MagicMan: Generative Novel View Synthesis of Humans with 3D-Aware Diffusion and Iterative Refinement

Xu He[1], Zhiyong Wu[1,5*], Xiaoyu Li[2*], Di Kang[2], Chaopeng Zhang[2], Jiangnan Ye[1], Liyang Chen[1], Xiangjun Gao[3], Han Zhang[4], Haolin Zhuang[1]

[1]Shenzhen International Graduate School, Tsinghua University
[2]Tencent
[3]The Hong Kong University of Science and Technology
[4]Stanford University
[5]The Chinese University of Hong Kong

hex22@mails.tsinghua.edu.cn, zywu@sz.tsinghua.edu.cn, xliea@connect.ust.hk, di.kang@outlook.com, cpz@pku.edu.cn

Abstract

Existing works in single-image human reconstruction suffer from weak generalizability due to insufficient training data or 3D inconsistencies for a lack of comprehensive multi-view knowledge. In this paper, we introduce MagicMan, a human-specific multi-view diffusion model to generate high-quality novel views from a single reference image. As its core, we leverage a pre-trained 2D diffusion model as the generative prior for generalizability, with the parametric SMPL-X model as the 3D body prior to promote 3D awareness. To maintain consistency while generating denser views for improved 3D human reconstruction, we introduce hybrid multi-view attention to facilitate efficient and thorough information interchange across views. Besides, we present a geometry-aware dual branch to perform concurrent generation in both RGB and normal domains, further enhancing consistency via geometry cues. Last but not least, to address ill-shaped issues arising from inaccurate SMPL-X estimation, we propose a novel iterative refinement strategy, which progressively optimizes SMPL-X accuracy while enhancing the quality and consistency of the generated multi-views. Extensive experimental results demonstrate that our method significantly outperforms existing approaches in both novel view synthesis and subsequent 3D human reconstruction tasks. Code and demos are available at https://thuhcsi.github.io/MagicMan.

Code — https://thuhcsi.github.io/MagicMan
Extended version — https://arxiv.org/abs/2408.14211

Figure 1: Given a reference human image in different poses, outfits, or styles (i.e. real and fictional characters) as input, MagicMan is able to generate consistent high-quality novel view images and normal maps, which are well-suited for downstream multi-view reconstruction applications.

1 Introduction

3D digital human creation is an important technique in computer vision and graphics. Traditional methods (Balan et al. 2007; Vlasic et al. 2009) usually utilize a dense camera array to capture synchronized posed multi-view images for human reconstruction. However, these methods typically require a tedious and time-consuming process which is not practical for general users. Therefore, creating 3D human models from a single image is a significant task to be investigated.

To this end, early works like PIFu (2019), PaMIR (2021), and ICON (2022) have been introduced to train feed-forward networks on scanned 3D human datasets, capable of reconstructing 3D humans from a single image. Nevertheless, these data are scarce with limited diversity, resulting in poor generalizability to varied poses and outfits. Besides, the weak generative capability from insufficient data also leads to overly-smoothed geometry and blurred textures.

Thanks to the abundant priors in text-to-image diffusion models trained on large-scale data, another category of works (Zhang et al. 2024; Gao et al. 2024) leverage pre-trained diffusion models to optimize 3D representations through a Score Distillation Sampling (SDS) loss to create human models from a single image. Although these methods yield impressive results in generalization and detailed textures, the absence of 3D awareness frequently leads to 3D inconsistencies like multi-face Janus problem (2022). More-

*Corresponding authors.

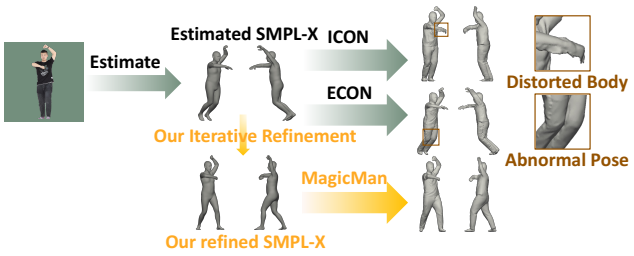

Figure 2: Ill-shaped geometry from inaccurate SMPL-X can be alleviated through our proposed iterative refinement.

over, text descriptions or CLIP embeddings, which are typically used to guide SDS, contain only global semantic information and thereby hinder the generation of fine-grained textures consistent with the reference (Huang et al. 2024).

In the field of 3D object generation, as exemplified by Zero123 (2023a), efforts have been made to use image diffusion models trained on multi-view data to directly generate novel views from a single image, which are subsequently used to reconstruct 3D models, achieving promising results.

Following this paradigm, we present a multi-view diffusion model to directly generate human novel views, encompassing both generalizability and 3D knowledge. Three challenges emerge: The primary is to ensure multi-view consistency as emphasized in 3D-object-related works. Both intricate geometry and detailed textures of humans pose greater difficulties. Secondly, existing works enhance multi-view consistency either by extending 2D self-attention in image diffusion models to 3D attention across all views (Shi et al. 2023), or by integrating 3D representations (Liu et al. 2023b). Both approaches are memory-consuming and only generate sparse views, insufficient for reliable 3D human reconstruction where self-occlusion typically occurs. Finally, most research (Zheng et al. 2021; Liu et al. 2024; Huang et al. 2024) demonstrates the importance of using parametric models, e.g., SMPL (-X) (2015; 2019), as 3D body priors in human-related tasks, which is also found to be beneficial for consistency and robustness in our work. However, SMPL-X estimated from monocular images often display inaccurate poses, featuring depth ambiguities and misalignment with the reference, which will lead to ill-shaped geometry in reconstructed human models, manifesting as abnormal overall poses or distorted body parts as illustrated in Fig. 2.

In this paper, we take the above challenges into account and propose **MagicMan** to produce dense, consistent, and pose-accurate multi-view human images from a single reference image. We utilize the powerful Stable Diffusion (SD) (Rombach et al. 2022) equipped with SMPL-X guidance as the backbone. To obtain consistent yet dense novel views, we introduce a hybrid multi-view attention mechanism to establish connections between different views. Specifically, we combine efficient 1D attention across all views and 3D attention spanning pixels of a sparse subset of selected views to enhance information interchange with minimal memory overhead, enabling us to generate significantly denser (i.e., 20 views at 512 resolution) novel views while maintaining consistency. Next, to deal with detailed geometry, we propose a geometry-aware dual branch with shared blocks to simultaneously generate aligned novel view RGB images and normal maps. The normal map prediction supplements structure information, further improving the consistency of geometric details in the RGB domain. Last but not least, to address the ill-shaped issues arising from inaccurate SMPL-X estimates, we propose an iterative refinement approach, wherein intermediate generated multi-views are employed to refine the SMPL-X parameters for more accurate poses, which in turn guide the multi-view generation process with more consistent and robust results.

To summarize, our main contributions are as follows:

- We present **MagicMan**, a method designed to generate high-quality multi-view images for humans from a single reference image, thereby facilitating seamless 3D human reconstruction.
- We propose an efficient hybrid multi-view attention to generate denser multi-view human images while maintaining better 3D consistency.
- A geometry-aware dual branch is introduced to perform generation in both RGB and normal domains, further enhancing multi-view consistency via geometry cues.
- An iterative refinement strategy is proposed to progressively enhance both the SMPL-X pose accuracy and the generated multi-view consistency, reducing ill-shaped issues arising from unreliable SMPL-X estimation.

2 Related Work

Our work focuses on the generative novel view synthesis using diffusion models, which is related to diffusion models, generative view synthesis, and human image synthesis.

Diffusion Models. Diffusion models (2015; 2020) have demonstrated promising results in recent image synthesis works (Shen et al. 2024). Ho and Salimans performs classifier-free guidance by combining the score estimates from conditional and unconditional generation. Based on diffusion models and large-scale data, various image synthesis tasks have achieved impressive results such as text-to-image synthesis (Saharia et al. 2022; Ramesh et al. 2022; Rombach et al. 2022; Xian et al. 2024). This paper explores the use of diffusion models for human novel view synthesis.

Generative Novel View Synthesis. Generative novel view synthesis requires synthesizing views far beyond the input. Compared with traditional regression-based methods (Yu et al. 2021a), it only has sparse or a single view as input to hallucinate unseen parts. With the development of generative models, this task has been studied by utilizing generative adversarial networks (Li et al. 2022) and transformer (Kulhánek et al. 2022). Recently, diffusion models have also been applied to this task. Zero-1-to-3 (2023a) proposes a viewpoint-conditioned diffusion model trained on a large-scale dataset that shows strong zero-shot generalizability. To incorporate 3D awareness into 2D diffusion models, GeNVS (2023) and SyncDreamer (2023b) use a 3D feature as condition. Tseng et al. uses an epipolar attention for consistent view synthesis. Wonder3D (2024) extends diffusion models with cross-domain attention and SV3D (2024) utilizes a video diffusion model to improve consistency. In this

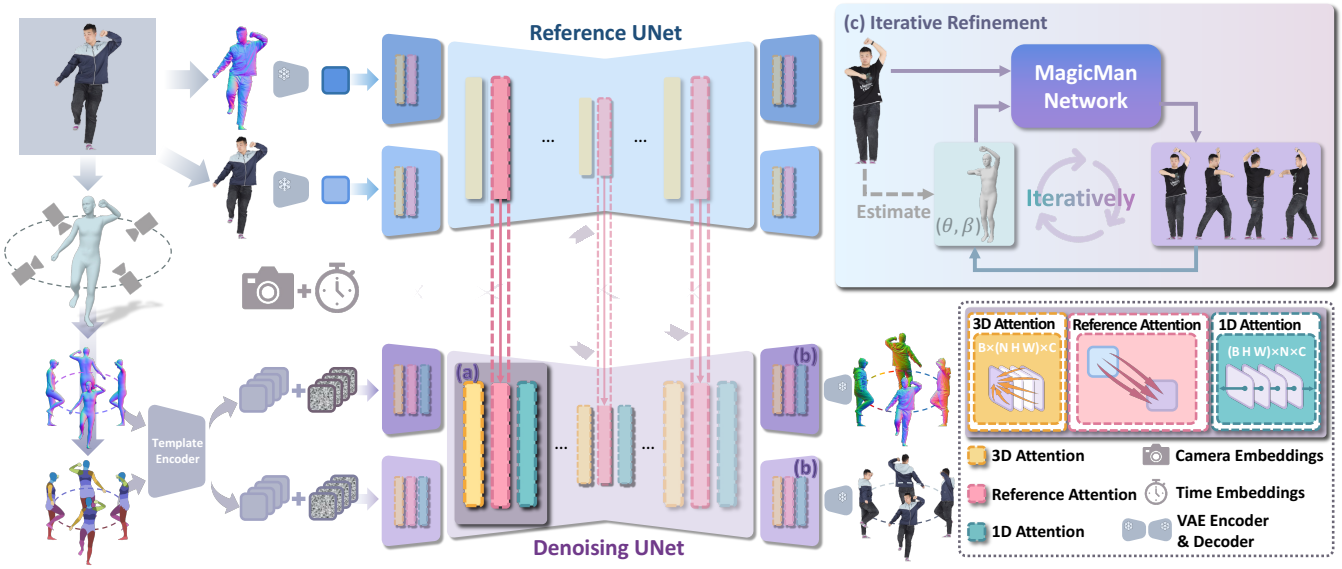

Figure 3: Given a single human image, our proposed MagicMan utilizes a pre-trained 2D diffusion model with a 3D human prior to generate novel view images for humans. First, the reference image is fed into the denoising UNet via a reference UNet, with the viewpoint condition incorporated through camera embeddings. The rendered normal and segmentation maps of the posed SMPL-X mesh that corresponds to the reference image are also provided as geometry guidance to facilitate 3D awareness and consistency. To obtain dense and consistent novel view images, we modify the attention module to a more efficient hybrid 1D-3D attention (a) to establish comprehensive connections between multi-views, and propose a geometry-aware dual branch (b) to also generate normal images complementary to RGB images via geometry cues. Last but not least, a novel iterative refinement strategy (c) is proposed in the inference stage to gradually update the initially estimated inaccurate SMPL-X pose and the synthesized novel view images, substantially reducing the ill-shaped issues arising from unreliable SMPL-X estimates.

work, we use diffusion models equipped with a parametric body model as prior to synthesize consistent human views.

Human Image Synthesis. Human image synthesis aims to synthesize novel views or poses given a source image as reference. This problem is explored by using generative adversarial networks (Ma et al. 2017) or optical-flow-based warping (He et al. 2024) to synthesize novel pose results. Liquid Warping GAN (2019) also demonstrates the human novel view results. Recently, diffusion models excel at modeling complex data distribution and also show promise in human image synthesis. Bhunia et al. introduces the first diffusion-based approach for pose-guided person synthesis. In addition, DreamPose (2023), Animate Anyone (2024), MagicAnimate (2024), and Champ (2024) generates animated videos from a source image using pre-trained SD. These methods could synthesize human images in novel views by giving the corresponding poses. However, without considering the 3D information between different views, it is hard to ensure view consistency. To address this problem, we utilize SMPL-X model to guide the synthesis process and explore more appropriate attention for multi-view consistency.

3 Method

Our proposed **MagicMan**, as illustrated in Fig. 3, takes a single human image as input and generates dense consistent multi-view images. To utilize human priors from abundant in-the-wild images, MagicMan adopts a pre-trained diffusion model (2022) as the backbone, and accepts one reference image along with SMPL-X pose and viewpoint as generation conditions (Sec. 3.1). We integrate an efficient hybrid-attention mechanism connecting different views, which contains 1D attention across all views and 3D attention across spatial locations and sparse selected views (Sec. 3.2). To achieve better geometric consistency, we propose a geometry-aware dual branch complementary to novel view image synthesis (Sec.3.3). Last but not least, we propose a novel iterative refinement strategy by updating both the accuracy of SMPL-X poses and the quality of generated multi-views across iterations, reducing the ill-shaped issues resulting from inaccurate pose estimation (Sec. 3.4). Further details on methodology are in our extended version.

3.1 Conditional Diffusion Model

Our backbone is a denoising UNet (2015) that inherits the structure and pre-trained weights from SD1.5 (2022). The vanilla SD UNet consists of downsampling, middle, and upsampling blocks, taking noise latents as input. Each block contains interleaved convolution layers, self-attention layers that perform feature aggregation spatially, and cross-attention layers that interact with CLIP embeddings. In our work, we incorporate the reference image and viewpoints into the network, and provide SMPL-X meshes as geometry guidance to promote 3D awareness and consistency.

Reference UNet. Inspired by recent advances in character animation (Hu 2024), we utilize a copy of the denoising

UNet, referred to as the reference UNet, to extract features from the reference image, which encompass both semantic and low-level reference information, and perform better than CLIP embeddings (Gu et al. 2024). Therefore, we replace the original CLIP cross attention with reference attention to interact with the features extracted by the reference UNet.

Let $x \in \mathbb{R}^{B\times N\times H\times W\times C}$ and $y \in \mathbb{R}^{B\times H\times W\times C}$ denote the corresponding feature maps from denosing UNet and reference UNet respectively, with batch size B, view count N, spatial size $H \times W$, and number of channels C. Since the reference image is shared for all views, feature maps y from the reference net are replicated N times. And then both x and y are reshaped to $\mathbb{R}^{(BN)\times(HW)\times C}$ for the following attention calculation. Mathematically,

$$Q^{\text{ref}} = W_Q^{\text{ref}} x, K^{\text{ref}} = W_K^{\text{ref}} (x \oplus y), V^{\text{ref}} = W_V^{\text{ref}} (x \oplus y), \quad (1)$$

where $\oplus$ denotes concatenation along the HW dimension.

Pose guidance and viewpoint control. A parametric SMPL-X mesh is estimated for the reference image and rendered according to the generated viewpoints to serve as pose and viewpoint conditions. We render normal and segmentation maps, which are encoded with a lightweight 4-layer convolution template encoder as proposed by Hu and added to the latent noise, to provide complementary geometric and semantic information (Zhu et al. 2024). In addition, view control is explicitly incorporated into the network through camera embeddings via an MLP and added to the denoising time embeddings.

3.2 Hybrid Multi-View Attention

To generate as many views as possible to capture comprehensive 3D information while maintaining multi-view consistency, the key problem lies in how to ensure thorough information exchange across a wide range of views in a memory-efficient manner. Therefore, we propose a novel hybrid attention mechanism that combines the strengths of two types of multi-view attention, i.e. efficiency of 1D attention and thoroughness of 3D attention in multi-view interaction.

1D multi-view attention. First, we insert an additional 1D attention layer behind the reference attention to establish connections between different views in a highly memory-efficient manner. It is calculated along view dimension only between identical pixel locations, allowing coherent generation of up to 20 views in a single forward pass. Specifically, the feature map is reshaped to $\mathbb{R}^{(BHW)\times N\times C}$ to calculate self-attention along N, and we employ relative positional encoding (2022) instead of the commonly used sinusoidal encoding to account for relative viewpoint differences.

3D multi-view attention. Relying solely on 1D attention leads to content drift issues (2023) between views after large viewpoint changes (Fig. 6) since 1D attention lacks interaction between pixels at different locations and cannot find the corresponding pixel from other views. Therefore, we further integrate 3D multi-view attention, facilitating more thorough information sharing across both spatial and view dimensions. Owing to the initial interactions established by 1D attention, 3D attention can be confined to a sparse subset of views without incurring excessive memory overhead.

Algorithm 1: Iterative Refinement

Input: Reference image $\mathcal{I}^{\text{ref}}$, target viewpoints $\mathcal{V}_{1:N}$
Parameter: Iterations K,
linearly increasing CFG scale $\{w_k\}_{k=1}^{K}$
Output: Novel view RGB images and normal maps $\hat{\mathcal{I}}_{1:N}$

Initialize SMPL-X params $(\theta, \beta, \psi) \leftarrow \text{Estimate}\left(\mathcal{I}^{\text{ref}}\right)$
for $k = 1, \ldots, K-1$ **do**
  `// Forward pass: Generate novel views with SMPL-X guidance and CFG scale` w_k
  $\hat{\mathcal{I}}_{1:N} = \mathcal{G}\left(\mathcal{I}^{\text{ref}}, \mathcal{M}\left(\theta, \beta, \psi\right), \mathcal{V}_{1:N}; w_k\right)$
  `// Backward pass: Refine SMPL-X supervised by generated novel views`
  Optimize (θ, β) with $\hat{\mathcal{I}}_{1:N}$ by minimizing Eq. 3
`// Generation with final refined SMPL-X`
return $\hat{\mathcal{I}}_{1:N} = \mathcal{G}\left(\mathcal{I}^{\text{ref}}, \mathcal{M}\left(\theta, \beta, \psi\right), \mathcal{V}_{1:N}; w_K\right)$

To be specific, we extend the origin self-attention of denoising UNet to 3D attention. For a reshaped feature map $x_i \in \mathbb{R}^{B\times(HW)\times C}$ of each view, 3D attention is efficiently conducted with a small number of feature maps $x_{j_1:j_M} \in \mathbb{R}^{B\times(MHW)\times C}$ in M views selected from the other views, calculated between all pixels across x_i and $x_{j_1:j_M}$ with:

$$Q = W_Q x_i, K = W_K\left(x_i \oplus x_{j_1:j_M}\right), V = W_V\left(x_i \oplus x_{j_1:j_M}\right). \quad (2)$$

Complete connections between different views are established by our hybrid 1D-3D attention without overwhelming computational cost, enabling the generation of dense and consistent multi-views. In practice, the sparse subset of views selected for 3D attention varies across different UNet blocks, making full use of different levels of information.

3.3 Geometry-Aware Dual Branch

Since geometry details are difficult to capture in the RGB domain, we introduce the dual branch to perform geometry-aware denoising, which generates the spatially aligned normal maps along with RGB images. To be specific, we replicate one of the UNet's input and output blocks to create two expert branches for RGB and normal images, both inheriting SD1.5 pre-trained weights, while the remaining blocks serve as shared components, as illustrated in Fig. 3(b). For the output of the normal branch, we use normal maps rendered from the ground truth (GT) human scans as training supervision. For the input to the reference normal branch, considering the unavailability of the GT during inference, we employ an off-the-shelf monocular normal estimator Marigold (Ke et al. 2024) to estimate the reference normal map from the input RGB image, unifying the training and inference setting. With these designs, the shared blocks facilitate feature fusion across domains. The normal branch enriches RGB by incorporating geometry awareness, improving structural stability and geometric consistency. Simultaneously, RGB enhances the accuracy and details of normal maps, significantly aiding subsequent 3D reconstruction.

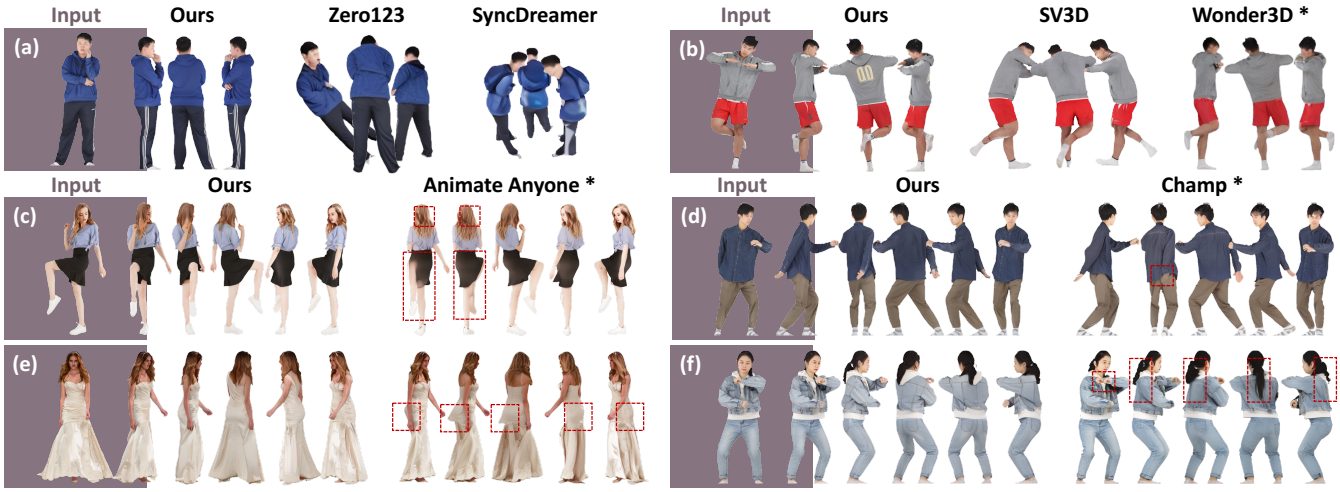

Figure 4: Qualitative results of human novel view synthesis. MagicMan generates the highest-quality dense novel view images with better consistency. Methods finetuned on THuman2.1 dataset are marked with *. Please zoom in for details.

3.4 Iterative Refinement

SMPL-X is a template human mesh $\mathcal{M}(\theta, \beta, \psi)$ parameterized by pose parameters θ, shape parameters β, and expression parameters ψ. The accuracy of the SMPL-X pose matters a lot since we employ its rendered normal and segmentation maps as geometry guidance to improve 3D consistency. However, monocular estimation could give inaccurate SMPL-X poses that conflict with the reference images, leading to the generation of distorted novel views and thus ill-shaped 3D reconstruction as illustrated in Fig. 7(b). On the other hand, generating novel view images without flawed SMPL-X guidance usually keeps its pose well matching the reference image, but inferior in terms of 3D consistency as shown in Fig. 7(a). Therefore, we propose that multi-view human images generated without flawed SMPL-X guidance can be used to optimize the accuracy of SMPL-X poses, while the refined SMPL-X meshes can then guide the generation of human muti-views with improved 3D consistency.

Therefore, we randomly drop SMPL-X guidance with a certain ratio during training, enabling guidance-free generation in line with classifier-free guidance (CFG) (2022). During inference, we introduce an iterative refinement process, as detailed in Algorithm 1. We set the initial CFG scale to 0, disabling SMPL-X guidance to preserve more accurate poses in the generated novel views matching the reference image. These images are then used to update the SMPL-X parameters. In subsequent iterations, we gradually increase the CFG scale to enhance the pose guidance of the refined SMPL-X estimation to further enhance 3D consistency.

Specifically, the iterative refinement process starts with estimating the initial SMPL-X parameters using PyMAF-X (2023). In each iteration, we use the current SMPL-X mesh as guidance, applying the corresponding CFG scale to generate human images. It's important to note that in the early iterations, the scale is kept small, resulting in weaker guidance that allows the generated images to better match the reference poses. Next, we use a differentiable renderer to produce SMPL-X mesh's normal maps $\mathcal{N}_{1:N}^{\text{SMPL-X}}$ and silhouettes $\mathcal{S}_{1:N}^{\text{SMPL-X}}$, as well as project 3D joints into 2D keypoints $\mathcal{J}_{1:N}^{\text{SMPL-X}}$ according to the camera poses. SMPL-X parameters are then optimized under the supervision of all the generated novel view images with the generated normal maps $\hat{\mathcal{N}}_{1:N}$, silhouettes $\hat{\mathcal{S}}_{1:N}$, and detected 2D joint keypoints $\hat{\mathcal{J}}_{1:N}$ from the generated novel views. The SMPL-X optimization is performed by minimizing the following loss:

$$\begin{aligned} \mathcal{L}_{\text{refine}} &= \lambda_{\text{normal}}\mathcal{L}_{\text{normal}} + \lambda_{\text{silhouette}}\mathcal{L}_{\text{silhouette}} + \lambda_{\text{joint}}\mathcal{L}_{\text{joint}}, \\ \mathcal{L}_{\text{normal}} = \left|\mathcal{N}_{1:N}^{\text{SMPL-X}} - \hat{\mathcal{N}}_{1:N}\right|, \mathcal{L}_{\text{silhouette}} &= \left|\mathcal{S}_{1:N}^{\text{SMPL-X}} - \hat{\mathcal{S}}_{1:N}\right|, \\ \mathcal{L}_{\text{joint}} &= \left|\mathcal{J}_{1:N}^{\text{SMPL-X}} - \hat{\mathcal{J}}_{1:N}\right|. \end{aligned} \tag{3}$$

After the optimization, SMPL-X parameters are refined to be more accurate and aligned with the reference, and will be fed back into the generation process with an increased CFG scale in the next iteration. In summary, during each iterative process, SMPL-X undergoes refinement supervised by all generated multi-views, and the multi-view generation is enhanced with the improved SMPL-X as guidance.

4 Experiments

Training data. We train MagicMan on 2347 human scans from THuman2.1 dataset (2021b). RGB and normal images are rendered using a weak perspective camera on 20 fixed viewpoints looking at the scan with evenly distributed azimuths from 0° to 360°, at 512×512 resolution.

Evaluation data. We test on 95 scans from THuman2.1 dataset and 30 scans from CustomHumans dataset (2023) and also evaluate on in-the-wild images, comprising 100 from SHHQ dataset (Fu et al. 2022) and 120 from the Internet featuring varied poses, outfits, and styles.

Metrics. Evaluation is conducted on two tasks: 1) Novel view synthesis. We use PSNR, SSIM, LPIPS, and CLIP scores to compare generated views w.r.t. the ground-truth images of corresponding views. For in-the-wild data, we calculate LPIPS of the generated reference view and CLIP

Method	THuman2.1 PSNR ↑	SSIM ↑	LPIPS ↓	CLIP ↑	CustomHumans PSNR ↑	SSIM ↑	LPIPS ↓	CLIP ↑	in-the-wild LPIPS ↓	CLIP ↑
Generative Novel View Synthesis Methods										
Zero123	15.5/12.3	0.829/0.798	0.215/0.247	0.792/0.773	13.9/10.3	0.811/0.766	0.211/0.250	0.818/0.800	0.082	0.722
SyncDreamer	13.2/ –	0.841/ –	0.209/ –	0.749/ –	11.5/ –	0.824/ –	0.211/ –	0.767/ –	0.175	0.630
SV3D	19.8/16.5	0.896/0.868	0.115/0.140	0.888/0.877	18.3/14.3	0.892/0.857	0.111/0.141	0.912/0.899	**0.030**	0.818
Wonder3D*	21.2/ –	0.906/ –	0.110/ –	0.882/ –	18.3/ –	0.889/ –	0.122/ –	0.863/ –	0.064	0.757
Character Animation Methods										
Animate Anyone*	23.6/22.1	0.923/0.910	0.070/0.078	0.920/0.917	22.0/20.4	0.919/0.905	0.060/0.068	0.931/0.929	0.061	0.850
Champ*	24.9/23.3	0.930/0.918	0.063/0.071	0.927/0.924	23.2/21.4	0.931/0.918	0.055/0.063	0.938/0.934	0.053	0.852
Ours	**26.0/24.9**	**0.946/0.929**	**0.049/0.054**	**0.947/0.938**	**24.7/22.9**	**0.950/0.937**	**0.044/0.052**	**0.947/0.947**	0.040	**0.871**

Table 1: Quantitative evaluation for novel view synthesis on "4 views / 20 views". Methods finetuned on THuman2.1 dataset are marked with *. Bold and underline indicate the best and the second, respectively.

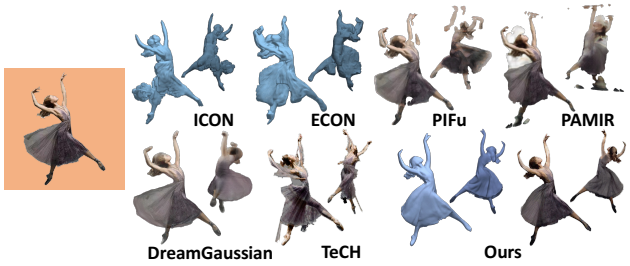

Figure 5: Reconstructed 3D human meshes. MagicMan produces the best geometry and textures in the case with challenging poses and loose outfits. Please zoom in for details.

Method	THuman2.1 Chamfer ↓	P2S ↓	NE ↓	CustomHumans Chamfer ↓	P2S ↓	NE ↓
PIFu	5.62	5.11	0.150	6.43	5.76	0.159
PAMIR	4.30	4.27	0.132	5.00	4.89	0.135
ICON	5.05	5.02	0.139	5.46	5.48	0.134
ECON	5.45	5.26	0.148	5.72	5.61	0.138
Ours	**2.35**	**2.44**	**0.093**	**2.34**	**2.43**	**0.095**

Table 2: Quantitative evaluation for 3D human reconstruction. Bold and underline indicate the best and the second.

scores of generated novel views w.r.t. the input image. 2) 3D human reconstruction. Following Xiu et al., we calculate Chamfer and P2S distance, and L2 normal errors (NE).

4.1 Novel View Synthesis

To evaluate novel view synthesis results, we compare MagicMan with generative object novel view synthesis methods, i.e. Zero123 (2023a), SyncDreamer (2023b), Wonder3D (2024), and SV3D (2024), and character animation methods with body priors, i.e. Animate Anyone (2024) and Champ (2024). We selected the best-performing Wonder3D as a representative of general novel view synthesis methods and fine-tuned it on THuman2.1. Additional comparison results finetuned on THuman2.1 can be found in our extended version. Examples of human novel view images and normal maps generated by MagicMan are shown in Fig.1, demonstrating that MagicMan can generate high-quality and 3D consistent human novel views across diverse poses, outfits, and styles. Fig. 4 presents qualitative comparisons between MagicMan and baseline methods. Zero123, SyncDreamer, and SV3D typically generate distorted human images. Wonder3D produces only six views at half of our resolution, leading to texture detail loss. Lack of a body prior also results in geometric errors. Animation methods yield unrealistic body structures for a lack of geometric awareness, sometimes encountering ambiguities between front and back views as shown in Fig. 4(c). Besides, they exhibit noticeable inconsistency between views with a large viewpoint change as illustrated in Fig. 4(e) and 4(f). In contrast, our method achieves stable structures, consistent geometry, and detailed textures while generating dense novel views for humans.

Quantitative comparisons are reported in Tab 1, showing that MagicMan outperforms baseline approaches in both pixel-level and semantic metrics, except for slightly higher LPIPS on in-the-wild data for reference view reconstruction, likely due to SV3D's better frontal details at a higher resolution. However, CLIP scores of novel views indicate that our method significantly excels in novel view synthesis.

4.2 3D Human Reconstruction

Fig. 5 displays our reconstructed human mesh, compared with those produced by baseline methods including feed-forward approaches PIFu (2019), PaMIR (2021), ICON (2022), ECON (2023), and SDS-based DreamGaussian (2023), TeCH (2024). Both feed-forward and SDS-based methods fail to produce reasonable geometry and detailed consistent textures for the challenging pose and outfit, while our 3D-aware diffusion model with refined body prior generates dense and consistent multi-views, which support reliable reconstruction with enhanced geometry and textures. Quantitative comparisons with PIFu, PAMIR, ICON, and ECON are presented in Tab. 2, illustrating that MagicMan outperforms previous approaches on all metrics by a significant margin. Note that we include our iterative refinement process and the SMPL-X optimization operations of

Method	THuman2.1 PSNR ↑	SSIM ↑	LPIPS ↓	CLIP ↑	in-the-wild LPIPS ↓	CLIP ↑
Baseline	22.84	0.915	0.070	0.923	**0.037**	0.843
+ 1D attn.	<u>23.94</u>	<u>0.924</u>	0.063	0.935	0.041	0.853
+ 3D attn.	23.77	0.923	<u>0.063</u>	<u>0.936</u>	0.042	<u>0.854</u>
w/o norm.	23.73	0.921	0.064	0.925	0.042	0.833
Ours	**24.85**	**0.929**	**0.054**	**0.938**	<u>0.040</u>	**0.871**

Table 3: Quantitative ablations on hybrid attention and dual branch. Bold and underline indicate the best and the second.

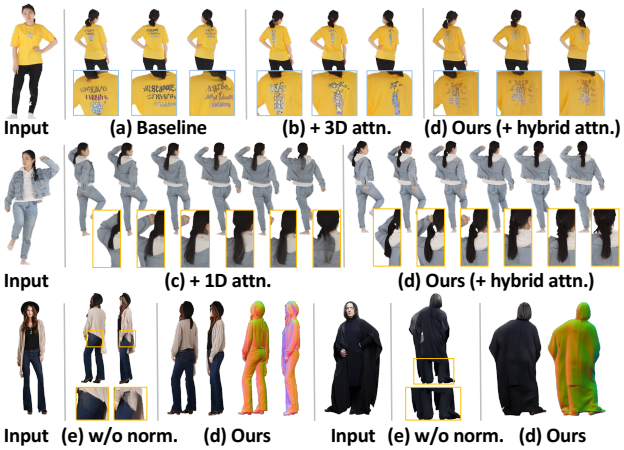

Figure 6: Ablations on hybrid attention and dual branch. Our full model presents the best multi-view consistency.

ICON, ECON, and PAMIR are retained for fair comparison.

4.3 Ablations and Discussions

Hybrid attention. With hybrid attention, MagicMan can generate up to 20 consistent multi-views in training, with an inference time of ∼40s on 1 A100 GPU, while traditional 3D attention across all views yields only 6 views under the same memory constraint and takes ∼60s for inference. Fig. 6 illustrates the effectiveness of different components of hybrid attention: (a) The baseline model without multi-view attention generates inconsistent views. (b) 3D attention across selected views still produces flickering cloth patterns. (c) 1D attention alone presents content drift, e.g., hair length that gradually changes with increasing viewpoint changes, indicating that information exchange via 1D attention alone improves similarity but is insufficient for comprehensive consistency. (d) Our full model with hybrid attention demonstrates the best consistency when generating dense multi-views, also confirmed by the quantitative results in Tab. 3.

Geometry-aware dual branch. In Fig. 6(e) and row 4 of Tab. 3, removing the normal branch leads to a degradation in multi-view consistency, especially in complex geometric deformations, e.g., fabric layers and folds. Our full model with normal prediction enhances geometry awareness, yielding improved structures and consistency.

Iterative refinement. We conduct an ablation study to validate the effectiveness of the iterative refinement. As shown in Fig. 7(a), generation without SMPL-X guidance produces seemingly satisfactory novel views with accurate poses, which, however, exhibits severe artifacts in reconstruction due to inconsistent poses between views without 3D prior. Directly utilizing the estimated inaccurate SMPL-X mesh as pose guidance in Fig. 7(b) leads to distorted novel view images and the ill-shaped reconstructed mesh (e.g., the missing and disjointed hand and foot) due to conflicts between incorrect SMPL-X and reference image. Impressive results can be achieved with the accurate ground-truth SMPL-X as presented in Fig. 7(c), which, however, is unavailable in practice. Our iterative refinement process progressively improves the novel view results for reconstruction with increasingly accurate SMPL-X guidance through successive iterations, as demonstrated by the green bubbles in Fig. 7(e). The finally refined multi-view images, encompassing both accurate poses and 3D consistency, yield comparable results to those produced with the ground-truth SMPL-X. The refined SMPL-X mesh with more accurate poses and reduced depth ambiguities, as a byproduct of our refinement process, suggests that the abundant priors in pre-trained image diffusion models can potentially aid in human body estimation.

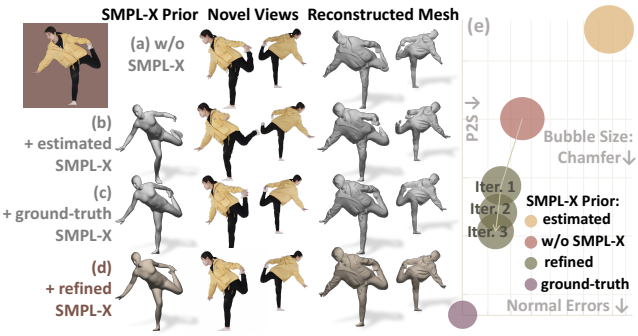

Figure 7: Qualitative (left) and quantitative (right) results of ablation studies on the iterative refinement.

5 Conclusion

In this paper, we introduce **MagicMan**, a method for generating human novel views from a single reference image by leveraging an image diffusion model as the 2D generative prior and the SMPL-X model as the 3D body prior. Building on this, our proposed efficient hybrid multi-view attention ensures the generation of denser multi-view images while maintaining high 3D consistency, which is further enhanced by the geometry-aware dual branch via geometry cues. Moreover, our novel iterative refinement process optimizes the initially estimated SMPL-X poses over successive iterations, guiding novel view generation with improved consistency and alleviating ill-shaped issues caused by inaccurate SMPL-X estimates. Extensive experimental results demonstrate that our method can generate dense, high-quality, and consistent human novel view images, which are also ideally suited for subsequent 3D human reconstruction tasks.

Ethical Statement

Human-centric generation raises concerns like privacy violations and intellectual property rights infringement, requiring a collective effort to develop ethical guidelines and legal standards. However, we still believe that the proper use of this technique will enhance the research of artificial intelligence and digital entertainment. In this work, considering the sensitivity of personal information, all processed data, models, and results will be strictly used for academic purposes and will not be authorized for commercial use.

Acknowledgments

This work is supported by National Natural Science Foundation of China (62076144), Shenzhen Key Laboratory of next generation interactive media innovative technology (ZDSYS20210623092001004) and Shenzhen Science and Technology Program (WDZC20220816140515001, JCYJ20220818101014030).

References

Balan, A. O.; Sigal, L.; Black, M. J.; Davis, J. E.; and Haussecker, H. W. 2007. Detailed human shape and pose from images. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, 1–8. IEEE.

Bhunia, A. K.; Khan, S.; Cholakkal, H.; Anwer, R. M.; Laaksonen, J.; Shah, M.; and Khan, F. S. 2023. Person image synthesis via denoising diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5968–5976.

Chan, E. R.; Nagano, K.; Chan, M. A.; Bergman, A. W.; Park, J. J.; Levy, A.; Aittala, M.; De Mello, S.; Karras, T.; and Wetzstein, G. 2023. Generative novel view synthesis with 3d-aware diffusion models. *arXiv preprint arXiv:2304.02602*.

Fu, J.; Li, S.; Jiang, Y.; Lin, K.-Y.; Qian, C.; Loy, C. C.; Wu, W.; and Liu, Z. 2022. Stylegan-human: A data-centric odyssey of human generation. In *European Conference on Computer Vision*, 1–19. Springer.

Gao, X.; Li, X.; Zhang, C.; Zhang, Q.; Cao, Y.; Shan, Y.; and Quan, L. 2024. Contex-human: Free-view rendering of human from a single image with texture-consistent synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10084–10094.

Gu, Y.; Xu, H.; Xie, Y.; Song, G.; Shi, Y.; Chang, D.; Yang, J.; and Luo, L. 2024. DiffPortrait3D: Controllable Diffusion for Zero-Shot Portrait View Synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10456–10465.

He, X.; Huang, Q.; Zhang, Z.; Lin, Z.; Wu, Z.; Yang, S.; Li, M.; Chen, Z.; Xu, S.; and Wu, X. 2024. Co-Speech Gesture Video Generation via Motion-Decoupled Diffusion Model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2263–2273.

Ho, H.-I.; Xue, L.; Song, J.; and Hilliges, O. 2023. Learning locally editable virtual humans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21024–21035.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Ho, J.; and Salimans, T. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*.

Hu, L. 2024. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8153–8163.

Huang, Y.; Yi, H.; Xiu, Y.; Liao, T.; Tang, J.; Cai, D.; and Thies, J. 2024. Tech: Text-guided reconstruction of lifelike clothed humans. In *2024 International Conference on 3D Vision (3DV)*, 1531–1542. IEEE.

Hwang, C.; Cui, W.; Xiong, Y.; Yang, Z.; Liu, Z.; Hu, H.; Wang, Z.; Salas, R.; Jose, J.; Ram, P.; Chau, J.; Cheng, P.; Yang, F.; Yang, M.; and Xiong, Y. 2022. Tutel: Adaptive Mixture-of-Experts at Scale. arXiv:2206.03382.

Karras, J.; Holynski, A.; Wang, T.-C.; and Kemelmacher-Shlizerman, I. 2023. Dreampose: Fashion image-to-video synthesis via stable diffusion. *arXiv preprint arXiv:2304.06025*.

Ke, B.; Obukhov, A.; Huang, S.; Metzger, N.; Daudt, R. C.; and Schindler, K. 2024. Repurposing Diffusion-Based Image Generators for Monocular Depth Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Kulhánek, J.; Derner, E.; Sattler, T.; and Babuška, R. 2022. Viewformer: Nerf-free neural rendering from few images using transformers. In *European Conference on Computer Vision*, 198–216. Springer.

Li, Z.; Wang, Q.; Snavely, N.; and Kanazawa, A. 2022. Infinitenature-zero: Learning perpetual view generation of natural scenes from single images. In *European Conference on Computer Vision*, 515–534. Springer.

Liu, R.; Wu, R.; Van Hoorick, B.; Tokmakov, P.; Zakharov, S.; and Vondrick, C. 2023a. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF international conference on computer vision*, 9298–9309.

Liu, W.; Piao, Z.; Min, J.; Luo, W.; Ma, L.; and Gao, S. 2019. Liquid warping gan: A unified framework for human motion imitation, appearance transfer and novel view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5904–5913.

Liu, X.; Zhan, X.; Tang, J.; Shan, Y.; Zeng, G.; Lin, D.; Liu, X.; and Liu, Z. 2024. Humangaussian: Text-driven 3d human generation with gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6646–6657.

Liu, Y.; Lin, C.; Zeng, Z.; Long, X.; Liu, L.; Komura, T.; and Wang, W. 2023b. Syncdreamer: Generating multiview-consistent images from a single-view image. *arXiv preprint arXiv:2309.03453*.

Long, X.; Guo, Y.-C.; Lin, C.; Liu, Y.; Dou, Z.; Liu, L.; Ma, Y.; Zhang, S.-H.; Habermann, M.; Theobalt, C.; et al. 2024. Wonder3d: Single image to 3d using cross-domain diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9970–9980.

Loper, M.; Mahmood, N.; Romero, J.; Pons-Moll, G.; and Black, M. J. 2015. SMPL: A Skinned Multi-Person Linear Model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6): 248:1–248:16.

Ma, L.; Jia, X.; Sun, Q.; Schiele, B.; Tuytelaars, T.; and Van Gool, L. 2017. Pose guided person image generation. *Advances in neural information processing systems*, 30.

Pavlakos, G.; Choutas, V.; Ghorbani, N.; Bolkart, T.; Osman, A. A.; Tzionas, D.; and Black, M. J. 2019. 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*, 10975–10985.

Poole, B.; Jain, A.; Barron, J. T.; and Mildenhall, B. 2022. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*.

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2): 3.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, 234–241. Springer.

Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E. L.; Ghasemipour, K.; Gontijo Lopes, R.; Karagol Ayan, B.; Salimans, T.; et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35: 36479–36494.

Saito, S.; Huang, Z.; Natsume, R.; Morishima, S.; Kanazawa, A.; and Li, H. 2019. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2304–2314.

Shen, F.; Ye, H.; Zhang, J.; Wang, C.; Han, X.; and Wei, Y. 2024. Advancing Pose-Guided Image Synthesis with Progressive Conditional Diffusion Models. In *The Twelfth International Conference on Learning Representations*.

Shi, Y.; Wang, P.; Ye, J.; Long, M.; Li, K.; and Yang, X. 2023. Mvdream: Multi-view diffusion for 3d generation. *arXiv preprint arXiv:2308.16512*.

Sohl-Dickstein, J.; Weiss, E.; Maheswaranathan, N.; and Ganguli, S. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, 2256–2265. PMLR.

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

Tseng, H.-Y.; Li, Q.; Kim, C.; Alsisan, S.; Huang, J.-B.; and Kopf, J. 2023. Consistent View Synthesis with Pose-Guided Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16773–16783.

Vlasic, D.; Peers, P.; Baran, I.; Debevec, P.; Popović, J.; Rusinkiewicz, S.; and Matusik, W. 2009. Dynamic shape capture using multi-view photometric stereo. In *ACM SIGGRAPH Asia 2009 papers*, 1–11.

Voleti, V.; Yao, C.-H.; Boss, M.; Letts, A.; Pankratz, D.; Tochilkin, D.; Laforte, C.; Rombach, R.; and Jampani, V. 2024. Sv3d: Novel multi-view synthesis and 3d generation from a single image using latent video diffusion. *arXiv preprint arXiv:2403.12008*.

Xian, X.; He, X.; Niu, Z.; Zhang, J.; Xie, W.; Song, S.; Yu, Z.; and Shen, L. 2024. CA-Edit: Causality-Aware Condition Adapter for High-Fidelity Local Facial Attribute Editing. *arXiv preprint arXiv:2412.13565*.

Xiu, Y.; Yang, J.; Cao, X.; Tzionas, D.; and Black, M. J. 2023. Econ: Explicit clothed humans optimized via normal integration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 512–523.

Xiu, Y.; Yang, J.; Tzionas, D.; and Black, M. J. 2022. Icon: Implicit clothed humans obtained from normals. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 13286–13296. IEEE.

Xu, Z.; Zhang, J.; Liew, J. H.; Yan, H.; Liu, J.-W.; Zhang, C.; Feng, J.; and Shou, M. Z. 2024. Magicanimate: Temporally consistent human image animation using diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1481–1490.

Yu, A.; Ye, V.; Tancik, M.; and Kanazawa, A. 2021a. pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4578–4587.

Yu, T.; Zheng, Z.; Guo, K.; Liu, P.; Dai, Q.; and Liu, Y. 2021b. Function4D: Real-time Human Volumetric Capture from Very Sparse Consumer RGBD Sensors. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR2021)*.

Zhang, H.; Tian, Y.; Zhang, Y.; Li, M.; An, L.; Sun, Z.; and Liu, Y. 2023. PyMAF-X: Towards Well-aligned Full-body Model Regression from Monocular Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zhang, J.; Li, X.; Zhang, Q.; Cao, Y.; Shan, Y.; and Liao, J. 2024. Humanref: Single image to 3d human generation via reference-guided diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1844–1854.

Zheng, Z.; Yu, T.; Liu, Y.; and Dai, Q. 2021. Pamir: Parametric model-conditioned implicit representation for image-based human reconstruction. *IEEE transactions on pattern analysis and machine intelligence*, 44(6): 3170–3184.

Zhu, S.; Chen, J. L.; Dai, Z.; Xu, Y.; Cao, X.; Yao, Y.; Zhu, H.; and Zhu, S. 2024. Champ: Controllable and consistent human image animation with 3d parametric guidance. *arXiv preprint arXiv:2403.14781*.