MoGA: 3D Generative Avatar Prior for Monocular Gaussian Avatar Reconstruction

Zijian Dong[1,4*] Longteng Duan[1*] Jie Song[3] Michael J. Black[4] Andreas Geiger[2]
[1]ETH Zürich, Department of Computer Science [2]University of Tübingen, Tübingen AI Center
[3]HKUST(GZ)&HKUST [4]Max Planck Institute for Intelligent Systems, Tübingen

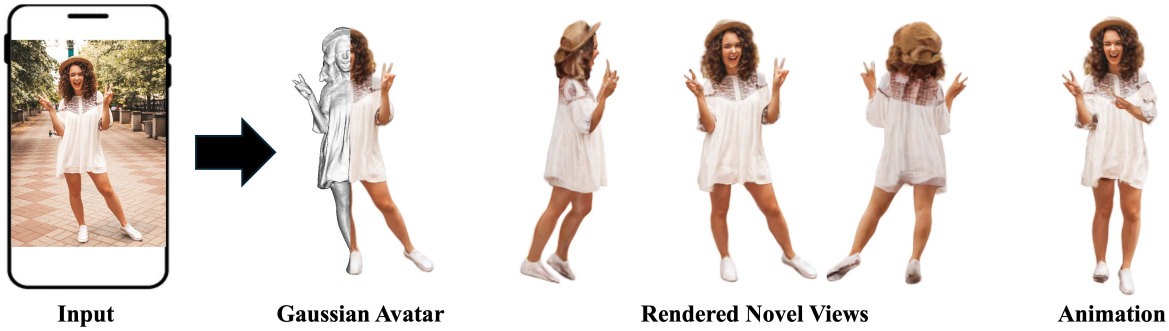

Figure 1. We propose MoGA, a method to genereate high-fidelity Gaussian avatars from a single image. Left: A challenging in-the-wild example. Middle: Unlike previous methods that struggle with such cases (Fig. 4), MoGA enables 3D-consistent full-body novel view synthesis and detailed geometry extraction. Right: Our reconstructed Gaussian avatar supports animation without any post-processing.

Abstract

We present MoGA, a novel method to reconstruct high-fidelity 3D Gaussian avatars from a single-view image. The main challenge lies in inferring unseen appearance and geometric details while ensuring 3D consistency and realism. Most previous methods rely on 2D diffusion models to synthesize unseen views; however, these generated views are sparse and inconsistent, resulting in unrealistic 3D artifacts and blurred appearance. To address these limitations, we leverage a generative avatar model, that can generate diverse 3D avatars by sampling deformed Gaussians from a learned prior distribution. Due to the limited amount of 3D training data such a 3D model alone cannot capture all image details of unseen identities. Consequently, we integrate it as a prior, ensuring 3D consistency by projecting input images into its latent space and enforcing additional 3D appearance and geometric constraints. Our novel approach formulates Gaussian avatar creation as a model inversion process by fitting the generative avatar to synthetic views from 2D diffusion models. The generative avatar provides a meaningful initialization for model fitting, enforces 3D regularization, and helps in refining pose estimation. Experiments show that our method surpasses state-of-the-art techniques and generalizes well to real-world scenarios. Our Gaussian avatars are also inherently animatable.

1. Introduction

Animatable and realistic avatar creation enables many applications in AR/VR, movies, and the gaming industry. It is hard to scale up this process since previous traditional approaches [22] require expensive multi-view systems and highly-specialized expertise to craft the avatar. To make digital avatars widely available and consumer-friendly, it is essential to develop methods for creating an avatar from in-the-wild images. However, this problem is very challenging due to the ill-posed nature of the monocular setting, which causes ambiguities in the appearance, depth, and body poses.

Embracing the challenging problem, previous methods [41, 42, 52] learn a large network to predict explicit or implicit 3D representations from 2D pixel-aligned features. These methods are trained on small-scale datasets due to the limited amount of available 3D data; this greatly restricts their generalization to diverse human poses and clothing styles. More recently, to enable more powerful generalization ability, some methods [14, 26, 53, 59] leverage a multi-view diffusion model to hallucinate back and side views. Despite impressive performance, most existing multi-view diffusion models can only generate very sparse views with high resolution due to memory constraints during training [23]. The sparsity of generated views leads to artifacts in self-occluded regions or unobserved side views. Furthermore, since such methods rely heavily on 2D priors

[*] Equal contribution

from 2D diffusion, the generated multi-view imagery often lacks 3D consistency [27, 38], which results in blurry appearance in 3D. To prevent the reconstruction of unnatural bodies, some methods [8, 14, 20, 26, 59] leverage parametric body models like SMPL [29] as full-body priors. Although this helps avoid abnormal shapes, it is restricted by the fixed topology and minimally-clothed SMPL body shape and cannot provide a 3D appearance prior.

To address these limitations, we propose a novel method, MoGA (Moncular Gaussian Avatar), that reconstructs a high-fidelity 3D Gaussian avatar from a single-view image (Fig. 1). At its core, our approach leverages a generative 3D avatar model as a powerful human body prior. Unlike the SMPL body prior, our model captures not only detailed geometry but also realistic human appearance, including hair and clothing, using deformed Gaussians. We harness multi-view diffusion to infer unseen views while ensuring 3D consistency and realism. This is achieved by projecting the synthetic images back to the learned latent space of the generative avatar and applying additional constraints. The creation of the Gaussian avatar is then formulated by fitting the generative avatar to these generated views. During this process, our generative avatar model plays a pivotal role in three ways: (i) **Initialization:** Sampling from the learned avatar prior enables a meaningful initialization of geometry and appearance for fitting, helping avoid local minima in few-shot reconstruction. (ii) **Regularization:** Rather than relying solely on inconsistent synthetic images, the avatar prior enforces strong 3D constraints, ensuring view consistency and preventing unrealistic artifacts. (iii) **Pose Optimization:** This avatar prior enables us to refine human and camera poses through an effective photometric rendering loss, improving alignment accuracy and reconstruction fidelity over baselines.

More specifically, we represent human bodies and clothing in canonical space using 2D Gaussian Splatting [19], anchored to a parametric body template [35], and integrate it with an efficient deformation module [11]. To enable generation, we model every human subject via a per-subject latent code and a shared decoder to interpret the latent code into Gaussian features. This generative avatar prior is learned through a single-stage pipeline [7] that jointly optimizes the latent code, the shared decoder, and a latent diffusion model in the latent space. At test time, we first employ image-guided sampling [7] to obtain a meaningful latent code. Given this initialization, we perform model inversion to compute the latent code for a novel target identity while freezing the decoder and learned diffusion model. During this process, the diffusion model serves as a constraint in the latent space via a score distillation sampling loss [37]. Throughout the fitting procedure, both the avatar model and camera/human pose parameters are optimized in an alternating manner to correct abnormal poses.

We experimentally demonstrate that our method significantly outperforms previous state-of-the-art methods both quantitatively and qualitatively (Table 1 and Fig. 3). The resulting Gaussian avatar has better 3D consistency and realism (Fig. 3), and adapts better to model complex structures like hair (Fig. 5). In summary, we contribute:

- An optimization-based model fitting framework to reconstruct a Gaussian avatar from a single image, by fitting a 3D generative avatar to synthetic images generated by 2D multi-view diffusion.
- A generative 3D Gaussian avatar prior that enables reconstructing a deformed Gaussian avatar from sparse and inconsistent generated images, by providing crucial support for meaningful initialization, regularization, and pose refinement during model fitting.
- Generalization to in-the-wild outdoor images with challenging pose and clothing. The resulting avatar can be animated without post-processing.

Code and models are available at https://zj-dong.github.io/MoGA/.

2. Related Work

2.1. Single-view Human Reconstruction

Creating realistic avatars from a *single RGB image* is a challenging problem. PIFu [41] pioneered a data-driven pipeline to learn a mapping from 2D pixel-aligned features to 3D implicit functions. More recent work builds upon this idea and improves the geometry by leveraging normal guidance and parametric human body models [3, 27, 28, 42, 52]. However, given only a frontal view, these methods struggle to reconstruct realistic full-body texture because of the unobserved back side. More recently, methods use 2D diffusion models to hallucinate the back view and incorporate this into the reconstruction pipeline [14, 59]. However, with only two observations, these methods suffer from strong artifacts in side-views. Several methods go further by leveraging a modified multi-view diffusion process to generate more views, improving rendering quality and geometry [26, 53, 59]. Unfortunately, 2D diffusion often produces synthetic images that are inconsistent in 3D. To address this, PSHuman [26] and SIFU [59] leverage SMPL [29, 35] as a 3D template to regularize the reconstruction. However, since SMPL does not provide an appearance prior and only has a minimally clothed body shape, the methods struggle when the clothing and hair differ from the SMPL body topology. Human3DDiffusion [53] uses a 3D diffusion model to guide sampling of a multi-view diffusion model, but is limited by the low resolution of the diffusion model. In contrast to the prior work, we combine an expressive generative 3D avatar model with synthetic 2D images generated from a multi-view diffusion model, achieving better reconstruction quality and robustness to self-occlusion.

2.2. 3D Avatar Generation

Several methods [1, 11, 16, 32] leverage 3D-aware GANs [5, 6, 44] to generate 3D humans from 2D image collections. The main idea is to leverage 3D human models [29, 35] to learn a 3D human GAN with an adversarial loss. Several techniques are then developed to improve geometric quality [11], the face region [11, 16], deformation [11] and efficiency [1]. Despite impressive results, these methods are all trained with 2D image discriminators that are unable to reason about cross-view relationships [7], making it challenging to exploit multi-view data. Recent 3D diffusion models [7, 15, 46, 48] show better generation capabilities due to their more expressive and high-dimensional latent space. Leveraging this, recent work [9, 18, 58] parameterizes the human via primitives [9] or a structured latent code [18] and learns a diffusion model in the latent space for unconditional generation. By leveraging multi-view data, these methods acheive 3D consistency in appearance generation. Unlike these methods we leverage a 3D generative avatar model for single-view avatar reconstruction. Although reconstruction from single images is possible through GAN inversion [6], the quality is limited by the expressiveness of the latent space, making it hard to apply to in-the-wild images. Rodin [50] and its extensions [57] employ an image-conditioned diffusion model for few-shot face reconstruction. These are trained on synthetic images and struggle to generalize to the real-world. We address the more challenging problem of reconstructing full-body avatars with diverse poses and clothing styles from in-the-wild images.

2.3. Gaussian Avatar

Representations for 3D avatars include 3D meshes [2, 29, 35], implicit functions [10, 12, 21, 30, 36, 51], and point clouds [49, 62]. Recently, 3D Gaussian Splatting [19, 24] has gained attention due to its high rendering quality and efficiency. Many methods [17, 31, 39, 56, 60] learn 3D Gaussian avatars from monocular videos. Despite strong performance, these methods typically fail when the number of observations becomes sparse. GPS-Gaussian [61] proposes a generalizable multi-view human Gaussian model with high-quality rendering, but it needs relatively dense views (16) and accurate camera poses. Concurrent methods [8, 33] leverage a pre-trained transformer to predict the 3D Gaussians from a single-view image, but struggle to reconstruct details and tend to generate artifacts on faces. In contrast, our method achieves a more detailed and realistic Gaussian avatar from only a single-view image. A detailed comparison with [8, 33] is not possible since the models or code have not been released at the time of writing.

3. Method

Given a single-view image, our goal is to reconstruct a high fidelity 3D Gaussian avatar. To address this ill-posed problem, our key idea is to leverage a generative 3D avatar model as a human prior and fit this generative model to synthetic images generated by multi-view diffusion. An overview of our method is shown in Fig. 2.

We first introduce an efficient and articulation-aware 3D human generator (Section 3.1 and Fig. 2(a)), which generates the appearance and shape in canonical space and leverages a deformation module to deform these into posed space. To learn this generator, we utilize a single-stage training pipeline [7] that jointly optimizes a Gaussian autodecoder and a latent diffusion model.

At test time, we fit the learned generative avatar model to 6 synthetic images generated using a pre-trained multi-view diffusion model (Section 3.2 and Fig. 2(b)). We show that the learned generative prior improves the performance by providing a good initialization and regularization for fitting to handle the inconsistency between synthetic views. During the fitting process, we optimize both the avatar model and the camera/human pose parameters alternately to correct abnormal poses.

3.1. Generative Avatar Prior Training

3.1.1. Canonical Gaussian Representation

To achieve high quality reconstruction of both appearance and geometry, we employ 2D Gaussian splatting [19] to represent the appearance and geometry of the avatar generated in the canonical space. Motivated by GGHead [25] and Relightable Gaussian Codec Avatars [43], we parameterize 2D Gaussians $\mathcal{G}$ on a UV map U of a template mesh [35]. Here, each Gaussian primitive $\mathcal{G}_k$ is parameterized by five attributes: an opacity $\sigma_k \in \mathbb{R}$, a Gaussian center $\mu_k \in \mathbb{R}^3$, RGB color $c_k \in \mathbb{R}^3$ for simplicity, a scale vector $s_k \in \mathbb{R}^2$ for 2D Gaussian Splatting, and a rotation matrix R_k represented by the axis angle vector $r_k \in \mathbb{R}^3$. Finally, based on the UV mapping, we can represent the Gaussian attributes $\mathcal{G}$ with a 2D UV map $U \in \mathbb{R}^{256\times256\times12}$.

To better leverage the template body prior, we model the Gaussian center μ_k, the scale vector s_k, and the rotation r_k as a residual from the canonical SMPL-X body [35]:

$$\begin{aligned} \mu_k &= \hat{\mu}_k + \delta_{\mu k} \\ s_k &= \hat{s}_k \cdot \delta_{sk} \\ r_k &= \hat{r}_k \cdot \delta_{rk}. \end{aligned} \tag{1}$$

Here, $\hat{\mu}_k$, $\hat{s}_k$, and $\hat{r}_k$ are the initial center, scale, and rotation of the Gaussian primitives, which are obtained from the SMPL-X mesh; for more details, see Sup. Mat. Finally, we predict the offset value $\delta_{\mu k}$, δ_{sk}, δ_{rk} to represent Gaussian attributes.

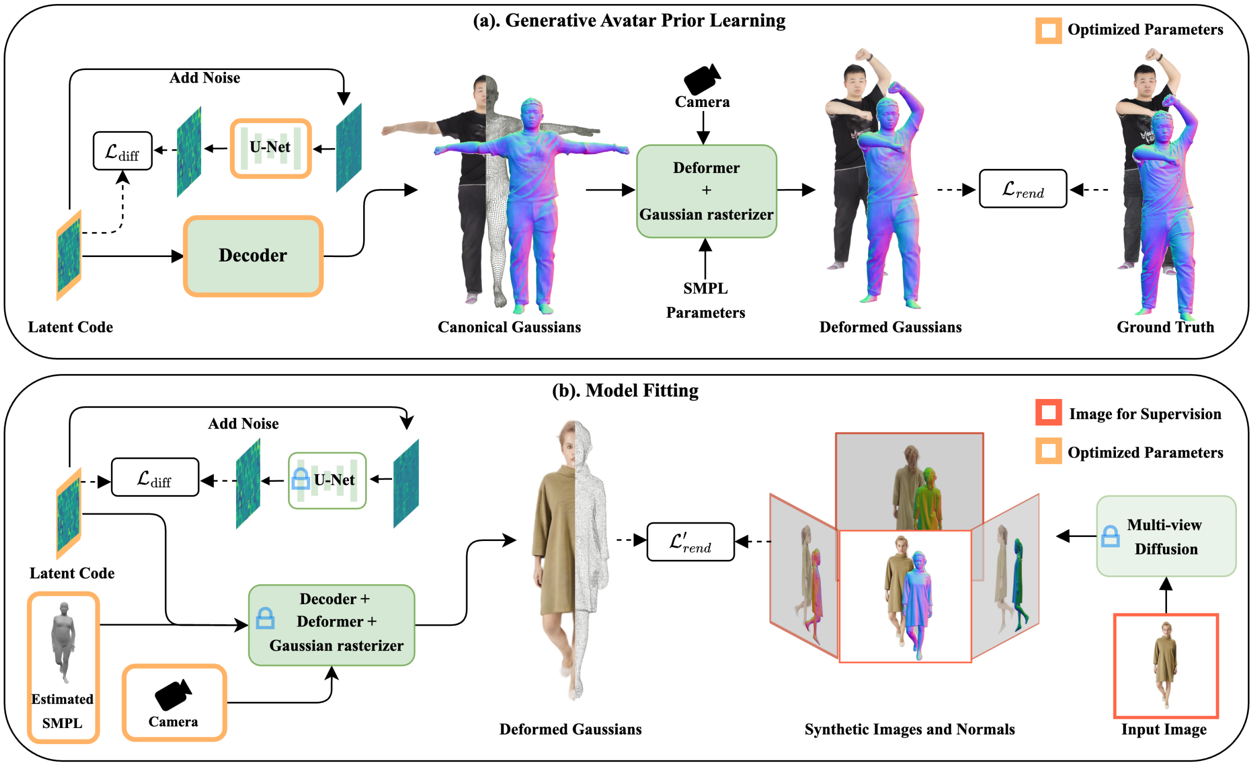

Figure 2. **Method Overview.** *Generative Avatar Prior Learning:* Our 3D human generator creates the appearance and geometry in canonical space represented by 3D Gaussians and leverages an efficient deformation module to deform these into posed space for Gaussian rasterization. To learn this generative avatar model from a 3D human dataset, we utilize a single-stage training pipeline that jointly optimizes a Gaussian auto-decoder (including a per-subject latent code and a shared decoder) and a latent diffusion model. *Model Fitting:* At test time, we fit the learned generative avatar to synthetic images generated from a pretrained multi-view diffusion model. During this process, we first initialize the latent code by image-guided sampling and perform model inversion to compute the latent code while freezing the decoder and learned diffusion model. Both the avatar model and camera/human pose parameters are optimized in an alternating manner to correct abnormal poses.

To make our model generalize to various people, we learn a shared auto-decoder across all the training people. For each identity, we model each person by a small compressed latent code $X_i \in \mathbb{R}^{64\times64\times32}$ and then decode the latent code to final UV map U_i using the shared CNN decoder. More details of the decoder can be found in Sup. Mat.

3.1.2. Deformer

To enable animation and learn from posed images, we use a deformer to transform the avatar $\mathcal{G}$ from the canonical space into posed space. For each Gaussian primitive $\mathcal{G}_k$, the deformed Gaussian center and rotation matrix μ'_k and R'_k are computed as:

$$\mu'_k = T\mu_k, R'_k = TR_k, \text{ where } T = \sum_{i=1}^{n_b} w_i B_i. \quad (2)$$

Here n_b is the number of joints, B_i is the bone transformation matrix for joint $i \in \{1, ..., n_b\}$, and w_i is the skinning weight, which determines the influence of the motion of each joint on μ_k. Following AG3D [11], the skinning weight is represented as a low-resolution voxel grid. More details can be found in Sup. Mat.

3.1.3. Rendering

After we obtain the deformed Gaussian attributes, we perform 2D Gaussian splatting as in [19]. For each pixel $\mathbf{x} = (x, y)$, the pixel color is obtained by:

$$c(\mathbf{x}) = \sum_{i=1}^{N} c_i \mathcal{G}_i(\mathbf{x})\sigma_i \prod_{j=1}^{i-1}(1 - \sigma_j \mathcal{G}_j(\mathbf{x})) \quad (3)$$

where c_i is the color of the i-th projected 2D Gaussian primitive sorted by depth. To render normal maps, we replace the color c_i with the normal of the Gaussian primitives. σ_i represents the opacity values. $\mathcal{G}(\mathbf{x})$ is the evaluated 2D Gaussian value. More details of the evaluation of $\mathcal{G}(\mathbf{x})$ can be seen in [19].

3.1.4. Generative Avatar Training

Fig. 2(a) illustrates the training process of this generative avatar model. We leverage the latent diffusion model (LDM) [40] to learn the generative prior in the latent space. Following SSDNeRF [7], we adopt a single-stage training pipeline to jointly optimize our auto-decoder and LDM. The training objective is:

$$\mathcal{L} = \lambda_{\text{rend}}\mathcal{L}_{\text{rend}}(\{X_i\}, \psi) + \lambda_{\text{diff}}\mathcal{L}_{\text{diff}}(\{X_i\}, \phi). \quad (4)$$

Here X_i is the latent feature code and ψ and ϕ denote the parameters of the decoder and denoising U-Net respectively. λ_* are the loss weights. $\mathcal{L}_{\text{rend}}$ and $\mathcal{L}_{\text{diff}}$ are the training objectives for the rendering and diffusion process. Compared to two-stage training [46, 50], the resulting learned latent space is smoother due to end-to-end optimization of the diffusion and decoder weights. The rendering loss is:

$$\mathcal{L}_{\text{rend}}(\{X_i\}, \psi) = \lambda_{\text{l2}}\mathcal{L}_{\text{l2}} + \lambda_{\text{vgg}}\mathcal{L}_{\text{vgg}} + \lambda_{\text{reg}}\mathcal{L}_{\text{reg}}. \quad (5)$$

$\mathcal{L}_{\text{l2}}$ is the L2 reconstruction loss between rendering and observations. Unlike [7], we compute the reconstruction loss on both RGB and normal images to improve the geometry quality. $\mathcal{L}_{\text{vgg}}$ is a perceptual loss based on the difference between the feature maps obtained from [47] and the rendered image. We define $\mathcal{L}_{\text{reg}} = \|\delta_{\mu k}\|$ to prevent the predicted offset from being too large.

To train the diffusion model, similar to [7], we compute $\mathcal{L}_{\text{diff}}$ as:

$$\mathcal{L}_{\text{diff}}\left(\{X_i\}, \phi\right) = \mathop{\mathbb{E}}_{i,t,\epsilon}\left[\frac{1}{2}w^{(t)}\left\|\hat{X}_i - X_i\right\|^2\right] \quad (6)$$

where $\hat{X}_i$ is the denoised latent code with time step $t \sim \mathcal{U}(0, T)$, $w^{(t)}$ is an empirical time-dependent weighting function, and ϵ is the added noise. More details can be found in Sup. Mat.

3.2. Model Fitting

Equipped with the learned generative avatar prior, we reconstruct a personalized avatar by fitting the generative model to synthetic views generated from multi-view diffusion.

3.2.1. Multi-view Hallucination and Pose Estimation

Since a single image is not enough for Gaussian reconstruction, we leverage a pre-trained multi-view diffusion model [26] to hallucinate 6 synthetic human images from a single image. After obtaining synthetic views, similar to [45], we leverage a human pose estimator to obtain the initial SMPL-X parameters. More details of preprocessing can be found in Sup. Mat.

3.2.2. Rendering Objective

To optimize the latent feature map, we define the rendering loss $\mathcal{L}'_{\text{rend}}$ during inference as:

$$\begin{aligned}\mathcal{L}'_{\text{rend}}(\{X_i\}, \psi) =& \lambda_{\text{l2}}\mathcal{L}_{\text{l2}} + \lambda_{\text{vgg}}\mathcal{L}_{\text{vgg}} + \lambda_{\text{reg}}\mathcal{L}_{\text{reg}} \\ &+ \lambda_{\text{nc}}\mathcal{L}_{\text{nc}} + \lambda_{\text{d}}\mathcal{L}_{\text{d}}.\end{aligned} \quad (7)$$

Here we use the same equation as in Eq. (4) to calculate the L2 reconstruction loss $\mathcal{L}_{\text{l2}}$ and perceptual loss $\mathcal{L}_{\text{vgg}}$ between predicted images (normals) and generated synthetic images (normals). To improve the geometry of the avatar, we additionally add the normal consistency loss $\mathcal{L}_{\text{nc}}$ and the depth distortion loss $\mathcal{L}_{\text{d}}$ from [19]. More details can be found in the Sup. Mat.

3.2.3. Prior-guided Optimization

Since the multi-view diffusion model generates sparse and inconsistent images, solely relying on the rendering loss makes the result blurry and unrealistic. Here, we tackle the problem by leveraging our pretrained generative avatar prior as a powerful human prior of 3D appearance and geometry. More concretely, this prior mainly contributes to three aspects including:

Initialization. Random initialization of the latent code can sometimes cause the optimization to converge to a bad local minimum. To solve this problem, we leverage our learned generative model to provide a good starting point for model fitting. To make our model generalizable to unseen test images, we follow [7] to use image-guided sampling. More specifically, for a noisy code $X^{(t)}$ at every denoising step t, we additionally compute an approximated rendering gradient g based on testing rendering loss $\mathcal{L}'_{\text{rend}}(X^{(t)})$ and add it to the denoised output $\hat{X}^{(t)}$ as an image-guided correction. More details of this computation can be found in Sup. Mat.

Regularization. Image-guided sampling provides a good initialization, but still cannot reconstruct all the details of test images. To solve this, we refine the sampled latent code by solving :

$$\min_X \lambda_{\text{rend}}\mathcal{L}'_{\text{rend}}(X) + \lambda'_{\text{diff}}\mathcal{L}_{\text{diff}}(X). \quad (8)$$

Here we optimize the diffusion loss defined in Eq. (6) jointly with the rendering loss, while freezing the weights of the diffusion and decoder models. Unlike [37], the diffusion model trained in UV feature space serves as a prior to regularize and inpaint the noisy and incomplete latent code during optimization.

Pose Optimization. The generated synthetic images from multi-view diffusion models are inconsistent, and this results in inaccurate camera and body pose estimation. Instead of using noisy 2D joint estimates [4, 34], we optimize pose parameters via a more effective photometric loss:

$$\mathcal{L}_{\text{pose}} = \lambda_{\text{l2}}\mathcal{L}_{\text{l2}} + \lambda_{\text{vgg}}\mathcal{L}_{\text{vgg}} + \lambda_{\text{mask}}\mathcal{L}_{\text{mask}} \quad (9)$$

where we combine a mask loss $\mathcal{L}_{\text{mask}}$ with predefined L2 loss $\mathcal{L}_{\text{l2}}$ and a perceptual loss. Here we optimize both SMPL

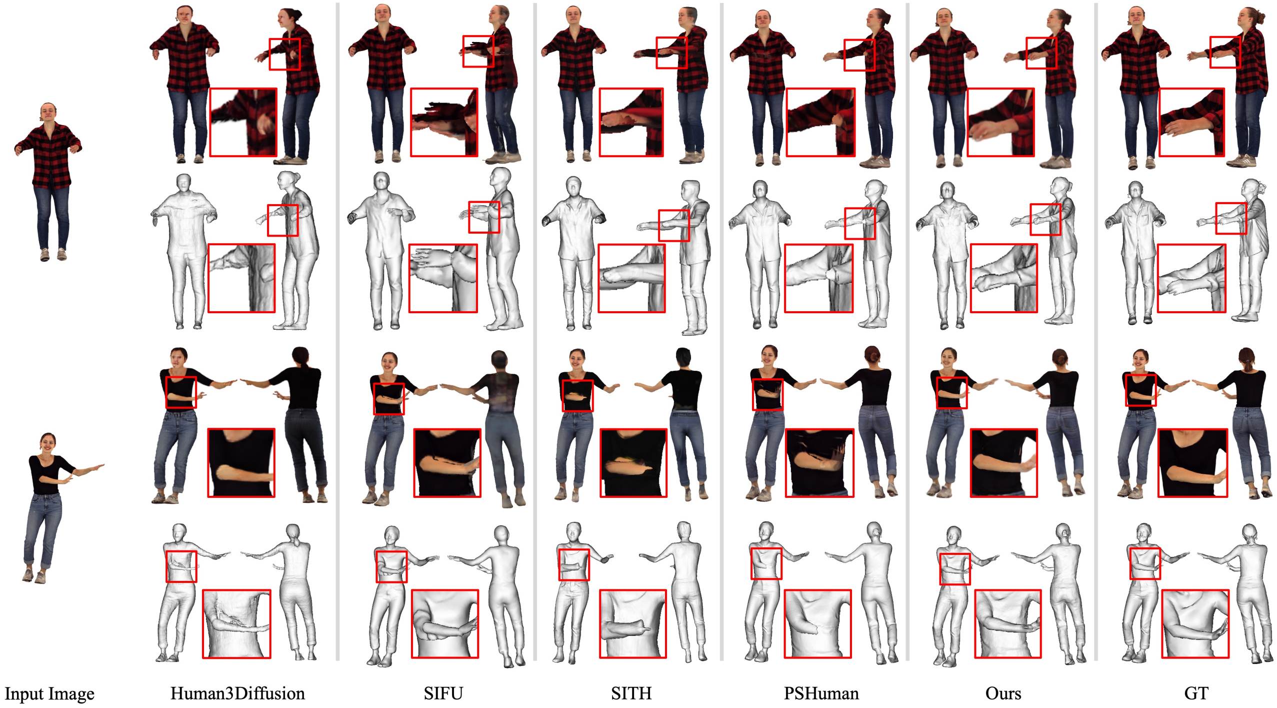

Figure 3. **Qualitative comparison to SotA methods on CustomHuman.** Our method achieves better image and shape quality, enables 3D consistency in side views, and avoids unrealistic reconstruction due to self-occlusion.

parameters and camera poses by back-propagation. We optimize the latent code, camera and human poses in an alternating manner to avoid falling into locally suboptimal results. More details can be found in Sup. Mat.

4. Experiments

In our experiments, we first compare our method to state-of-the-art (SotA) baselines on two public datasets and then test the generalization capability of our method on in-the-wild images. In addition, we provide an ablation study to investigate the importance of each component in our model.

Datasets. We evaluate our proposed method on THuman2 [55] and CustomHumans [13]. To test the generalization ability, we also collect some in-the-wild images from the Internet for qualitative comparison. For details see Sup. Mat.

Metrics. Following previous methods [14, 26], we evaluate appearance using peak signal-to-noise ratio (PSNR), structural similarity index (SSIM), and perceptual similarity (LPIPS). For geometry comparison, we compute Chamfer Distance (CD), Point to Surface (P2S) distance, and normal consistency (NC). For details see Sup. Mat.

Baselines. We conduct experiments on current state-of-the-art methods for single-view human reconstruction, including Human3Diffusion [54], SIFU [59], SiTH [14], and a concurrent method, PSHuman [26]. More details about baselines can be found in Sup. Mat.

4.1. Comparison to SotA.

Table 1 summarizes our quantitative comparisons. Since Human3Diffusion [54] only supports low-resolution rendering, we mainly compare it qualitatively in Fig. 3. Table 1 shows that our method largely outperforms other baselines in appearance, especially in PSNR. As shown in Fig. 3, our method generates overall sharper images with more details. Our method also reconstructs better geometry both quantitatively and qualitatively with better local geometric detail. Here, we discuss the main reason for the improvements:

Side views. The improvements of appearance and geometry are particularly pronounced for side views. This is because most baselines rely on 2D multi-view diffusion models to hallucinate side views, which are not 3D consistent. In contrast, our learned generative avatar model provides additional appearance and geometry constraints on multi-view consistency, yielding better results.

Self-occlusion. With a single input view, self-occlusion inevitably happens in the arm and hand regions due to the articulated nature of the human body as illustrated in Fig. 3. In the first example, despite the use of the SMPL body prior, all previous methods fail to accurately reconstruct the left hand and arm. This is because these baselines tend to overfit to the input view, leading to incomplete reconstructions in occluded regions. In contrast, our method effectively reconstructs occluded arms and hands by leveraging the 3D

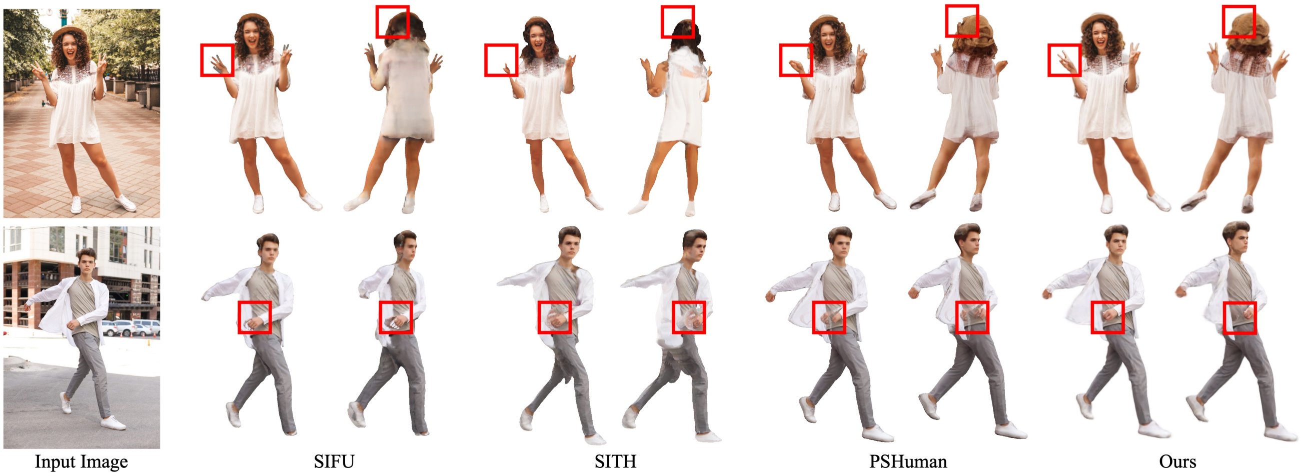

Figure 4. **Qualitative comparison to SotA methods on in-the-wild images:** Ours outperforms baselines on in-the-wild images by generating more plausible back/side views, reconstructing finer details such as fingers and hats, and avoids artifacts due to self-occlusion.

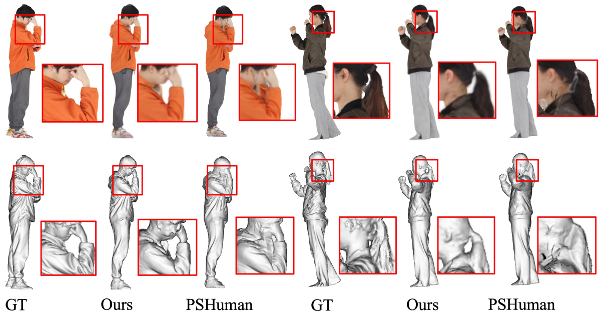

Figure 5. **Qualitative comparison to PSHuman on fine-sclae structures.** Our method reconstructs complex topologies, like a ponytail, that deviate from the body topology.

appearance prior from our model. In the second example, the baselines exhibit artifacts on the clothing due to color misalignment from the arm, which is exacerbated by self-occlusion and depth ambiguity. Instead, our method reconstructs both the arm and clothing, despite the occlusion.

Topology changes. Figure 5 compares our method with the concurrent PSHuman [26] on two more challenging subjects. Due to fixed topology of the template mesh, PSHuman fails to model areas between the arm and face, as well as the ponytail. These artifacts are present in other SMPL-based methods [59]. In contrast, our method reconstructs 3D Gaussians and this flexible representation allows it to reconstruct more complex structures.

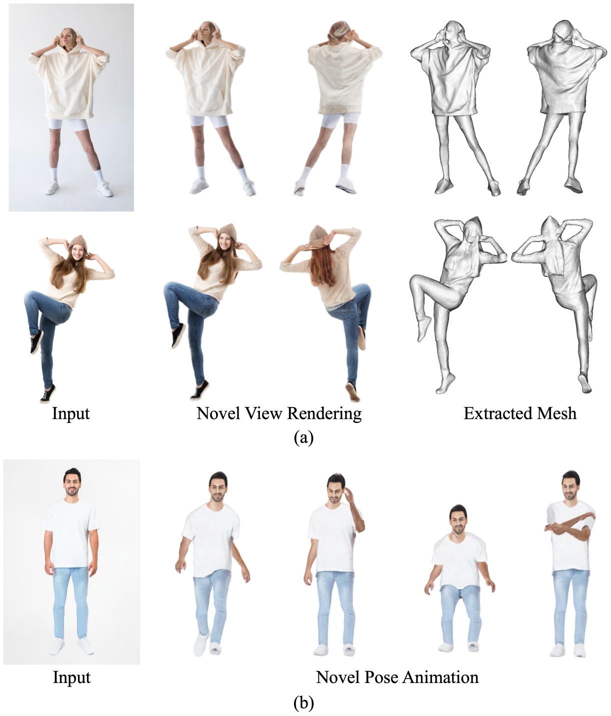

Figure 6. **Qualitative results on in-the-wild images**. (a) From an in-the-wild image with challenging poses and clothing, our method reconstructs a high-quality Gaussian avatar, that enables realistic novel view synthesis and detailed geometry reconstruction. (b) The resulting avatar can be animated with SMPL-X poses.

4.2. In-the-wild Performance

Figure 6(a) shows qualitative results on in-the-wild images. MoGA generalizes to loose clothing and challenging poses. The reconstructed Gaussian avatar can be posed or animated (Fig. 6(b)), because it is based on SMPL-X. Figure 4 compares results of SotA methods on in-the-wild images. Both SiTH and SIFU struggle to produce reasonable reconstructions, often generating blurry back views and unrealistic appearance. PSHuman works better, but fails to capture fine details such as fingers and hats, while also introducing artifacts on clothing. In contrast, by leveraging the learned

Method	THuman2.1						CustomHuman					
	PSNR↑	SSIM↑	LPIPS↓	CD↓	P2S↓	NC↑	PSNR↑	SSIM↑	LPIPS↓	CD↓	P2S↓	NC↑
SIFU[59]	17.5271	0.9219	0.1019	2.6220	2.4450	0.787	16.0198	0.9056	0.1146	2.3717	2.2206	0.809
SiTH [14]	19.4020	0.9344	0.0796	2.2370	1.8459	0.808	17.7986	0.9211	0.0921	2.9142	2.0426	0.788
PSHuman [26]	19.9595	0.9350	0.0778	1.4128	1.2320	0.837	18.6704	0.9223	0.0850	1.9197	**1.4695**	0.828
Ours	**24.0926**	**0.9455**	**0.0732**	**1.3608**	**1.2226**	**0.850**	**23.4383**	**0.9351**	**0.0791**	**1.8086**	1.4821	**0.834**

Table 1. **Quantitative comparison with SotA Methods on Thuman2.1 and CustomHuman.** Our method outperforms other baselines by a large margin on appearance and also demonstrates a clear improvement in geometry.

generative avatar prior, our MoGA model faithfully reconstructs Gaussian avatars in this challenging setting.

4.3. Ablation Study

Since the generative avatar prior is the key to our method, we focus on ablations that evaluate its effectiveness. An analysis of pose optimization, unconditional generation, and robustness to the number of views appears in Sup. Mat. All experiments are conducted on the CustomHuman Dataset.

Method	PSNR ↑	SSIM ↑	LPIPS ↓
w/o Initialization	22.6996	0.9278	0.0891
w/o Avatar Prior	22.5838	0.9288	0.0863
Full Model	**23.4383**	**0.9351**	**0.0791**

Table 2. **Ablation.** We compare our method with ablated baselines in which we remove the generative avatar prior for initialization and regularization.

Effect of initialization. Our generative avatar plays a crucial role in initializing the fitting process. We compare our method against ablated versions where the latent code is initialized randomly. As shown in Table 2, MoGA improves all appearance metrics. Furthermore, qualitative results in Fig. 7 demonstrate that, without proper initialization, the reconstructed appearance becomes blurry, particularly in the face region. This degradation occurs because the optimization process converges to a poor local minimum. In contrast, our method produces sharper and more accurate facial reconstructions, highlighting the importance of a well-initialized latent code.

Effect of generative avatar prior. To evaluate the effect of the generative avatar prior, we create an ablated version of our method that directly optimizes SMPL-anchored Gaussians without using the learned decoder and latent diffusion model. Table 2 shows that the generative avatar prior is important. In the first example of Fig. 8, removing the generative avatar prior results in a blurry side view and introduces a visible white crack between the front and back. This artifact arises due to the 3D inconsistency of images generated by the multi-view diffusion model. In contrast, our method enforces better 3D consistency. The second example in Fig. 8 highlights another issue without our model. The ablated baseline tends to produce artifacts in occluded regions. In comparison, incorporating the generative prior enables our model to inpaint the missing areas effectively, resulting in a more natural and complete appearance.

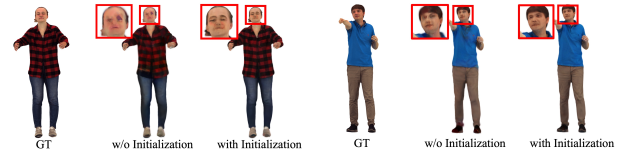

Figure 7. **Ablation of Initialization.** The good initialization provided by our generative avatar model enhances appearance quality, reducing blurriness and producing a more detailed face

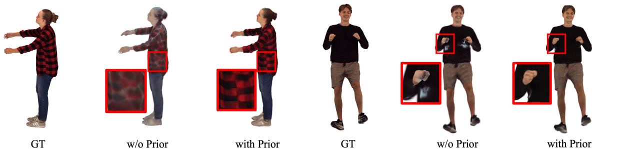

Figure 8. **Ablation of generative avatar prior.** The generative avatar model serves as an important 3D regularization to ensure 3D consistency and inpaint missing regions.

5. Conclusion

In this paper, we propose MoGA, a novel approach for reconstructing Gaussian avatars from a monocular image. Unlike previous methods that rely solely on multi-view diffusion, we integrate a 3D generative avatar model as a complementary prior, ensuring 3D consistency by projecting images into its latent space and enforcing both 3D appearance and geometry constraints. We formulate Gaussian avatar creation as model inversion by fitting the generative avatar model to synthetic images from 2D diffusion models. Our method sets a new state-of-the-art in reconstruction quality and 3D consistency, generalizing well to in-the-wild images while producing animatable avatars without post-processing. Limitations and discussions appear in Sup. Mat.

Acknowledgements: Andreas Geiger was supported by the ERC Starting Grant LEGO-3D (850533) and the DFG EXC number 2064/1 - project number 390727645. We thank Chengwei Zheng, Chen Guo, Juan Zarate, Zhiyi Chen and Jiarui Gao for their help and feedback.

Disclosure. While MJB is a co-founder and Chief Scientist at Meshcapade, his research in this project was performed solely at, and funded solely by, the Max Planck Society.

References

[1] Rameen Abdal, Wang Yifan, Zifan Shi, Yinghao Xu, Ryan Po, Zhengfei Kuang, Qifeng Chen, Dit-Yan Yeung, and Gordon Wetzstein. Gaussian shell maps for efficient 3d human generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9441–9451, 2024. 3

[2] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Video based reconstruction of 3d people models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8387–8397, 2018. 3

[3] Thiemo Alldieck, Mihai Zanfir, and Cristian Sminchisescu. Photorealistic monocular 3d reconstruction of humans wearing clothing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1506–1515, 2022. 2

[4] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J. Black. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *Computer Vision – ECCV 2016*. Springer International Publishing, 2016. 5

[5] Eric R Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5799–5809, 2021. 3

[6] 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, et al. Efficient geometry-aware 3d generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16123–16133, 2022. 3

[7] Hansheng Chen, Jiatao Gu, Anpei Chen, Wei Tian, Zhuowen Tu, Lingjie Liu, and Hao Su. Single-stage diffusion nerf: A unified approach to 3d generation and reconstruction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2416–2425, 2023. 2, 3, 5

[8] Jinnan Chen, Chen Li, Jianfeng Zhang, Lingting Zhu, Buzhen Huang, Hanlin Chen, and Gim Hee Lee. Generalizable human gaussians from single-view image. *arXiv preprint arXiv:2406.06050*, 2024. 2, 3

[9] Zhaoxi Chen, Fangzhou Hong, Haiyi Mei, Guangcong Wang, Lei Yang, and Ziwei Liu. Primdiffusion: Volumetric primitives diffusion for 3d human generation. *Advances in Neural Information Processing Systems*, 36:13664–13677, 2023. 3

[10] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20470–20480, 2022. 3

[11] Zijian Dong, Xu Chen, Jinlong Yang, Michael J Black, Otmar Hilliges, and Andreas Geiger. Ag3d: Learning to generate 3d avatars from 2d image collections. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 14916–14927, 2023. 2, 3, 4

[12] Chen Guo, Tianjian Jiang, Xu Chen, Jie Song, and Otmar Hilliges. Vid2avatar: 3d avatar reconstruction from videos in the wild via self-supervised scene decomposition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12858–12868, 2023. 3

[13] Hsuan-I Ho, Lixin Xue, Jie Song, and Otmar Hilliges. Learning locally editable virtual humans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21024–21035, 2023. 6

[14] I Ho, Jie Song, Otmar Hilliges, et al. Sith: Single-view textured human reconstruction with image-conditioned diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 538–549, 2024. 1, 2, 6, 8

[15] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 3

[16] Fangzhou Hong, Zhaoxi Chen, Yushi Lan, Liang Pan, and Ziwei Liu. Eva3d: Compositional 3d human generation from 2d image collections. *arXiv preprint arXiv:2210.04888*, 2022. 3

[17] Liangxiao Hu, Hongwen Zhang, Yuxiang Zhang, Boyao Zhou, Boning Liu, Shengping Zhang, and Liqiang Nie. Gaussianavatar: Towards realistic human avatar modeling from a single video via animatable 3d gaussians. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 634–644, 2024. 3

[18] Tao Hu, Fangzhou Hong, and Ziwei Liu. Structldm: Structured latent diffusion for 3d human generation. *arXiv preprint arXiv:2404.01241*, 2024. 3

[19] Binbin Huang, Zehao Yu, Anpei Chen, Andreas Geiger, and Shenghua Gao. 2d gaussian splatting for geometrically accurate radiance fields. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024. 2, 3, 4, 5

[20] Yangyi Huang, Hongwei Yi, Yuliang Xiu, Tingting Liao, Jiaxiang Tang, Deng Cai, and Justus Thies. Tech: Text-guided reconstruction of lifelike clothed humans. In *2024 International Conference on 3D Vision (3DV)*, pages 1531–1542. IEEE, 2024. 2

[21] Tianjian Jiang, Xu Chen, Jie Song, and Otmar Hilliges. Instantavatar: Learning avatars from monocular video in 60 seconds. 2023. 3

[22] Hanbyul Joo, Hao Liu, Lei Tan, Lin Gui, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social motion capture. In *Proceedings of the IEEE international conference on computer vision*, pages 3334–3342, 2015. 1

[23] Yash Kant, Ethan Weber, Jin Kyu Kim, Rawal Khirodkar, Su Zhaoen, Julieta Martinez, Igor Gilitschenski, Shunsuke Saito, and Timur Bagautdinov. Pippo: High-resolution multi-view humans from a single image. *arXiv preprint arXiv:2502.07785*, 2025. 1

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

[25] Tobias Kirschstein, Simon Giebenhain, Jiapeng Tang, Markos Georgopoulos, and Matthias Nießner. Gghead: Fast and generalizable 3d gaussian heads. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–11, 2024. 3

[26] Peng Li, Wangguandong Zheng, Yuan Liu, Tao Yu, Yangguang Li, Xingqun Qi, Mengfei Li, Xiaowei Chi, Siyu Xia, Wei Xue, et al. Pshuman: Photorealistic single-view human reconstruction using cross-scale diffusion. *arXiv preprint arXiv:2409.10141*, 2024. 1, 2, 5, 6, 7, 8

[27] 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. *Advances in Neural Information Processing Systems*, 36:22226–22246, 2023. 2

[28] 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 *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9298–9309, 2023. 2

[29] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 851–866. 2023. 2, 3

[30] Feichi Lu, Zijian Dong, Jie Song, and Otmar Hilliges. Avatarpose: Avatar-guided 3d pose estimation of close human interaction from sparse multi-view videos. In *European Conference on Computer Vision*, pages 215–233. Springer, 2024. 3

[31] Gyeongsik Moon, Takaaki Shiratori, and Shunsuke Saito. Expressive whole-body 3d gaussian avatar. In *European Conference on Computer Vision*, pages 19–35. Springer, 2024. 3

[32] Atsuhiro Noguchi, Xiao Sun, Stephen Lin, and Tatsuya Harada. Unsupervised learning of efficient geometry-aware neural articulated representations. In *European Conference on Computer Vision*, pages 597–614. Springer, 2022. 3

[33] Panwang Pan, Zhuo Su, Chenguo Lin, Zhen Fan, Yongjie Zhang, Zeming Li, Tingting Shen, Yadong Mu, and Yebin Liu. Humansplat: Generalizable single-image human gaussian splatting with structure priors. *Advances in Neural Information Processing Systems*, 37:74383–74410, 2025. 3

[34] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10975–10985, 2019. 5

[35] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3D hands, face, and body from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10975–10985, 2019. 2, 3

[36] Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9054–9063, 2021. 3

[37] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 2, 5

[38] Guocheng Qian, Jinjie Mai, Abdullah Hamdi, Jian Ren, Aliaksandr Siarohin, Bing Li, Hsin-Ying Lee, Ivan Skorokhodov, Peter Wonka, Sergey Tulyakov, et al. Magic123: One image to high-quality 3d object generation using both 2d and 3d diffusion priors. *arXiv preprint arXiv:2306.17843*, 2023. 2

[39] Zhiyin Qian, Shaofei Wang, Marko Mihajlovic, Andreas Geiger, and Siyu Tang. 3dgs-avatar: Animatable avatars via deformable 3d gaussian splatting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5020–5030, 2024. 3

[40] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 5

[41] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2304–2314, 2019. 1, 2

[42] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 84–93, 2020. 1, 2

[43] Shunsuke Saito, Gabriel Schwartz, Tomas Simon, Junxuan Li, and Giljoo Nam. Relightable gaussian codec avatars. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 130–141, 2024. 3

[44] Katja Schwarz, Yiyi Liao, Michael Niemeyer, and Andreas Geiger. Graf: Generative radiance fields for 3d-aware image synthesis. *Advances in Neural Information Processing Systems*, 33:20154–20166, 2020. 3

[45] Kaiyue Shen, Chen Guo, Manuel Kaufmann, Juan Jose Zarate, Julien Valentin, Jie Song, and Otmar Hilliges. X-avatar: Expressive human avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16911–16921, 2023. 5

[46] J Ryan Shue, Eric Ryan Chan, Ryan Po, Zachary Ankner, Jiajun Wu, and Gordon Wetzstein. 3d neural field generation using triplane diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20875–20886, 2023. 3, 5

[47] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 5

[48] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 3

[49] Yingzhi Tang, Qijian Zhang, Junhui Hou, and Yebin Liu. Human as points: Explicit point-based 3d human reconstruction from single-view rgb images. *arXiv preprint arXiv:2311.02892*, 2023. 3

[50] Tengfei Wang, Bo Zhang, Ting Zhang, Shuyang Gu, Jianmin Bao, Tadas Baltrusaitis, Jingjing Shen, Dong Chen, Fang Wen, Qifeng Chen, et al. Rodin: A generative model for sculpting 3d digital avatars using diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4563–4573, 2023. 3, 5

[51] 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*, pages 16210–16220, 2022. 3

[52] Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J Black. Icon: Implicit clothed humans obtained from normals. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13286–13296. IEEE, 2022. 1, 2

[53] Yuxuan Xue, Xianghui Xie, Riccardo Marin, and Gerard Pons-Moll. Human 3diffusion: Realistic avatar creation via explicit 3d consistent diffusion models. *arXiv preprint arXiv:2406.08475*, 2024. 1, 2

[54] Yuxuan Xue, Xianghui Xie, Riccardo Marin, and Gerard. Pons-Moll. Human 3Diffusion: Realistic Avatar Creation via Explicit 3D Consistent Diffusion Models. 2024. 6

[55] 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)*, 2021. 6

[56] Ye Yuan, Xueting Li, Yangyi Huang, Shalini De Mello, Koki Nagano, Jan Kautz, and Umar Iqbal. Gavatar: Animatable 3d gaussian avatars with implicit mesh learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 896–905, 2024. 3

[57] Bowen Zhang, Yiji Cheng, Chunyu Wang, Ting Zhang, Jiaolong Yang, Yansong Tang, Feng Zhao, Dong Chen, and Baining Guo. Rodinhd: High-fidelity 3d avatar generation with diffusion models. In *European Conference on Computer Vision*, pages 465–483. Springer, 2024. 3

[58] Weitian Zhang, Yichao Yan, Yunhui Liu, Xingdong Sheng, and Xiaokang Yang. e^3gen: Efficient, expressive and editable avatars generation. *arXiv preprint arXiv:2405.19203*, 2024. 3

[59] Zechuan Zhang, Zongxin Yang, and Yi Yang. Sifu: Side-view conditioned implicit function for real-world usable clothed human reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9936–9947, 2024. 1, 2, 6, 7, 8

[60] Haoyu Zhao, Hao Wang, Chen Yang, and Wei Shen. Chase: 3d-consistent human avatars with sparse inputs via gaussian splatting and contrastive learning. *arXiv preprint arXiv:2408.09663*, 2024. 3

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

[62] Yufeng Zheng, Wang Yifan, Gordon Wetzstein, Michael J Black, and Otmar Hilliges. Pointavatar: Deformable point-based head avatars from videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 21057–21067, 2023. 3