

AniGS: Animatable Gaussian Avatar from a Single Image with Inconsistent Gaussian Reconstruction

Lingteng Qiu^{1*} Shenhao Zhu^{1,3*} Qi Zuo^{1*} Xiaodong Gu^{1*}
 Yuan Dong¹ Junfei Zhang¹ Chao Xu¹ Zhe Li^{1,4} Weihao Yuan¹ Liefeng Bo¹
 Guanying Chen^{2†} Zilong Dong^{1†}

¹Alibaba Group ²Sun Yat-sen University
³Nanjing University ⁴Huazhong University of Science and Technology



Figure 1. **3D Avatar Reconstruction and Animation Results of AniGS.** Given a single human image as input, AniGS is capable of reconstructing a high-fidelity 3D avatar in a canonical pose, which can be used for both photorealistic rendering and real-time animation.

Abstract

Generating animatable human avatars from a single image is essential for various digital human modeling applications. Existing 3D reconstruction methods often struggle to capture fine details in animatable models, while generative approaches for controllable animation, though avoiding ex-

plicit 3D modeling, suffer from viewpoint inconsistencies in extreme poses and computational inefficiencies. In this paper, we address these challenges by leveraging the power of generative models to produce detailed multi-view canonical pose images, which help resolve ambiguities in animatable human reconstruction. We then propose a robust method for 3D reconstruction of inconsistent images, enabling real-time rendering during inference. Specifically, we adapt a transformer-based video generation model to generate

*Equal contribution.

†Corresponding author.

multi-view canonical pose images and normal maps, pre-training on a large-scale video dataset to improve generalization. To handle view inconsistencies, we recast the reconstruction problem as a 4D task and introduce an efficient 3D modeling approach using 4D Gaussian Splatting. Experiments demonstrate that our method achieves photorealistic, real-time animation of 3D human avatars from in-the-wild images, showcasing its effectiveness and generalization capability. Our code will be available on <https://lingtengqiu.github.io/2024/AniGS/>.

1. Introduction

Generating animatable human avatars has become increasingly important for a wide range of applications, such as virtual reality, gaming, and human-robot interaction. However, creating an animatable human avatar with diverse shapes, appearances, and clothing from a single image remains a challenging problem.

Despite significant progress in human reconstruction from a single image [50, 62, 76, 82], the reconstructed models are often difficult to animate. This is because the reconstructed human pose is aligned with the input pose, which is usually non-canonical, requiring complex rigging to enable animation. For animatable avatar reconstruction, methods based on predicting parametric human models (*e.g.*, SMPL [46]) with geometry offset refinement often struggle to capture fine details [3]. Additionally, methods relying on implicit surface reconstruction face challenges in generalization, primarily due to the lack of large-scale, rigged 3D human datasets for training [20, 29].

Recently, controllable human image animation has made significant progress with diffusion models [27, 85], which generate animated images directly. These methods achieve realistic results and are easy to animate by input control poses, but suffer from inconsistencies across views due to the lack of a global representation. In addition, these methods also face efficiency issues due to the high computational effort per animation frame.

Motivated by the success of diffusion models in generating multi-view images of objects [45, 67], several methods have explored fine-tuning these models to generate multi-view human images from a single input, followed by 3D reconstruction using multi-view techniques [21, 44]. Specifically, CharacterGen [53] demonstrates the ability to generate *cartoon-style avatars* from a single image by first producing multi-view canonical pose images. To reconstruct the 3D model from these potentially inconsistent multi-view images, a transformer-based 3D reconstruction model is trained. However, training both the multi-view diffusion model and the reconstruction model requires a synthetic 3D *rigged* human dataset to render multi-view canonical images, limiting the generalization ability of these models.

To address these challenges, we adapt a transformer-based video generation model [26] to predict multi-view images and normal maps, by incorporating guidance support from the reference image and human poses. Our generation model can be trained on large-scale, in-the-wild video data, thereby enhancing its generalization.

Despite the high-quality multi-view images generated by our method, inconsistencies still arise, which can affect the 3D reconstruction. By considering these inconsistencies as the dynamic variations within a temporal sequence, we can reformulate the problem of 3D reconstruction from inconsistent images as a 4D reconstruction task. Observing the effectiveness of 4D Gaussian splatting (4DGS) in dynamic scene modeling [47, 73, 79], we introduce an efficient 4DGS to fit the multi-view images. After optimization, the high-fidelity 3D avatar model can be obtained as the model in the canonical space of the 4DGS (see Fig. 1).

In summary, the key contributions of this paper are as follows:

- We introduce a method for multi-view canonical pose image generation using a video generation model, trained on unconstrained human pose video data, without the need for synthetic 3D rigged human datasets.
- We introduce a new perspective to the problem of 3D animatable avatar reconstruction from inconsistent images, formulating it as a 4D reconstruction task and introducing an efficient 4D Gaussian Splatting model.
- Experiments demonstrate that our method generates high-fidelity animatable avatars from a single image, enabling photorealistic and real-time animation during inference.

2. Related Work

Single-Image Human Reconstruction and Generation

Early methods for single-image human reconstruction primarily formulated this problem as geometry offset prediction for mesh-based statistical models for naked [15, 32, 36, 37, 63, 68, 70] or clothed [2, 4, 5, 8, 30, 40, 55, 74, 84] human body, with some approaches extending this to texture prediction [5, 8]. While a consistent naked topology simplifies animation, it is less effective for modeling diverse clothing styles. To deal with the cloth-style variation in the wild, numerous notable approaches [6, 10, 16, 61, 62, 76, 77, 82, 83] utilize implicit functions as representations for 3D human models, enabling them to capture complex topologies without suffering from resolution constraints.

Recently, the rise of generative models has blurred the boundaries between reconstruction and generation processes. Some methods [21, 44, 48] utilize the input image as a conditional element, leveraging generative techniques such as GANs [17, 25, 48, 78], Image Diffusion Models [12], and Video Diffusion Models [44, 66] to synthesize 3D human models. The concurrent work MagicMan [21]

has explored the idea of generating multi-view human images and normal maps given the input image. Unlike these methods, which generate static models, our approach generates multi-view canonical pose images, followed by 3D modeling to reconstruct animatable human models.

Animatable Human Generation Reconstructing animatable avatars remains a challenging problem. Early methods often adopt a parametric-model-based solution [3]. To better model the clothed human body, methods like ARCH [20, 29] adopts an implicit function representation to represent the geometry of the human body. Other works focus on animatable avatar reconstruction from monocular videos [31, 43, 58, 72] or multi-view videos [13, 41, 54]. With the advent of text-to-3D techniques [56], several methods have explored generating 3D avatars from text prompts [11, 28, 38].

Recently, CharacterGen [53] achieves cartoon-style avatar generation from a single image by generating multi-view images with a diffusion model, followed by transformer-based shape reconstruction. In contrast, our work focuses on in-the-wild animatable human avatars, using a video generation model pretrained on large-scale in-the-wild videos to improve generalization. Additionally, we propose a robust 3D reconstruction method based on 4DGs, avoiding the need for training a transformer reconstruction model.

Controllable Human Image Animation Original 2D diffusion models [60] primarily focus on generating single-view images and do not support human animation. Animate Anyone [27] introduces a reference net into the diffusion models to preserve the identity of the input image and incorporates a lightweight, pose-guided network for guidance. Champ [85], MIMO [49], and Human4DiT [66] employ 3D-level shape guidance [46, 51] rather than sparse 2D keypoints, enabling more accurate, controllable image generation under human attributes conditions.

Although they can produce vivid human-centric animation videos, body distortion, and identity mutation often occur when the animated human turns back at wide angles. Additionally, these methods require several minutes to generate a video sequence given the human poses [66, 85], limiting their practicality in interactive applications. In contrast, by leveraging an explicit high-fidelity 3D Gaussian model, our methods achieves real-time photorealistic rendering at inference time.

3. Preliminary

Human Parametric Model The SMPL [46] and SMPL-X [51] parametric models are widely used for human body representation. These models utilize skinning techniques and blend shapes derived from a dataset of thousands of 3D body scans. Specifically, SMPL-X employs shape parame-

ters $\beta \in \mathbb{R}^{20}$ and pose parameters $\theta \in \mathbb{R}^{55 \times 3}$ to represent body mesh deformations.

Diffusion Transformer Model for Video Generation

Diffusion Probabilistic Models [22, 69] use a forward Markov chain to gradually transform a sample x_0 drawn from the data distribution $p(x)$ into a noisy equivalent $q(x)$:

$$q(x_t) = \sqrt{\alpha_t} \cdot x_0 + \sqrt{1 - \alpha_t} \cdot \epsilon, t \in (0, T), \quad (1)$$

where $\epsilon \sim \mathcal{N}(0, I)$ denotes Gaussian noise, T is the final time step, t is the current time step, and α_t is the noisy schedule parameter.

The CogVideo model [26, 80] is an open-source Text-to-Video (T2V) diffusion model that employs a Transformer architecture [52], referred to as μ_θ , to model the reverse diffusion process. The transformer model processes the current noisy sample x_t , the associated time step t , and optional conditioning inputs c to estimate the noise ϵ . The loss function for training the denoising model is:

$$\mathcal{L}_\theta = \mathbb{E}_{x_0, c, t} \|x_0 - \mu_\theta(x_t, c, t)\|^2. \quad (2)$$

3D Gaussian Splatting 3D Gaussian Splatting [33] represents 3D data using a collection of 3D Gaussians. Each Gaussian is defined by a center $\mathbf{x} \in \mathbb{R}^3$, a scaling factor $\mathbf{s} \in \mathbb{R}^3$, and a rotation quaternion $\mathbf{q} \in \mathbb{R}^4$. Additionally, it includes an opacity value $\alpha \in \mathbb{R}$ and a color feature $\mathbf{c} \in \mathbb{R}^C$ for rendering purposes, with spherical harmonics capturing view-dependent effects. Rendering these Gaussians involves projecting them onto the image plane as 2D Gaussians and applying alpha compositing for each pixel in a front-to-back order. Recent methods [47, 73, 79] extend 3D Gaussian Splatting to capture dynamic 4D scenes by incorporating a temporal embedding.

4. Method

4.1. Overview

Given an input human image $I \in \mathbb{R}^{H \times W \times 3}$, our goal is to create an animatable 3D avatar represented by a 3DGs model. This avatar can be animated by applying new human pose conditions during inference. To facilitate animation, it is crucial to reconstruct the 3D avatar in a canonical pose, simplifying the rigging process. The 3D avatar in this canonical pose can be rigged using the aligned SMPL-X model or other off-the-shelf rigging methods [1]. As shown in Fig. 2, our framework consists of two main stages.

In the first stage, we employ a reference image-guided video generation model to produce high-quality multi-view canonical human images and their corresponding normals from the input image. In the second stage, we reconstruct the 3D model using these generated images. However, despite the high quality of the generated images, multi-view inconsistencies still arise due to the nature of the

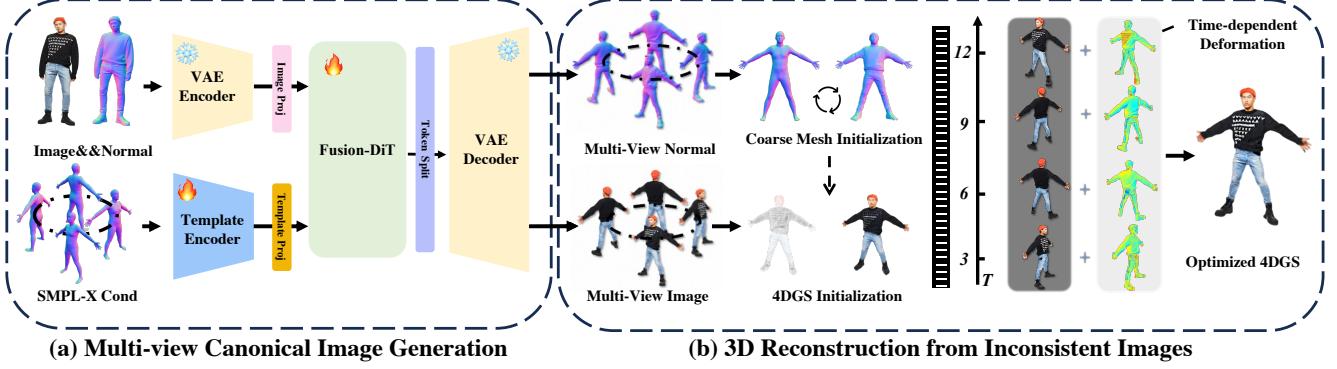


Figure 2. **Overview of the proposed AniGS.** In the first stage, a reference image-guided video generation model is employed to produce high-quality multi-view canonical human images along with their corresponding normals, based on the input image. In the second stage, a robust 3D model reconstruction method is applied, using 4D Gaussian Splatting (4DGS) optimization to handle subtle appearance variations across the generated views.

diffusion-based video generation model. Consequently, directly applying traditional multi-view reconstruction methods to these images often results in the loss of detail and the introduction of artifacts.

To tackle the issue of view inconsistencies, we formulate this problem as a 4D reconstruction task and introduce an efficient 4D Gaussian Splatting model to account for the appearance variations at different timesteps (*i.e.*, viewpoints). After optimization, a high-fidelity 3D model can be obtained as the model in the canonical space of the 4DGS.

4.2. Multi-view Canonical Image Generation

Given a reference human image in an *arbitrary pose*, our goal in the first stage is to generate multi-view RGB images of the same subject in a *canonical pose*. Motivated by recent successes in controllable image generation through video models [27, 80], we adapt a diffusion transformer-based video generation model to reposition the human subject to a canonical pose and generate multi-view images.

Specifically, the video generation model takes as input the reference image and SMPL-X pose conditions to produce multi-view images. Here, the rotation of the camera in relation to the subject is treated as equivalent to subject rotation.

Reference-guided Canonical Video Generation We extend the CogVideo model [26] to achieve reference image-guided and SMPL-X pose-guided video generation. This model includes a transformer-based denoiser and a Variational Autoencoder (VAE) that maps input videos or images into a high-dimensional latent space.

The reference image I is first encoded into VAE features $\mathbb{F}_i \in \mathbb{R}^{B \times 1 \times C \times HW}$, while the latent representations for video noise are denoted by $\mathbb{F}_v \in \mathbb{R}^{B \times f \times C \times HW}$, where B represents the batch size, f the number of output frames, C the number of latent feature channels, and HW the total

number of input feature tokens.

To ensure that generated multi-view images retain the identity of the reference image, we fuse the reference image features and latent features during the denoising process. We achieve this by concatenating the reference image features and latent features along the frame channel, yielding $\mathbb{F}_c \in \mathbb{R}^{B \times (f+1) \times C \times HW}$ at each DiT block, which allows for feature interaction through self-attention.

To guide the video generation with input human poses, we integrate a lightweight pose guidance network inspired by CHAMP [85]. This network extracts guidance features from the canonical SMPL-X normal, N_{smplx} , which are added to the corresponding noisy latents to direct the denoising process.

Joint Multi-view RGB and Normal Generation To enhance multi-view reconstruction with normal supervision [14, 45], we further extend the video generation framework to simultaneously produce multi-view RGB images and normal maps, conditioned on a reference image and its corresponding normal map, as predicted by an existing method [14].

We employ the CogVideo-2B architecture [80] as our base model, which comprises 30 DiT blocks. We modify the first three DiT blocks into two branches, one for RGB and the other for normal inputs. Similarly, we modify the last three DiT blocks into two branches that simultaneously output multi-view RGB images and normal maps.

To effectively integrate image and normal features, we first share the weights of the middle 24 DiT blocks for both RGB and normal feature processing. Secondly, we insert a multi-modal attention module between every three shared DiT blocks and the head of the middle DiT blocks. For this multi-modal attention block, we concatenate the RGB and normal features at the token level, resulting in $\mathbb{F}_m \in \mathbb{R}^{B \times f \times C \times 2HW}$.

Training Strategy To enhance the generalization capabilities of our model in the face of limited large-scale synthetic datasets, we first pre-train the video generation model on a large-scale, in-the-wild dataset consisting of 100,000 single-human animation videos. As ground-truth normal maps are unavailable for in-the-wild data, we use Sapiens [34] and Multi-HMR [7] to generate pseudo labels for both normal maps and SMPL-X human poses. During training, we randomly sample frames from these videos to predict short video segments.

Following pre-training, we generate synthetic 3D data assets to obtain self-rotated RGB images along with corresponding ground-truth normal maps. To maintain the model’s generalizability, we use a training strategy that allocates 10% probability to in-the-wild data and 90% to synthetic data.

4.3. 3D Reconstruction from Inconsistent Images

Once multi-view images are generated by the diffusion model, we can reconstruct a 3D Gaussian human model in canonical space. However, due to subtle appearance variations across views in the generated images, directly applying 3D Gaussian Splatting (3DGS) optimization degrades the quality of the reconstructed avatar (see Fig. 3).

Problem Formulation To address the challenges of 3D reconstruction from inconsistent views, it is necessary to handle the shape and appearance variations in each view. Viewing these inconsistencies as analogous to dynamic variations within a temporal sequence, we can reformulate the problem of 3D reconstruction from inconsistent images as a 4D reconstruction task. Inspired by the recent success of 4D Gaussian Splatting in dynamic scene modeling [47, 73, 79], we adopt a 4DGS approach to achieve efficient optimization and rendering.

The 4DGS framework is composed of a canonical space and a per-frame deformation module. The canonical space represents a static 3D model (*e.g.*, defined as the first frame), while the per-frame deformation module estimates shape and color variations of each 3D Gaussian, conditioned on the frame index, to fit the video sequence.

Our goal is to optimize both the canonical space and the deformation module based on the generated inconsistent images. Once optimized, this process yields a multi-view consistent Gaussian avatar model representing the shape in canonical space.

4D Gaussian Splatting Model Following existing dynamic Gaussian splatting methods [73], the deformation of 3D Gaussians is modeled by a deformation field network. We employ an efficient spatial-temporal encoder architecture consisting of a multi-resolution HexPlane and a compact MLP [9, 18, 65]. This structure encodes both the temporal and spatial features information of 3D Gaussians across six 2D voxel planes, incorporating temporal effects.

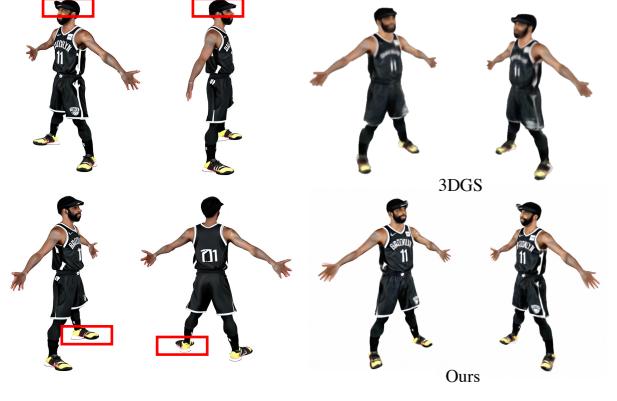


Figure 3. Inconsistencies caused by subtle variations in the generated multi-view images, which will degrade the 3D reconstruction quality. The red boxes highlights the inconsistent areas.

Once the 3D Gaussians features are encoded, separate MLPs are employed to compute the deformation for position $\Delta\mathbf{x}$, rotation $\Delta\mathbf{r}$, and scaling $\Delta\mathbf{s}$. Then, the deformed 3D Gaussian $(\mathbf{x}', \mathbf{r}', \mathbf{s}')$ can be expressed as:

$$(\mathbf{x}', \mathbf{r}', \mathbf{s}') = (\mathbf{x} + \Delta\mathbf{x}, \mathbf{r} + \Delta\mathbf{r}, \mathbf{s} + \Delta\mathbf{s}). \quad (3)$$

Shape Regularization In multi-view generation, both RGB images and normal maps are produced, allowing us to regularize the 4DGS optimization process using surface normal regularization. Specifically, we apply an L_1 loss on the rendered normals, $\mathcal{L}_{\text{normal}}$, to guide the optimization.

To mitigate the occurrence of spikes artifacts during avatar animation, which is often due to excessively large or elongated 3D Gaussian ellipsoids, we incorporate an anisotropy regularizer, \mathcal{L}_{ar} , to constrain the shape of the 3D Gaussians, following [75].

4DGS Optimization In line with existing reconstruction methods [33, 57, 64], we employ L1 loss for both color $\mathcal{L}_{\text{color}}$ and mask supervision $\mathcal{L}_{\text{mask}}$. We also include a grid-based total variation loss, \mathcal{L}_{tv} , as proposed in [73], to promote smooth deformation across views. For the human mask labels, we obtain pseudo ground-truth masks using SAM2 [59]. The total reconstruction loss is formulated as:

$$\mathcal{L}_r = \mathcal{L}_{\text{color}} + \lambda_m \mathcal{L}_{\text{mask}} + \lambda_n \mathcal{L}_{\text{normal}} + \lambda_{\text{tv}} \mathcal{L}_{\text{tv}} + \lambda_{\text{ar}} \cdot \mathcal{L}_{\text{ar}}, \quad (4)$$

where the weights for different loss terms are set as $\lambda_m = 0.1$, $\lambda_n = 0.05$, $\lambda_{\text{ar}} = 0.001$, and $\lambda_{\text{tv}} = 1$.

Point Cloud Initialization of 4DGS Point cloud initialization is critical for 3DGS optimization. We begin by generating a coarse mesh from the multi-view images and sample points on its surface to initialize the 3DGS points.

Specifically, we deform the predicted SMPL-X mesh to fit the generated multi-view RGB masks and normal maps.

We use the Nvidiffrast rasterizer [39] to render both the foreground and normal map of the deformed mesh. To reduce artifacts caused by inconsistencies in the multi-view images, we apply a Laplacian loss and an edge loss to regularize the mesh deformation. Further details are provided in the supplementary materials.

4.4. Animation

The reconstructed avatar is represented in a canonical space, aligned spatially with the canonical pose of the human body parametric model, SMPL-X. This alignment enables us to use SMPL-X driving parameters to animate the reconstructed avatar, making it fully animatable in 3D. To better handle cloth with large deformation, we apply a diffusion-based skinning method [42, 58], which transfers SMPL-X skinning weights throughout the entire human canonical space. During animation, skinning weights are obtained by querying the weights in space via bilinear interpolation. Additional details are provided in the supplementary materials.

5. Experiments

In this section, we thoroughly evaluate the effectiveness of our proposed methods by conducting a comprehensive comparison with state-of-the-art approaches.

Training Datasets For training the multi-view generation model, we first conduct a pretraining phase using a large dataset of dynamic human videos. Specifically, we collect approximately 200,000 dynamic human videos from various online sources. From this, we manually select single-person videos to create our in-the-wild training dataset, which consists of around 100,000 video samples. During the fine-tuning phase, we leverage a combination of public synthetic 3D datasets to render multi-view images. These datasets include 2K2K [19], Thuman2.0, Thuman2.1 [81], and CustomHumans [23], along with commercial datasets such as Thwindom and RenderPeople. Note that no rigged human models are used for training. In total, we utilize 6,124 synthetic human scans.

Inference Our model can generate an animatable 3D avatar in a canonical pose within a few minutes using a single RTX-3090 GPU. Specifically, it takes approximately 5 minutes to generate 30 frames of multi-view RGB and normal images. The 4DGS optimization process takes around 5 minutes. Once the multi-view reconstruction is complete, we set the time parameter $t = 0$ to obtain the final Gaussian point clouds. After optimization, the avatar can be animated and rendered in real time.

5.1. Comparison with Existing Methods

Evaluation Dataset We choose 50 rigged human avatars from Human4DiT [66] to evaluate our performance on multi-view canonical-pose image generation and 3D recon-

Table 1. Quantitative results on multi-view canonical image generation. ⁺ denotes using our generated canonical pose image as input, and ^{*} indicates that we fine-tune this approach on the multi-view synthetic dataset. Normal Loss is the cosine similarity loss between the predicted and ground-truth normal.

Methods	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	Normal Loss \downarrow
MagicMan ⁺ [85]	23.775	0.911	0.1077	0.134
CHAMP* [85]	20.025	0.894	0.1458	-
Ours	23.125	0.907	0.1023	0.101

Table 2. Quantitative comparison of 3D modeling on Human4DiT Synthetic Datasets.

Methods	LPIPS \downarrow	CLIP score \uparrow	FID \downarrow	User \uparrow
SiTH [82]	0.1607	86.854	86.895	2.134
LGM [71]	0.1567	86.686	82.121	2.617
MagicMan [21]	0.1479	82.693	144.642	2.095
CharacterGen [53]	0.1638	87.154	88.751	2.481
En3D [48]	0.1576	82.975	131.32	2.549
Ours	0.1085	90.370	77.879	4.199

struction metrics. For animation metrics, we use Blender software to obtain ground-truth video sequences and export the motion sequence to drive the created human models. We then compute photometric metrics in the foreground area to evaluate our performance on animation.

Baselines Our pipeline includes canonical multi-view generation, multi-view reconstruction, and human animation. For the multi-view generation task, we choose the state-of-the-art MagicMan [21] and CHAMP [85] as baselines. Notably, since CHAMP does not specifically train on self-rotated synthetic human datasets, we fine-tune this approach on our rendering datasets for a fair comparison.

For multi-view canonical human reconstruction tasks, we compare with the animatable human generation methods [48], CharacterGen [53]. Additionally, we also conducted comparison experiments with static clothed human reconstruction methods, including SiTH [24], MagicMan [21], and LGM [71] using our generated canonical pose image input. For human animation, we conduct comparison experiments with En3D and CharacterGen on the synthetic dataset.

Evaluation on Multi-view Canonical Generation As reported in Table 1, our method outperforms CHAMP in multi-view image generation, clearly demonstrating its effectiveness. It is important to note that MagicMan [21] is specifically designed for multi-view generation of static humans and cannot generate canonical pose images. Therefore, we use our generated canonical pose images as input for MagicMan. While MagicMan achieves better PSNR metrics, it is not suitable for animatable avatar generation. Additionally, our method achieves lower normal errors in the generated normal maps, further emphasizing its supe-



Figure 4. Visual comparison of animation results for the reconstructed 3D avatars. Best viewed with zoom-in.

rior performance.

Evaluation on Canonical Shape Reconstruction To assess the quality of canonical shape reconstruction, we render each reconstructed model from 24 different views and compute LPIPS and FID scores. The average CLIP score is computed by comparing the input RGB image with all rendered views.

Since this evaluation is closely related to generation tasks, we also conduct a user study to assess the similarity between the generated model and the ground truth. We randomly sample 20 cases and carry out an anonymous ranking survey with 40 participants across all baselines. The final metric represents the average ranking score, with a maximum score of 5 for each model. As shown in Table 2, our method consistently outperforms the baselines on all the metrics, justifying its design.

As demonstrated in Figure 5, our method produces significantly better canonical pose shape reconstructions than previous methods, with higher fidelity to the input image and enhanced sharpness in both texture and body shape details (e.g., clothing, faces, and hands). Note that since the texture refinement code for MagicMan [21] has not been released, the 3D reconstruction results for MagicMan do not include texture refinement.

Evaluation on Human Animation We compare the animation sequences of different methods with the rendered

Table 3. Quantitative results on human animation.

Methods	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow
CharacterGen [53]	17.570	0.644	0.205
En3D [21]	19.244	0.751	0.174
Ours	21.475	0.857	0.137

ground-truth sequences. As is demonstrated in Table 3, our method outperforms the baseline methods in terms of rendering quality in the animation sequences. Compared to the second-best method En3D, our method achieves improvements of 2.231, 0.106, and 0.037 in PSNR, SSIM, and LIPIS, respectively. As is visualized in Fig. 4, our method produces accurate and photorealistic animation results than the baseline methods. More results are included in the supplementary materials.

5.2. Ablation Study

Shape Regularization We conduct an ablation study to evaluate the design of the shape regularization. Specifically, Fig. 6 (a) demonstrates that normal regularization effectively reduces random noise and enhances surface details, while Fig. 6 (b) shows that anisotropic regularization helps eliminate spikes in novel pose animations.

Initialization of 4DGS Figure 7 compares the results of 4DGS optimization using random points, SMPL mesh, and



Figure 5. Visual comparison on canonical pose 3D avatar reconstruction from the single-view image. Since SiTH, MagicMan, and LGM cannot reconstruct canonical pose shapes from the input, we take our generated front-view canonical pose image as input to these methods.

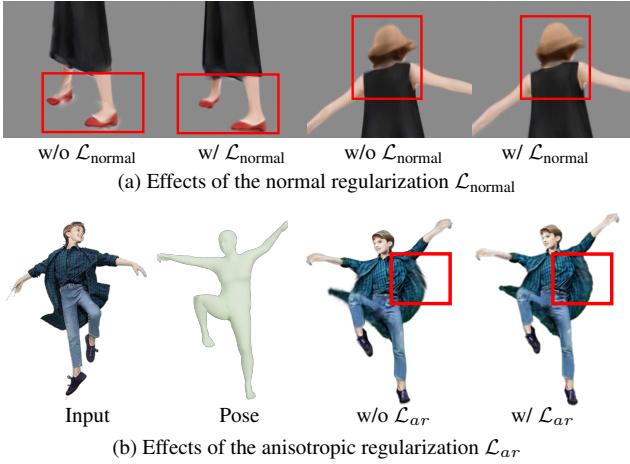


Figure 6. Ablation study for the shape regularization.

the proposed coarse mesh as initialization.

Without any shape prior, the Gaussian model initialized

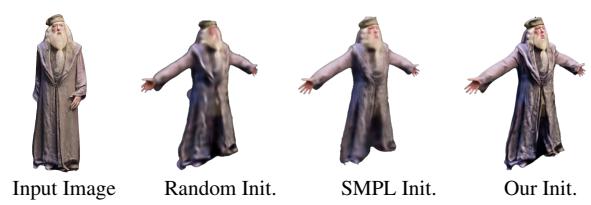


Figure 7. Ablation study for shape initialization strategy.

with random points struggles to optimize, resulting in blurry outputs. Initializing with the SMPL mesh provides a better shape prior for body representation but faces difficulties in accurately modeling points that are farther from the body. In contrast, the coarse mesh initialization provides a good starting point for the optimization.

6. Conclusion

In this work, we present a robust approach to generate animatable human avatars from a single image. We introduce

a reference image-guided video generation model to produce high-quality multi-view canonical human images and their corresponding normal maps. To handle view inconsistencies, we propose a 4D Gaussian Splatting (4DGS)-based method for reconstructing high-fidelity 3D avatars. Comprehensive evaluation demonstrates that our method enables photorealistic, real-time animation of 3D human avatars from in-the-wild images.

Limitations and Future Work While our method supports real-time inference, it still requires several minutes to optimize an animatable avatar. In future work, we aim to explore feed-forward 3D reconstruction techniques that are robust to multi-view inconsistencies.

References

- [1] actorcore. accurig, a software for automatic character rigging, 2023. 3
- [2] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Detailed human avatars from monocular video. In *3DV*, 2018. 2
- [3] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Video based reconstruction of 3d people models. In *CVPR*, 2018. 2, 3
- [4] Thiemo Alldieck, Marcus Magnor, Bharat Lal Bhatnagar, Christian Theobalt, and Gerard Pons-Moll. Learning to reconstruct people in clothing from a single rgb camera. In *CVPR*, 2019. 2
- [5] Thiemo Alldieck, Gerard Pons-Moll, Christian Theobalt, and Marcus Magnor. Tex2shape: Detailed full human body geometry from a single image. In *CVPR*, 2019. 2
- [6] Thiemo Alldieck, Mihai Zanfir, and Cristian Sminchisescu. Photorealistic monocular 3d reconstruction of humans wearing clothing. In *CVPR*, 2022. 2
- [7] Fabien Baradel*, Matthieu Armando, Salma Galaaoui, Romain Brégier, Philippe Weinzaepfel, Grégory Rogez, and Thomas Lucas*. Multi-hmr: Multi-person whole-body human mesh recovery in a single shot. In *ECCV*, 2024. 5
- [8] Bharat Lal Bhatnagar, Garvita Tiwari, Christian Theobalt, and Gerard Pons-Moll. Multi-garment net: Learning to dress 3d people from images. In *CVPR*, 2019. 2
- [9] Ang Cao and Justin Johnson. Hexplane: A fast representation for dynamic scenes. *CVPR*, 2023. 5
- [10] Yukang Cao, Guanying Chen, Kai Han, Wenqi Yang, and Kwan-Yee K Wong. Jiff: Jointly-aligned implicit face function for high quality single view clothed human reconstruction. In *CVPR*, 2022. 2
- [11] Yukang Cao, Yan-Pei Cao, Kai Han, Ying Shan, and Kwan-Yee K Wong. Dreamavatar: Text-and-shape guided 3d human avatar generation via diffusion models. In *CVPR*, 2024. 3
- [12] Mingjin Chen, Junhao Chen, Xiaojun Ye, Huan-ang Gao, Xiaoxue Chen, Zhaoxin Fan, and Hao Zhao. Ultraman: Single image 3d human reconstruction with ultra speed and detail. *arXiv preprint arXiv:2403.12028*, 2024. 2
- [13] Yushuo Chen, Zerong Zheng, Zhe Li, Chao Xu, and Yebin Liu. Meshavatar: Learning high-quality triangular human avatars from multi-view videos. *arXiv preprint arXiv:2407.08414*, 2024. 3
- [14] Ye Chongjie, Qiu Lingteng, Gu Xiaodong, Zuo Qi, Wu Yushuang, Dong Zilong, Bo Liefeng, Xiu Yuliang, and Han Xiaoguang. Stablenormal: Reducing diffusion variance for stable and sharp normal. *TOG*, 2024. 4
- [15] Vasileios Choutas, Lea Muller, Chun-Hao P. Huang, Siyu Tang, Dimitrios Tzionas, and Michael J. Black. Accurate 3d body shape regression using metric and semantic attributes. In *CVPR*, 2022. 2
- [16] Enric Corona, Mihai Zanfir, Thimo Alldieck, Eduard Gabriel Bazavan, Andrei Zanfir, and Cristian Sminchisescu. Structured 3d features for reconstructing controllable avatars. In *CVPR*, 2023. 2
- [17] 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 *ICCV*, 2023. 2
- [18] Jiemin Fang, Taoran Yi, Xinggang Wang, Lingxi Xie, Xiaopeng Zhang, Wenyu Liu, Matthias Nießner, and Qi Tian. Fast dynamic radiance fields with time-aware neural voxels. In *SIGGRAPH Asia*, 2022. 5
- [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 *CVPR*, 2023. 6, 12
- [20] Tong He, Yuanlu Xu, Shunsuke Saito, Stefano Soatto, and Tony Tung. Arch++: Animation-ready clothed human reconstruction revisited. In *ICCV*, 2021. 2, 3
- [21] Xu He, Xiaoyu Li, Di Kang, Jiangnan Ye, Chaopeng Zhang, Liyang Chen, Xiangjun Gao, Han Zhang, Zhiyong Wu, and Haolin Zhuang. Magicman: Generative novel view synthesis of humans with 3d-aware diffusion and iterative refinement. *arXiv preprint arXiv:2408.14211*, 2024. 2, 6, 7
- [22] Ajay Ho, Jonathan dand Jain and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 2020. 3
- [23] Hsuan-I Ho, Lixin Xue, Jie Song, and Otmar Hilliges. Learning locally editable virtual humans. In *CVPR*, 2023. 6, 12
- [24] I Ho, Jie Song, Otmar Hilliges, et al. Sith: Single-view textured human reconstruction with image-conditioned diffusion. In *CVPR*, 2024. 6
- [25] Fangzhou Hong, Zhaoxi Chen, Yushi Lan, Liang Pan, and Ziwei Liu. Eva3d: Compositional 3d human generation from 2d image collections. In *ICLR*, 2023. 2
- [26] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022. 2, 3, 4
- [27] Li Hu. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *CVPR*, 2024. 2, 3, 4
- [28] Yukun Huang, Jianan Wang, Ailing Zeng, He Cao, Xianbiao Qi, Yukai Shi, Zheng-Jun Zha, and Lei Zhang. Dreamwaltz: Make a scene with complex 3d animatable avatars. *NeurIPS*, 2024. 3

- [29] Zeng Huang, Yuanlu Xu, Christoph Lassner, Hao Li, and Tony Tung. Arch: Animatable reconstruction of clothed humans. In *CVPR*, 2020. 2, 3
- [30] Boyi Jiang, Juyong Zhang, Yang Hong, Jinhao Luo, Ligang Liu, and Hujun Bao. Bcnets: Learning body and cloth shape from a single image. In *ECCV*, 2020. 2
- [31] Boyi Jiang, Yang Hong, Hujun Bao, and Juyong Zhang. Selfrecon: Self reconstruction your digital avatar from monocular video. In *CVPR*, 2022. 3
- [32] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *CVPR*, 2018. 2
- [33] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *TOG*, 2023. 3, 5
- [34] Rawal Khirodkar, Timur Bagautdinov, Julieta Martinez, Su Zhaoen, Austin James, Peter Selednik, Stuart Anderson, and Shunsuke Saito. Sapiens: Foundation for human vision models. In *ECCV*, 2024. 5, 12
- [35] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, 2014. 12
- [36] Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. Vibe: Video inference for human body pose and shape estimation. In *CVPR*, 2020. 2
- [37] Nikos Kolotouros, Georgios Pavlakos, Michael J. Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *ICCV*, 2019. 2
- [38] Nikos Kolotouros, Thimo Alldieck, Andrei Zanfir, Eduard Bazavan, Mihai Fieraru, and Cristian Sminchisescu. Dreamhuman: Animatable 3d avatars from text. *NeurIPS*, 2024. 3
- [39] Samuli Laine, Janne Hellsten, Tero Karras, Yeongho Seol, Jaakko Lehtinen, and Timo Aila. Modular primitives for high-performance differentiable rendering. *TOG*, 2020. 6
- [40] Verica Lazova, Eldar Insafutdinov, and Gerard Pons-Moll. 360-degree textures of people in clothing from a single image. In *3DV*, 2019. 2
- [41] Zhe Li, Zerong Zheng, Lizhen Wang, and Yebin Liu. Animatable gaussians: Learning pose-dependent gaussian maps for high-fidelity human avatar modeling. In *CVPR*, 2024. 3
- [42] Siyou Lin, Hongwen Zhang, Zerong Zheng, Ruizhi Shao, and Yebin Liu. Learning implicit templates for point-based clothed human modeling. In *ECCV*, 2022. 6, 12
- [43] Shiyu Liu, Zibo Zhao, Yihao Zhi, Yiqun Zhao, Binbin Huang, Shuo Wang, Ruoyu Wang, Michael Xuan, Zhengxin Li, and Shenghua Gao. Heromaker: Human-centric video editing with motion priors. In *ACM Multimedia 2024*, 2024. 3
- [44] Zhibin Liu, Haoye Dong, Aviral Chharia, and Hefeng Wu. Human-vdm: Learning single-image 3d human gaussian splatting from video diffusion models. *arXiv preprint arXiv:2409.02851*, 2024. 2
- [45] Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, et al. Wonder3d: Single image to 3d using cross-domain diffusion. In *CVPR*, 2024. 2, 4
- [46] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: a skinned multi-person linear model. *TOG*, 2015. 2, 3
- [47] Jonathon Luiten, Georgios Kopanas, Bastian Leibe, and Deva Ramanan. Dynamic 3d gaussians: Tracking by persistent dynamic view synthesis. In *3DV*, 2024. 2, 3, 5
- [48] Yifang Men, Biwen Lei, Yuan Yao, Miaomiao Cui, Zhouhui Lian, and Xuansong Xie. En3d: An enhanced generative model for sculpting 3d humans from 2d synthetic data. In *CVPR*, 2024. 2, 6
- [49] Yifang Men, Yuan Yao, Miaomiao Cui, and Liefeng Bo. Mimo: Controllable character video synthesis with spatial decomposed modeling. *arXiv preprint arXiv:2409.16160*, 2024. 3
- [50] 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. *arXiv preprint arXiv:2406.12459*, 2024. 2
- [51] 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 *CVPR*, 2019. 3, 12
- [52] William Peebles and Saining Xie. Scalable diffusion models with transformers. *arXiv preprint arXiv:2212.09748*, 2022. 3
- [53] Hao-Yang Peng, Jia-Peng Zhang, Meng-Hao Guo, Yan-Pei Cao, and Shi-Min Hu. Charactergen: Efficient 3d character generation from single images with multi-view pose canonicalization. *TOG*, 2024. 2, 3, 6, 7
- [54] Sida Peng, Junting Dong, Qianqian Wang, Shangzhan Zhang, Qing Shuai, Xiaowei Zhou, and Hujun Bao. Animatable neural radiance fields for modeling dynamic human bodies. In *ICCV*, 2021. 3
- [55] Gerard Pons-Moll, Sergi Pujades, Sonny Hu, and Michael J. Black. Clothcap. *TOG*, 2017. 2
- [56] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *ICLR*, 2023. 3
- [57] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. *arXiv preprint arXiv:2011.13961*, 2020. 5
- [58] Lingteng Qiu and Guanying Chen. Rec-mv: Reconstructing 3d dynamic cloth from monocular videos. In *CVPR*, 2023. 3, 6
- [59] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 5
- [60] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 3
- [61] 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 *ICCV*, 2019. 2
- [62] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *CVPR*, 2020. 2
- [63] Shunsuke Saito, Jinlong Yang, Qianli Ma, and Michael J. Black. Scanimate: Weakly supervised learning of skinned clothed avatar networks. In *CVPR*, 2021. 2
- [64] Sara Fridovich-Keil and Giacomo Meanti, Frederik Rahbæk Warburg, Benjamin Recht, and Angjoo Kanazawa. K-planes: Explicit radiance fields in space, time, and appearance. In *CVPR*, 2023. 5
- [65] Ruizhi Shao, Zerong Zheng, Hanzhang Tu, Boning Liu, Hongwen Zhang, and Yebin Liu. Tensor4d: Efficient neural 4d decomposition for high-fidelity dynamic reconstruction and rendering. In *CVPR*, 2023. 5
- [66] Ruizhi Shao, Youxin Pang, Zerong Zheng, Jingxiang Sun, and Yebin Liu. Human4dit: 360-degree human video generation with 4d diffusion transformer. *TOG*, 2024. 2, 3, 6
- [67] Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. In *ICLR*, 2023. 2
- [68] David Smith, Matthew Loper, Xiaochen Hu, Paris Mavroidis, and Javier Romero. Facsimile: Fast and accurate scans from an image in less than a second. In *ICCV*, 2019. 2
- [69] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *ICLR*, 2021. 3
- [70] Yu Sun, Wu Liu, Qian Bao, Yili Fu, Tao Mei, and Michael J. Black. Putting people in their place: Monocular regression of 3d people in depth. In *CVPR*, 2022. 2
- [71] 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. In *ECCV*, 2025. 6
- [72] Chung-Yi Weng, Brian Curless, Pratul P Srinivasan, Jonathan T Barron, and Ira Kemelmacher-Shlizerman. Hu-mannerf: Free-viewpoint rendering of moving people from monocular video. In *CVPR*, 2022. 3
- [73] Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Xinggang Wang. 4d gaussian splatting for real-time dynamic scene rendering. In *CVPR*, 2024. 2, 3, 5
- [74] Donglai Xiang, Fabian Prada, Chenglei Wu, and Jessica Hodgins. Monoclothcap: Towards temporally coherent clothing capture from monocular rgb video. In *3DV*, 2020. 2
- [75] Tianyi Xie, Zeshun Zong, Yuxing Qiu, Xuan Li, Yutao Feng, Yin Yang, and Chenfanfu Jiang. Physgaussian: Physics-integrated 3d gaussians for generative dynamics. In *CVPR*, 2024. 5
- [76] Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J. Black. Icon: Implicit clothed humans obtained from normals. In *CVPR*, 2022. 2
- [77] Yuliang Xiu, Jinlong Yang, Xu Cao, Dimitrios Tzionas, and Michael J. Black. Econ: Explicit clothed humans optimized via normal integration. In *CVPR*, 2023. 2
- [78] Zhuoqian Yang, Shikai Li, Wayne Wu, and Bo Dai. 3dhumangan: 3d-aware human image generation with 3d pose mapping. In *ICCV*, 2023. 2
- [79] Ziyi Yang, Xinyu Gao, Wen Zhou, Shaohui Jiao, Yuqing Zhang, and Xiaogang Jin. Deformable 3d gaussians for high-fidelity monocular dynamic scene reconstruction. In *CVPR*, 2024. 2, 3, 5
- [80] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024. 3, 4
- [81] 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 *CVPR*, 2021. 6, 12
- [82] Zechuan Zhang, Zongxin Yang, and Yi Yang. Sifu: Side-view conditioned implicit function for real-world usable clothed human reconstruction. In *CVPR*, 2024. 2, 6
- [83] Zerong Zheng, Tao Yu, Yebin Liu, and Qionghai Dai. Pamir: Parametric model-conditioned implicit representation for image-based human reconstruction, 2020. 2
- [84] Hao Zhu, Xinxin Zuo, Sen Wang, Xun Cao, and Ruigang Yang. Detailed human shape estimation from a single image by hierarchical mesh deformation. In *CVPR*, 2019. 2
- [85] Shenhao Zhu, Junming Leo Chen, Zuozhuo Dai, Yinghui Xu, Xun Cao, Yao Yao, Hao Zhu, and Siyu Zhu. Champ: Controllable and consistent human image animation with 3d parametric guidance. *ECCV*, 2024. 2, 3, 4, 6

A. Demo Video

Please kindly check the [Demo Video](#) for animation results of the reconstructed 3D avatar.

B. More Details for the Method

B.1. Implementation Details

Multi-view Generation

For training the multi-view canonical image generation model, we first pre-train our RGB-Normal DiT model on in-the-wild video clips. To supervise the normal map output, we utilize Sapiens [34], an off-the-shelf normal estimation prior, to generate pseudo ground-truth normals from in-the-wild data. The model is trained using the Adam optimizer [35] with a learning rate of 2×10^{-4} and a batch size of 1. We employ 16 Nvidia A100 80G GPUs for training, with the pre-training process comprising 100,000 optimization iterations. Subsequently, the model is fine-tuned on a synthetic dataset using the same hyperparameters, performing an additional 50,000 iterations of optimization. To preserve the model’s generalizability, we adopt a data-mixing strategy during fine-tuning, assigning a 10% probability to sampling in-the-wild data and a 90% probability to synthetic data.

3D Reconstruction from Inconsistent Images. In the multi-view reconstruction phase, after obtaining the deformed coarse mesh from the original SMPL-X as the initialization for 4DGS, we first performed 3,000 iterations of optimization the 3DGS parameters. Sequentially, we continue to conduct 4,000 iterations of optimization in the temporal dimension to address multi-view inconsistency. In the multi-view reconstruction phase, we initialize with a deformed coarse mesh derived from the original SMPL-X model for the 4DGS process. The first step is optimizing the 3DGS parameters over 3,000 iterations. Subsequently, we perform 4,000 iterations of optimization considering the temporal dimension to address multi-view inconsistency.

B.2. RGB-Normal Diffusion Transformer

Figure 8 illustrates the architecture of our multi-view diffusion transformer model for canonical image and normal map generation. For simplicity, we omit SPML-X conditioning in the figure. Both ‘I-DiT-E’ and ‘N-DiT-E’ denote two independent DiT encoder blocks conditioned on image and normal input, respectively, while ‘I-DiT-D’ and ‘N-DiT-D’ refer to two independent decoders responsible for generating multi-view canonical images and normal maps. Additionally, ‘I-N’ within the intermediate DiT blocks represents a multi-modal attention module that effectively encodes joint image and normal features.

B.3. Coarse Shape Initialization

We optimize the following objective function to obtain the initial coarse mesh \mathcal{M}' for 3DGS initialization:

$$\begin{aligned} \mathcal{L}_{init} = & \lambda_{mask} \cdot \mathcal{L}_{mask} + \lambda_n \cdot \mathcal{L}_{normal} \\ & + \lambda_{lap} \cdot \mathcal{L}_{lap}(\mathcal{M}') + \lambda_{edge} \cdot \mathcal{L}_{edge}(\mathcal{M}'). \end{aligned} \quad (5)$$

where $\lambda_{mask} = 1.0$, $\lambda_n = 0.5$, $\lambda_{lap} = 0.1$, and $\lambda_{edge} = 0.05$.

Figure 9 demonstrates the coarse mesh results reconstructed from the generated images. As illustrated in the figure, the coarse mesh provides only a rough geometric surface, with several noticeable artifacts remaining on its surface.

B.4. Skinning-based Animation

We model large body motions using linear blend skinning (LBS) transformations based on the SMPL-X [51] model. Specifically, given an SMPL body with shape parameter β and pose parameter θ_i in the i -th frame, a point p on the body surface in canonical space with skinning weights $w(p)$ can be warped to camera view space via the skinning transformation W .

Notably, the skinning weights $w(p)$ are only defined for points on the SMPL-X surface. To handle shapes with large deformations (e.g., skirts) and to better facilitate the warping of arbitrary points in canonical space to the camera view, we employ the diffused skinning strategy [42] to propagate the skinning weights of the SMPL-X body vertices to the entire canonical space. These weights are stored in a voxel grid of size $256 \times 256 \times 256$. Skinning weights for arbitrary points are then obtained through trilinear interpolation.

B.5. More Details for the Synthetic Dataset

We leverage a combination of public synthetic 3D datasets to render multi-view images for fine-tuning the multi-view canonical image and normal generation model. These datasets include 2K2K [19], Thuman2.0, Thuman2.1 [81], and CustomHumans [23], along with commercial datasets such as Thwindom and RenderPeople. In total, we utilize 6,124 synthetic human scans.

For the synthetic data, we render each object from 30 different viewpoints by rotating the object. To improve the quality of multi-view reconstruction, images are rendered at varying elevations, which helps to regularize the optimization of the 3D Gaussian Splatting (3DGS) method. Specifically, the elevation range oscillates between -20° and 20° , following a sine function over a cycle of 30 views.

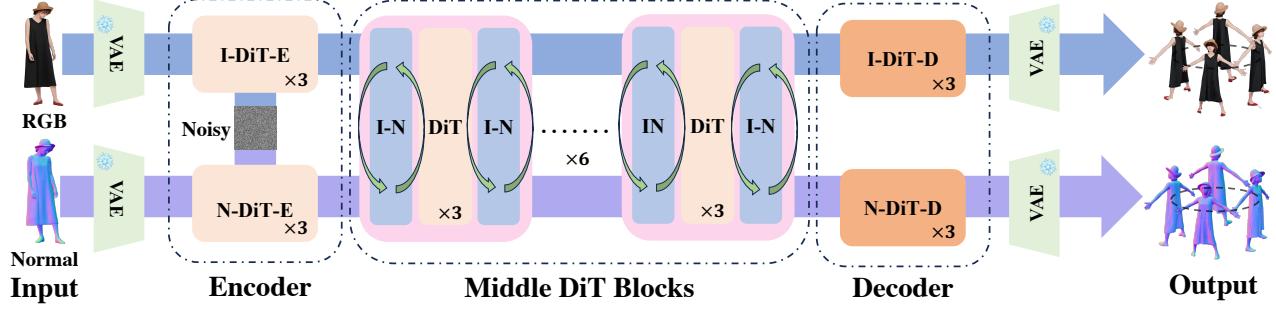


Figure 8. The architecture of the joint RGB-Normal Diffusion Transformer designed for generating multi-view canonical images and normal maps. For simplicity, SPML-X conditioning is omitted from the depiction.

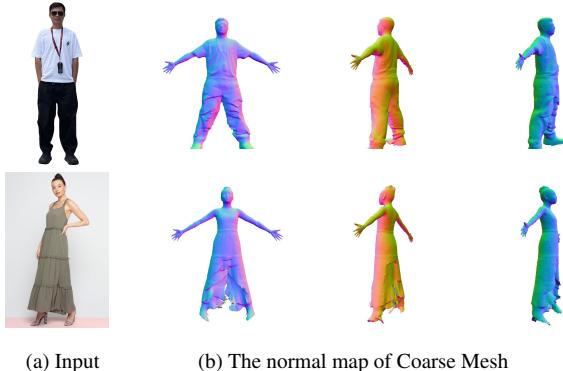


Figure 9. Sample results for the about coarse mesh reconstruction from multi-view images.

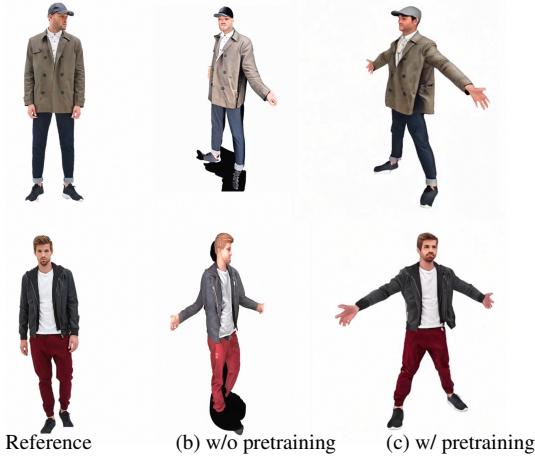


Figure 10. Effectiveness of pre-training on in-the-wild videos.

C. More Results

C.1. Pre-training on In-the-wild Data

Figure 10 underscores the critical role of pre-training on in-the-wild data. Models pre-trained on diverse and real-world datasets demonstrate substantially enhanced gener-

alization capabilities compared to models trained without pre-training, verifying the training strategy of our method.

C.2. Animation Results

Figure 11–Figure 12 showcase the animation results of input human images with diverse appearances and a wide range of poses. Our method demonstrates the ability to generate animations that are both robust and photorealistic, preserving fine details of the human appearance while ensuring smooth and natural motion transitions. These results highlight the generalizability and effectiveness of our approach in handling varying levels of complexity in human avatars.

C.3. Reconstruction and Animation from Any Input

Figure 15 and Fig. 16 illustrate reconstructions and animation results from a diverse set of images collected from the internet. Notably, the reference image is a non-human image input, demonstrating the model’s still maintain original diffusion model’s generalizability.

C.4. Canonical Shape Reconstruction

To further validate the effectiveness of the proposed method, we provide additional results for canonical shape reconstruction from single images. Figure 17–Fig. 19 present reconstruction results on the DeepFashion dataset, showcasing accurate recovery of canonical shapes from fashion images. Meanwhile, Figure 20–Fig. 22 illustrate reconstructions from a diverse set of images collected from the internet, demonstrating the model’s adaptability to various image sources and styles.



Reference

Animation results

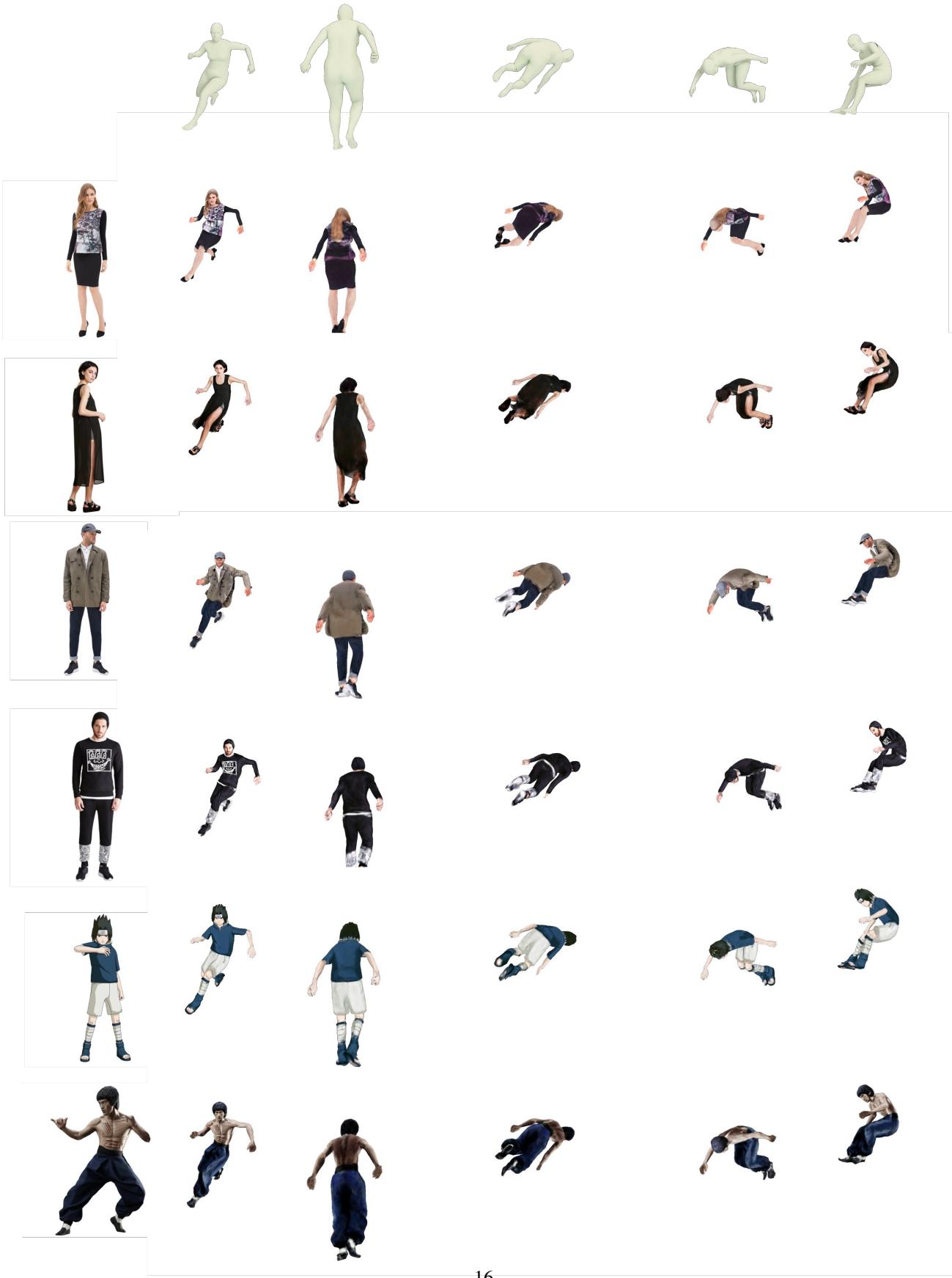
Figure 11. Visual results of human animation results (Part I) from any input. Best viewed with zoom-in.



Reference

Animation results

Figure 12. Visual results of human animation results (Part II) from any input. Best viewed with zoom-in.



Reference

16
Animation results

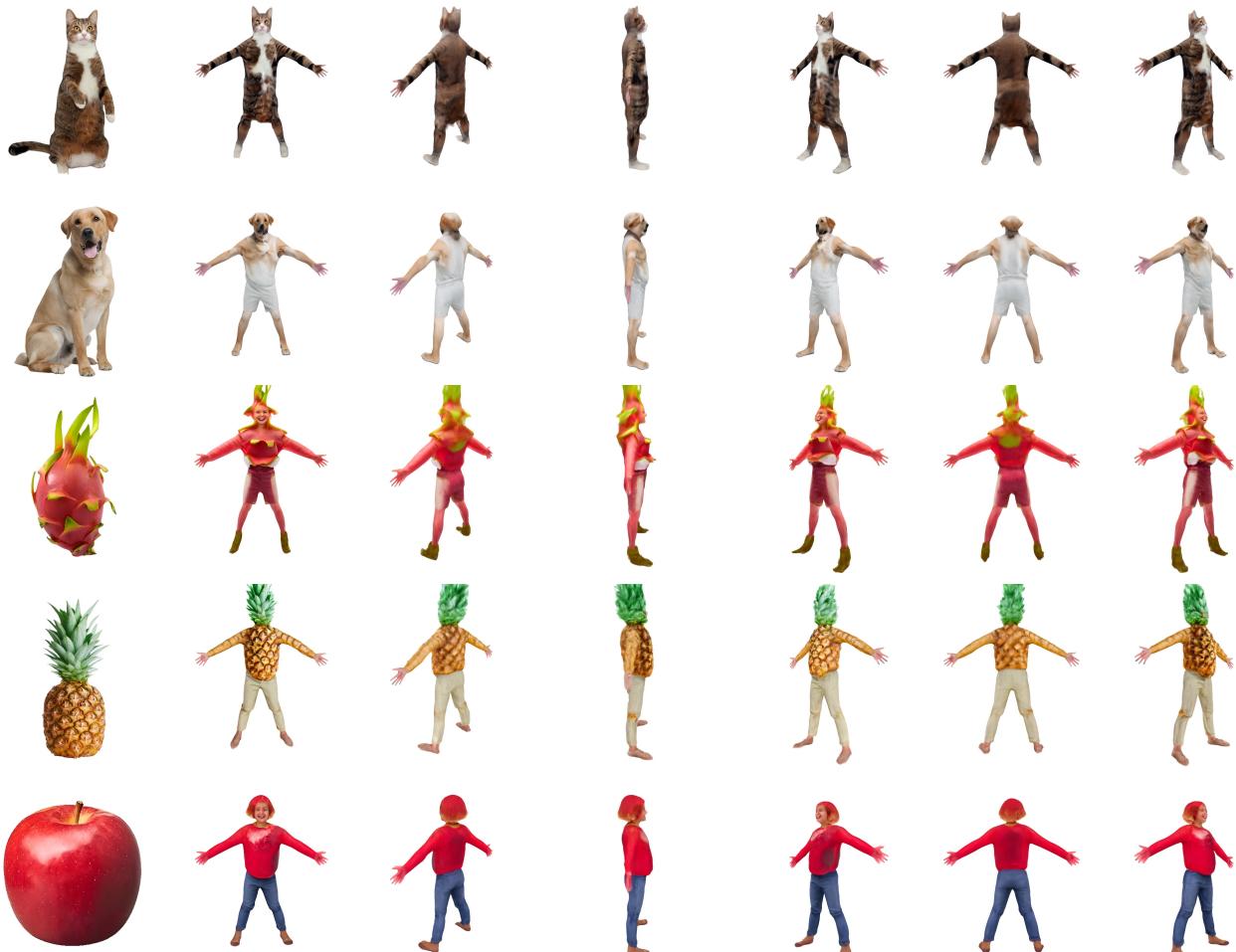
Figure 13. Visual results of human animation results (Part III) from any input. Best viewed with zoom-in.



Reference

Animation results

Figure 14. Visual results of human animation results (Part IV) from any input. Best viewed with zoom-in.



Reference

Multi-view Reconstruction

Figure 15. Visual results of canonical shape reconstruction from “Any Input”. Best viewed with zoom-in.



Figure 16. Visual results of canonical shape reconstruction from “Any Input”. Best viewed with zoom-in.



Reference

Multi-view Reconstruction

Figure 17. Visual results of canonical shape reconstruction (Part I). Best viewed with zoom-in.



Reference

Multi-view Reconstruction

Figure 18. Visual results of canonical shape reconstruction (Part II). Best viewed with zoom-in.



Reference

Multi-view Reconstruction

Figure 19. Visual results of canonical shape reconstruction (Part III). Best viewed with zoom-in.



Figure 20. Visual results of canonical shape reconstruction (Part IV). Best viewed with zoom-in.



Reference

Multi-view Reconstruction

Figure 21. Visual results of canonical shape reconstruction (Part V). Best viewed with zoom-in.



Reference

Multi-view Reconstruction

Figure 22. Visual results of canonical shape reconstruction (Part VI). Best viewed with zoom-in.