] Z. Wang, E.P. Simoncelli, and A.C. Bovik. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, volume 2, pages 1398–1402 Vol.2, 2003. doi: 10.1109/ACSSC.2003.1292216. 7

[77] Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *arXiv preprint arXiv:2305.16213*, 2023. 3

[78] Chung-Yi Weng, Brian Curless, Pratul P. Srinivasan, Jonathan T. Barron, and Ira Kemelmacher-Shlizerman. HumanNeRF: Free-viewpoint rendering of moving people from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16210–16220, June 2022. 2

[79] Jiajun Wu, Chengkai Zhang, Xiuming Zhang, Zhoutong Zhang, William T Freeman, and Joshua B Tenenbaum. Learning 3D Shape Priors for Shape Completion and Reconstruction. In *European Conference on Computer Vision (ECCV)*, 2018. 3

[80] Tong Wu, Jiarui Zhang, Xiao Fu, Yuxin Wang, Jiawei Ren, Liang Pan, Wayne Wu, Lei Yang, Jiaqi Wang, Chen Qian, Dahua Lin, and Ziwei Liu. Omniobject3d: Large-vocabulary 3d object dataset for realistic perception, reconstruction and generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 803–814. IEEE, 2023. doi: 10.1109/CVPR52729.2023.00084. URL https://doi.org/10.1109/CVPR52729.2023.00084. 2, 3, 34

[81] Yongqin Xian, Julian Chibane, Bharat Lal Bhatnagar, Bernt Schiele, Zeynep Akata, and Gerard Pons-Moll. Any-shot gin: Generalizing implicit networks for reconstructing novel classes. In *2022 International Conference on 3D Vision (3DV)*. IEEE, 2022. 3

[82] Xianghui Xie, Bharat Lal Bhatnagar, and Gerard Pons-Moll. CHORE: contact, human and object reconstruction from a single RGB image. *CoRR*, abs/2204.02445, 2022. doi: 10.48550/ARXIV.2204.02445. URL https://doi.org/10.48550/arXiv.2204.02445. 2

[83] Xianghui Xie, Bharat Lal Bhatnagar, and Gerard Pons-Moll. Visibility aware human-object interaction tracking from single RGB camera. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 4757–4768. IEEE, 2023. doi: 10.1109/CVPR52729.2023.00461. URL https://doi.org/10.1109/CVPR52729.2023.00461. 2

[84] Xianghui Xie, Bharat Lal Bhatnagar, Jan Eric Lenssen, and Gerard Pons-Moll. Template free reconstruction of human-object interaction with procedural interaction generation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024. 2, 9, 26, 32

[85] Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J. Black. ICON: implicit clothed humans obtained from normals. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 13286–13296. IEEE, 2022. doi: 10.1109/CVPR52688.2022.01294. URL https://doi.org/10.1109/CVPR52688.2022.01294. 2, 34, 35, 37

[86] Yuliang Xiu, Jinlong Yang, Xu Cao, Dimitrios Tzionas, and Michael J. Black. ECON: explicit clothed humans optimized via normal integration. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 512–523. IEEE, 2023. doi: 10.1109/CVPR52729.2023.00057. URL https://doi.org/10.1109/CVPR52729.2023.00057. 2, 34, 35

[87] Jiale Xu, Weihao Cheng, Yiming Gao, Xintao Wang, Shenghua Gao, and Ying Shan. Instantmesh: Efficient 3d mesh generation from a single image with sparse-view large reconstruction models. *arXiv preprint arXiv:2404.07191*, 2024. 3, 5, 7, 8

[88] Yinghao Xu, Hao Tan, Fujun Luan, Sai Bi, Peng Wang, Jiahao Li, Zifan Shi, Kalyan Sunkavalli, Gordon Wetzstein, Zexiang Xu, and Kai Zhang. DMV3D: denoising multi-view diffusion using 3d large reconstruction model. *CoRR*, abs/2311.09217, 2023. doi: 10.48550/ARXIV.2311.09217. URL https://doi.org/10.48550/arXiv.2311.09217. 3, 21

[89] Yinghao Xu, Zifan Shi, Wang Yifan, Sida Peng, Ceyuan Yang, Yujun Shen, and Wetzstein Gordon. Grm: Large gaussian reconstruction model for efficient 3d reconstruction and generation. *arxiv: 2403.14621*, 2024. 3

[90] Yuxuan Xue, Haolong Li, Stefan Leutenegger, and Joerg Stueckler. Event-based non-rigid reconstruction from contours. In *33rd British Machine Vision Conference 2022, BMVC 2022, London, UK, November 21-24, 2022*, page 78. BMVA Press, 2022. URL https://bmvc2022.mpi-inf.mpg.de/78/. 2

[91] Yuxuan Xue, Bharat Lal Bhatnagar, Riccardo Marin, Nikolaos Sarafianos, Yuanlu Xu, Gerard Pons-Moll, and Tony Tung. NSF: neural surface fields for human modeling from monocular depth. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 15004–15014. IEEE, 2023. doi: 10.1109/ICCV51070.2023.01382. URL https://doi.org/10.1109/ICCV51070.2023.01382. 2

[92] Yuxuan Xue, Haolong Li, Stefan Leutenegger, and Jörg Stückler. Event-based non-rigid reconstruction of low-rank parametrized deformations from contours. In *International Journal of Computer Vision (IJCV)*. Springer Science and Business Media LLC, February 2024. doi: 10.1007/s11263-024-02011-z. URL http://dx.doi.org/10.1007/s11263-024-02011-z. 2

[93] Xueting Yang, Yihao Luo, Yuliang Xiu, Wei Wang, Hao Xu, and Zhaoxin Fan. D-IF: Uncertainty-aware Human Digitization via Implicit Distribution Field. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9088–9098, Paris, France, October 2023. IEEE. ISBN 9798350307184. doi: 10.1109/ICCV51070.2023.00837. URL https://ieeexplore.ieee.org/document/10377954/. 2

[94] Kim Youwang, Tae-Hyun Oh, and Gerard Pons-Moll. Paint-it: Text-to-texture synthesis via deep convolutional texture map optimization and physically-based rendering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[95] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 4578–4587. Computer Vision Foundation / IEEE, 2021. doi: 10.1109/CVPR46437.2021.00455. URL https://openaccess.thecvf.com/content/CVPR2021/html/Yu_pixelNeRF_Neural_Radiance_Fields_From_One_or_Few_Images_CVPR_2021_paper.html. 4, 21

[96] Tao Yu, Zerong Zheng, Kaiwen Guo, Pengpeng Liu, Qionghai Dai, and Yebin Liu. Function4d: Real-time human volumetric capture from very sparse consumer rgbd sensors. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR2021)*, June 2021. 7, 34, 36

[97] Xianggang Yu, Mutian Xu, Yidan Zhang, Haolin Liu, Chongjie Ye, Yushuang Wu, Zizheng Yan, Chenming Zhu, Zhangyang Xiong, Tianyou Liang, Guanying Chen, Shuguang Cui, and Xiaoguang Han. Mvimgnet: A large-scale dataset of multi-view images. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 9150–9161. IEEE, 2023. doi: 10.1109/CVPR52729.2023.00883. URL https://doi.org/10.1109/CVPR52729.2023.00883. 2, 3, 34

[98] Zehao Yu, Torsten Sattler, and Andreas Geiger. Gaussian opacity fields: Efficient high-quality compact surface reconstruction in unbounded scenes. *arXiv:2404.10772*, 2024. 5, 21, 23

[99] Polina Zablotskaia, Aliaksandr Siarohin, Bo Zhao, and Leonid Sigal. Dwnet: Dense warp-based network for pose-guided human video generation. In *30th British Machine Vision Conference 2019, BMVC 2019, Cardiff, UK, September 9-12, 2019*, page 51. BMVA Press, 2019. URL https://bmvc2019.org/wp-content/uploads/papers/1039-paper.pdf. 10, 26, 30

[100] Andy Zeng, Shuran Song, Matthias Nießner, Matthew Fisher, Jianxiong Xiao, and Thomas Funkhouser. 3dmatch: Learning local geometric descriptors from rgb-d reconstructions. In *CVPR*, 2017. 21, 23

[101] Chao Zhang, Sergi Pujades, Michael J. Black, and Gerard Pons-Moll. Detailed, accurate, human shape estimation from clothed 3d scan sequences. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 5484–5493. IEEE Computer Society, 2017. doi: 10.1109/CVPR.2017.582. URL https://doi.org/10.1109/CVPR.2017.582. 34, 35, 37

[102] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 586–595. Computer Vision Foundation / IEEE Computer Society, 2018. doi: 10.1109/CVPR.2018.00068. URL http://openaccess.thecvf.com/content_cvpr_2018/html/Zhang_The_Unreasonable_Effectiveness_CVPR_2018_paper.html. 7

[103] Xiaohan Zhang, Bharat Lal Bhatnagar, Sebastian Starke, Vladimir Guzov, and Gerard Pons-Moll. COUCH: towards controllable human-chair interactions. In Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part V*, volume 13665 of *Lecture Notes in Computer Science*, pages 518–535. Springer, 2022. doi: 10.1007/978-3-031-20065-6_30. URL https://doi.org/10.1007/978-3-031-20065-6_30. 2

[104] Xiaohan Zhang, Bharat Lal Bhatnagar, Sebastian Starke, Ilya Petrov, Vladimir Guzov, Helisa Dhamo, Eduardo Pérez-Pellitero, and Gerard Pons-Moll. FORCE: dataset and method for intuitive physics guided human-object interaction. *CoRR*, abs/2403.11237, 2024. doi: 10.48550/ARXIV.2403.11237. URL https://doi.org/10.48550/arXiv.2403.11237. 2

[105] Xiuming Zhang, Zhoutong Zhang, Chengkai Zhang, Joshua B Tenenbaum, William T Freeman, and Jiajun Wu. Learning to Reconstruct Shapes From Unseen Classes. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. 3

[106] Zechuan Zhang, Zongxin Yang, and Yi Yang. SIFU: side-view conditioned implicit function for real-world usable clothed human reconstruction. *CoRR*, abs/2312.06704, 2023. doi: 10.48550/ARXIV.2312.06704. URL https://doi.org/10.48550/arXiv.2312.06704. 2, 7, 8, 33, 34, 35

[107] Zhizhuo Zhou and Shubham Tulsiani. Sparsefusion: Distilling view-conditioned diffusion for 3d reconstruction. In *CVPR*, 2023. 3

[108] Zi-Xin Zou, Zhipeng Yu, Yuan-Chen Guo, Yangguang Li, Ding Liang, Yan-Pei Cao, and Song-Hai Zhang. Triplane meets gaussian splatting: Fast and generalizable single-view 3d reconstruction with transformers. *CoRR*, abs/2312.09147, 2023. doi: 10.48550/ARXIV.2312.09147. URL https://doi.org/10.48550/arXiv.2312.09147. 2, 3

Appendix

Table of Contents

A Implementation Details

A.1 Training Details

As described in Sec. 5.1, we use an effective batch size of 256. Each batch involved sampling 4 orthogonal images with zero elevation angle as target views $\mathbf{x}_0^{\text{tgt}}$, and 12 additional images as novel views $\mathbf{x}_0^{\text{novel}}$ to supervise the 3D generative model Eq. (8). The hyperparameters for training Eq. (8) were set as follows: $\lambda_1 = 1.0$, $\lambda_2 = 1.0$, and $\lambda_3 = 100.0$.

During training, we employed the standard DDPM scheduler [23] to construct noisy target images $\mathbf{x}_t^{\text{tgt}}$. The maximum diffusion step T is set to 1000. At inference time, we use DDIM scheduler [61] to perform faster reverse sampling. The reverse steps is set to 50 in following experiments. The text prompt y used in our multi-view diffusion model(Eq. (3)) is set to "Photorealistic 3D human" for both training and inference across all subjects.

A.2 Joint Framework

Implementation. Our 2D multi-view diffusion model $\boldsymbol{\epsilon}_\theta$ is a latent diffusion model [75]. Thus, we use the frozen VAE in [53] to obtain input $\mathbf{x}_t^{\text{tgt}}$ in image space for the 3D generative model g_ϕ and encode refined $\tilde{\mathbf{x}}_t^{\text{tgt}}$ back to latent space for $\boldsymbol{\epsilon}_\theta$. We extract triangle mesh from predicted Gaussian splats using Gaussian Opacity field [98] and TSDF [100]. Please refer to Appendix A.4 for more details.

A.3 Generative 3D-GS Reconstruction Model

In this section, we provide details about our 3D generative model g_ϕ in Eq. (5) and Eq. (7) as well as the `renderer`$(\circ)$. Following [7, 33, 68], we learn the 3D generative model by adding and removing noise on the rendered 2D images from a 3D representation. A pseudo algorithm of the training and sampling process of our 3D generative model can be found in Algorithm 3.

Since we integrate both function into the reverse sampling process iteratively (eq. 2), we expect them be efficient and fast to execute. Tewari et al. [68] base their model on pixelNeRF [95], which is a generalizable NeRF [47] conditioned on a context view image. We adopt 3D Gaussian Splats [34] as our 3D state representation $\mathcal{G}$ due to its efficiency and simplicity. Our `renderer`$(\circ)$ is the differentiable rasteraizer accelerated and implemented in CUDA, which achieves around 2700 times faster rendering than volume-rendering-based `renderer`$(\circ)$ in [47, 68, 95].

For sampling the 3D State $\mathbf{S}$ from $\mathbf{x}_t^{\text{tgt}}$, $\tilde{\mathbf{x}}_0^{\text{tgt}}$, $\mathbf{x}^{\text{c}}$, and t (eq. 7), we adopt the time-conditioned UNet-Transformer architecture [53] due to the efficiency of convolutional layers and the scalability of transformers. For enabling the awareness of camera poses in the encoding process, we concatenate the Plücker Camera Ray Embedding $\{\mathbf{o}_i \times \mathbf{d}_i, \mathbf{d}_i\}$ [65, 88] with the image $\mathbf{x}_t^{\text{tgt}}$ and $\tilde{\mathbf{x}}_0^{\text{tgt}}$. To enhance the control ability of context view in the 3D generation process, we additionally concatenate the clear context view $\mathbf{x}^{\text{c}}$ with target images $\mathbf{x}_t^{\text{tgt}}$ follow [75]. This operation enables 3D dense self-attention process between the input multi-view target images and the clear context view image, provides pixel-level local conditional signal. Since the camera pose of context view $\mathbf{x}^c$ is unknown, we use the 0-vector as its embedding.

Algorithm 3 Learn 3D distribution

Input: Dataset of posed multi-view images $\mathbf{x}_0^{\text{tgt}}$, π^{tgt}, $\mathbf{x}_0^{\text{novel}}$, π^{novel}, a context image $\mathbf{x}^{\text{c}}$
Output: Optimized 3D State diffusion network g_ϕ

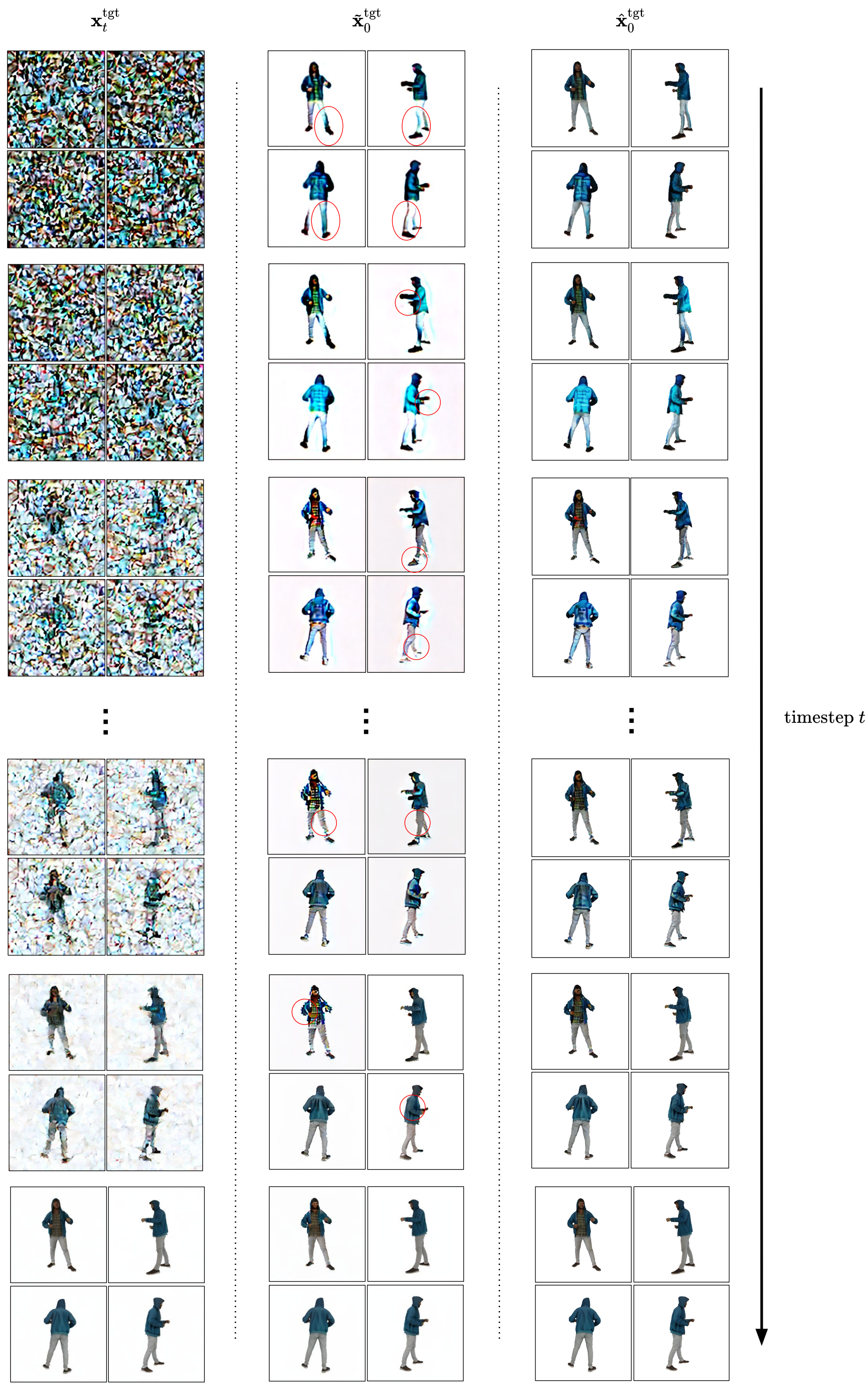
```
repeat
  {x_0^tgt, x_0^novel, x^c, y} ~ q({x_0^tgt, x_0^novel, x^c, y})
  t ~ Uniform({1, ..., T}); ε ~ N(0, I)
  x_t^tgt = sqrt(ᾱ_t) x_0^tgt + sqrt(1 − ᾱ_t) ε
  Ĝ = g_φ(x^c, x_t^tgt, t)
  {x̂_0^tgt, x̂_0^novel} = renderer(Ĝ, {π^tgt, π^novel})
  Compute loss L_gs ( Eq. (6))
  Gradient step to update g_φ
until converged
```

Algorithm 4 Sample from 3D distribution

Input: A context image $\mathbf{x}^c$; Converged 3D diffusion model g_ϕ
Output: A 3D Gaussian Avatar $\mathcal{G}$ of the 2D image $\mathbf{x}^c$

```
x_T^tgt ~ N(0, I)
for t = T, ..., 1 do
  Ĝ = g_φ(x^c, x_t^tgt, t)
  x̂_0^tgt = renderer(Ĝ, π^tgt)
  μ_{t−1}(x_t^tgt, x̂_0^tgt) = (sqrt(α_t)(1−ᾱ_{t-1}) / (1−ᾱ_t)) x_t^tgt + (sqrt(ᾱ_{t-1}) β_t / (1−ᾱ_t)) x̂_0^tgt
  x_{t−1}^tgt ~ N(x_{t−1}^tgt; μ̃_t(x_t^tgt, x̂_0^tgt), β̃_{t−1} I)
end for
return G = g_φ(x_0^tgt, x^c, t = 0)
```

Figure 7: Visualization intermediate sampling steps from a Gaussian Noise ($t = 1000$) to the last denoising step ($t = 0$). From top to bottom: current state $\mathbf{x}_t^{\text{tgt}}$, estimated clear view by 2D diffusion models $\tilde{\mathbf{x}}_0^{\text{tgt}}$, and corrected clear view by generated 3D Gaussian Splatting $\hat{\mathbf{x}}_0^{\text{tgt}}$. Our 2D diffusion model $\boldsymbol{\epsilon}_\phi(\circ)$ already provides strong multi-view prior at an early stage with large t. Our 3D reconstruction model $\mathbf{g}_\phi(\circ)$ can correct the inconsistency in $\tilde{\mathbf{x}}_0^{\text{tgt}}$ illustrated in red circle.



A.4 Textured Mesh Extraction

Gaussian Opacity Fields [98] enables extraction of triangle meshes from an existing 3D Gaussian Splatting. However, because the location of 3D-Gaussian Splats is not necessary to be on the real surface, we observe that the extracted meshes as well as the rendered depth maps are noisy. Since our method generate realistic RGB images, we use PiFU-HD [55] to estimate the normals and use Bilateral Normal Integration (BiNI) [10] to refine the noisy rendered depth with the estimated normal. As we only want the estimated normal to denoise the rendered depth map instead of modifying geometry, we set up the hyperparameter in BiNI with $\lambda = 1 \times 10^4$. Such a large number ensures that the normal map is not used to modify the geometry but just regularize the depth map.

Assuming we have a generated 3D-Gaussian Splats $\mathcal{G}$ from $g_\phi(\circ)$ and n camera views $\pi^1, \pi^2, ...\pi^n$, we obtain n paris of posed RGB-D images by *Gaussian Splatting*, *Normal Estimation*, and *Bilateral Normal Integration*. Finally, we perform volumetric TSDF fusion [100] to obtain high quality textured mesh from n pairs RGB-D images. Given generated 3D-GS $\mathcal{G}$, we set up 36 views to obtain the refined RGB-D image pairs. The rendering view of each camera i can be calculated as:

$$\text{elevation}_i = -\frac{1}{4}\pi + \frac{1}{4}\pi * \frac{i}{36}, \tag{10}$$

$$\text{azimuth}_i = 0 + 3\pi * \frac{i}{36}. \tag{11}$$

B Comparison

B.1 Qualitative Comparison

Figure 8: Qualitative comparison on Sizer [70] and IIIT [30].

Figure 9: Qualitative comparison on IIIT [30].

C More Qualitative Results

In this section, we show more qualitative results on in-the-wild data, UBC fashion dataset [99], GSO dataset [16], and human-object interaction data [84].

C.1 In-the-wild Data

Figure 10: Qualitative results on unseen data during training. Input image is in left column. Our method successfully reconstructs different degree of loose clothing.

Figure 11: Qualitative results on more unseen data during training. Input image is in left column. Our method successfully reconstructs different types of clothing, including casual, sport, suits, custom, etc., in both appearance and geometry.

Figure 12: Qualitative results on more unseen data during training. Input image is in left column. Our method successfully reconstructs clothing and interacting objects (racket and bag here) in both appearance and geometry.

Figure 13: Qualitative results on more unseen data during training. Input image is in left column. Our method successfully reconstructs rarely seen suits and objects, in both appearance and geometry.

C.2 UBC Fashion Dataset

In this section, we show qualitative result of our model on UBC fashion [99] dataset. The input images are the first frame extracted from each video in the dataset.

Figure 14: Qualitative results on UBC fashion [99] dataset. Results demonstrate that our model generalizes well to real world images in both geometry and appearance.

C.3 Google Scan Objects (GSO)

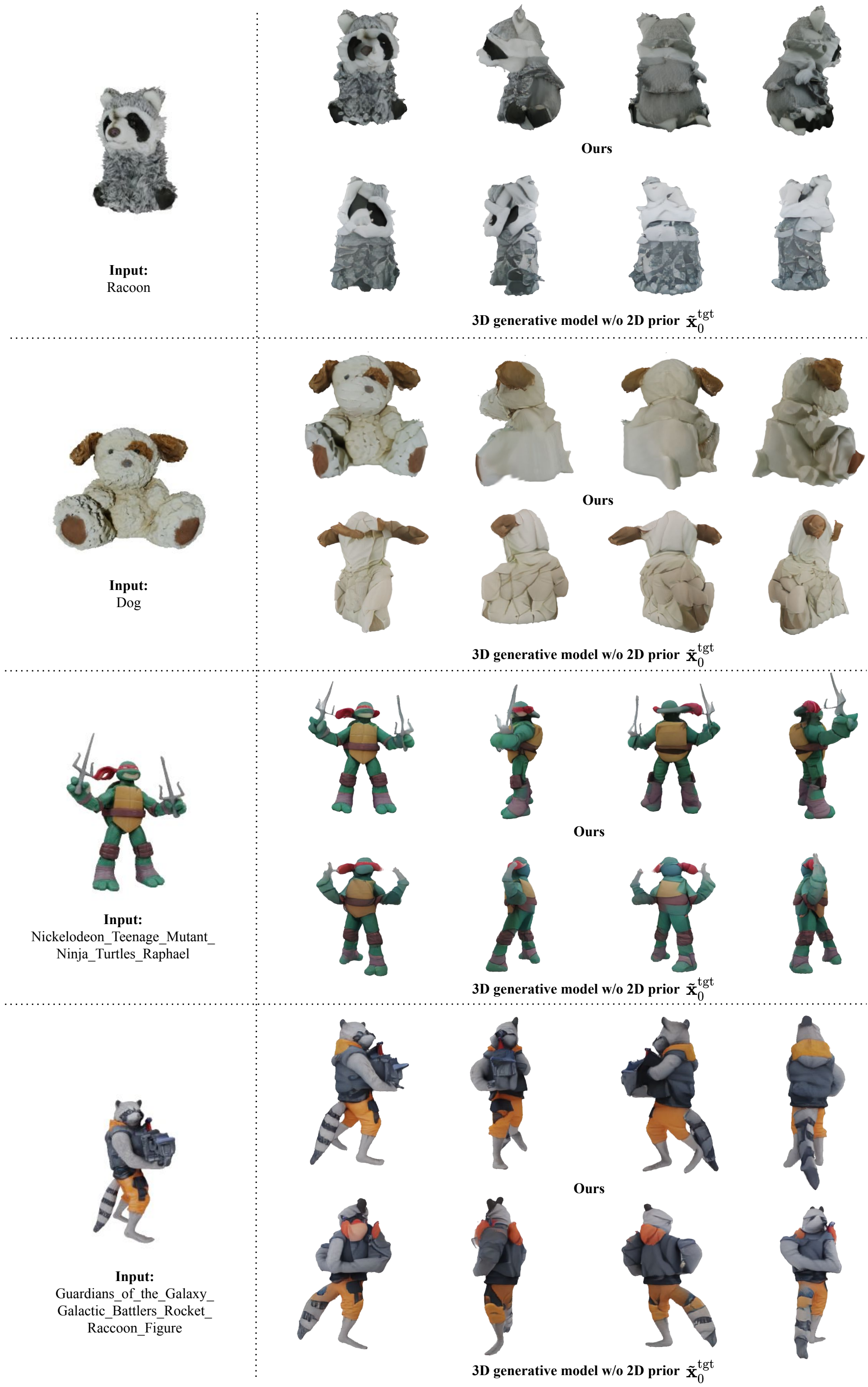


Figure 15: Ablation study: benefit of 2D multi-view prior $\tilde{\mathbf{x}}_0^{\mathrm{tgt}}$ in 3D generation. The 2D prior from 2D diffusion model is essntial for generalization on general objects dataset GSO [16].

C.4 Human Object Interaction

Figure 16: Qualitative results of Human-Object Interaction reconstruction on online stock images. Results show that our model is able to generalize to casual human-object-interactions.

Figure 17: Qualitative results of Human-Object Interaction reconstruction on ProciGen [84] dataset. Results demonstrate that our model can reconstruct some simple Human-Object-Interaction images with large objects.

C.5 Generative Power in Reconstruction

Our model learns a conditional distribution of the 3D representation given 2D context image. Thus, by sampling from the distribution with different seed, we obtain diverse yet plausible 3D representation. As illustrated in Fig. 18, the appearance of the occluded region (back side of subject) is different with different sampling in hair style, texture, and cloth wrinkles.

The generative power of our approach is the key to generate clear self-occluded regions, which is impossible by non-generative reconstruction methods [54, 55, 72, 106]. As shown in Fig. 3 and Fig. 8, non-generative approaches tend to generate blurry self-occluded results because they cannot sample from distribution but only regress to a mean value of the training datasets.

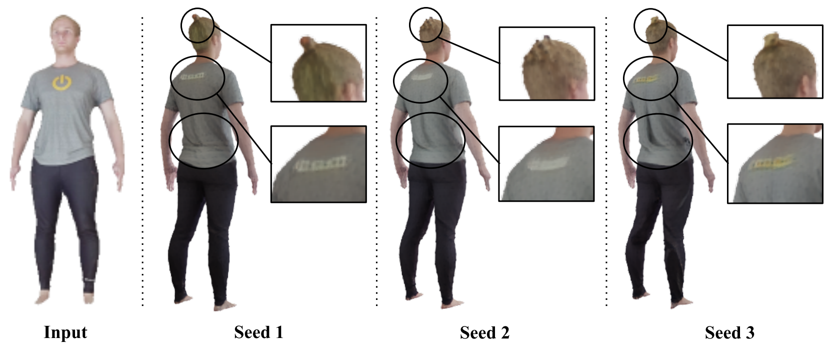


Figure 18: Our model learns 3D distribution. By different sampling from the learned distribution, we obtain diverse yet plausible 3D representations. The generative power is a key to generate clear self-occluded region, which is impossible in non-generative reconstruction approaches [54, 55, 72, 106].

D Dataset Overview

To ensure robust performance and generalization, we train our model on a combined dataset comprising 3520 scans from publicly available datasets [19, 25, 63, 96] and 2320 scans from commercial 3D human datasets [1, 3, 4, 2]. These datasets encompass a diverse range of body shapes, genders, ages, clothing, accessories, and interacting objects. For further details and examples of our training datasets, please refer to section D.1. All 3D scans are rendered into RGB-A images using BlenderProc [13] along a spiral path as described in Eq. (13).
For evaluation, we note that the commonly-used CAPE dataset [45, 50, 101] in previous works [21, 85, 86, 106] often contains artifacts in scans, such as holes, and not all 3D scans are fully publicly available. To effectively and fairly evaluate performance, we propose using Sizer [70] and IIIT-Human [30] datasets, from which we randomly sample 150 scans each for evaluation. While Sizer [70] provides scans with normal human appearance similar to our training datasets, IIIT-Human [30] can be considered as out-of-distribution (o.o.d.) evaluation dataset due to its inclusion of unseen clothing types, such as traditional Indian suits. For more additional examples and analysis, please see Fig. 3 and Appendix D.2.

D.1 Training Dataset

Datasets To prevent the overfitting of our large neural network, namely the 2D multi-view diffusion models $\boldsymbol{\epsilon}_\theta(\circ)$ and 3D generative models $\mathbf{g}_\phi(\circ)$, we train on as much data as we can. Unlike general objects community which has massive dataset such as Objaverse (800K) and Objaverse-XL (10M) [12]) OmniObject3D (6K) [80], MVImageNet (87K) [97], we don't have a single 3D human dataset available at such a scale. To collect data as much as possible, we collect both following public datasets and commercial human scans.

We collect several publicly available datasets inluding 2k2k (2K) [19], CustomHuman (640) [25], Thuman2.0 (520) [96], and Thuman3.0 (360) [63]). Among them, CustomHuman, Thuman2.0, and Thuman3.0 have more repeating subjects with different poses, which have less diverse subject appearance compared to 2k2k. It is worth mentioning that 2k2k [19] is a high quality dataset which contains human with diverse clothing (such as skirt) and accessories (such as cap, hat, scarf).

We also utilize in total 2320 high quality commercial scans from AXYZ [1], Treedy [3], Twindom [4], and RenderPeople [2]. All of these scans are with casual clothing and without interaction with objects.

Rendering For each scan, we render 100 views following a spiral trajectory with each view i:

$$\text{elevation}_i = -\frac{1}{4}\pi + \frac{7}{8}\pi * \frac{i}{100}, \tag{12}$$

$$\text{azimuth}_i = 0 + 5\pi * \frac{i}{100}. \tag{13}$$

Additionally, we render 32 views uniformly around z-axis with each view j:

$$\text{elevation}_j = 0, \tag{14}$$

$$\text{azimuth}_j = 0 + \pi * \frac{j}{32}. \tag{15}$$

To protect the privacy of subjects in the training dataset, we only use the frontal view (with $\text{azimuth}_j \in [-\frac{\pi}{2}, \frac{\pi}{2}]$) as the input context view during training. Thus, we expect the model will not learn faces of subjects when it takes the back view as input.

D.2 Evaluation Dataset

For quantitative evaluation, we use Sizer [70] and IIIT 3D human dataset [30]. In this section, we start with introducing the two used evaluation datasets, and explain why we omit the commonly used CAPE dataset [45, 50, 101] in our experiments. Finally, we provide a summary of the evaluation datasets.

Sizer Sizer [70] is a high quality 3D human scan dataset which contains 100 different subjects wearing casual clothing items in various sizes. We randomly sample 150 scans from Sizer [70] as one of our evaluation dataset.

IIIT IIIT 3D Humans [30] is a high quality dataset from IIIT Hyderabad in India. Different from the casual clothing setup in Sizer [70], IIIT dataset mainly focuses on subjects wearing traditional India custom suits, including ethnicity, diverse color pattern and extremely loose clothing (Fig. 21). It brings the huge variety of the subject appearance which can be considered as o.o.d. evaluation set to our model and baselines. We randomly sample 150 scans from IIIT dataset [30] for evaluation.

Why not CAPE Unlike previous methods [21, 85, 86, 106] which evaluate on CAPE dataset [45, 50, 101], we omitted CAPE in our evaluation mainly because of the limitation in appearance variety, geometrical artifacts as well as the publicly unavailability. CAPE only contains simple clothing such as T-shirts and jeans, but no garments of loose clothing. As illustrated in Fig. 22, CAPE contains several artifacts, such as holes on the head, missing hands, and wrong mesh geometry. Moreover, CAPE doesn't have most original scans publicly available, but only the SMPL+D fitting. Due to this reason, we cannot render the CAPE scan to RGB images at desired camera view to evaluate the appearance performance such as PSNR, SSIM, and LPIPS.

Summary We observe that the high quality datasets Sizer [70] and IIIT 3D Human [30] are unexplored for the community of 3D avatars reconstruction. In fact, Sizer [70] contains casual clothing which is suitable to evaluate performance, and IIIT [30] contains challenging texture and loose clothing which is suitable to evaluate robustness. All high quality scans in [30, 70] have no severe artifacts and are fully publicly available, which are the benefits unprovided in CAPE dataset [45, 50, 101]. By evaluating on these datasets [30, 70] and release our randomly sampled subjects which are used in our experiments, we hope the 3D avatars community can discover and benefit from them.

Figure 19: Example scans in training datasets [1–4, 19, 25, 63, 96].

Figure 20: Example scans in Sizer [70] dataset. Sizer contains human in casual clothing.

Figure 21: Example scans in IIIT [30] dataset. IIIT contains subjects with diverse color pattern and loose garments, which rarely appear in training datasets.

Figure 22: Example artifacts in CAPE [45, 50, 101] dataset. Images shown here are rendered by ICON [85] due to the inaccessibility of original scans.

E Failure Cases

Limited by low resolution (256 × 256) of our multi-view diffusion model [75], our model can often fail in reconstructing fine details such as text on the cloth as illustrated in Fig. 23. One potential solution is to switch to a recent powerful high-resolution multi-view diffusion models [17, 66].

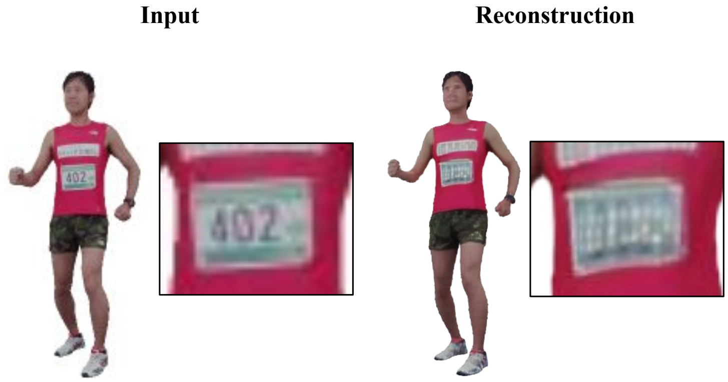


Figure 23: Failure Case: our model cannot reconstruct the numbers on the cloth.

In addition, we observe that our model can fail when reconstructing human extremely challenging poses. As shown in Fig. 24, our model cannot infer head geometry and appearance accurately due to the challenging pose in input image.

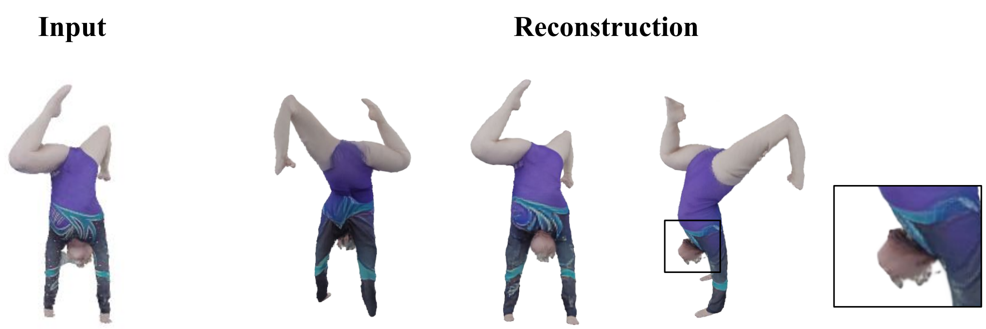


Figure 24: Failure Case: our model fails in infer appearance of human with challenging pose.

F Broader Impacts

Our work shows generality across different ethnicities and humans, providing a useful tool for a fair representation of different cultures. Having a robust method to synthesize realistic 3D geometry from a single RGB image may be used in surveillance and inappropriate content generation.