This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

FISHuman: Fine-grained Single-image 3D Human Reconstruction via Multi-view 4D Remeshing

Hanxi Liu[1], Yifang Men[2], Zhouhui Lian[1]*
[1]Wangxuan Institute of Computer Technology, Peking University, China
[2]Institute for Intelligent Computing, Alibaba Tongyi Lab

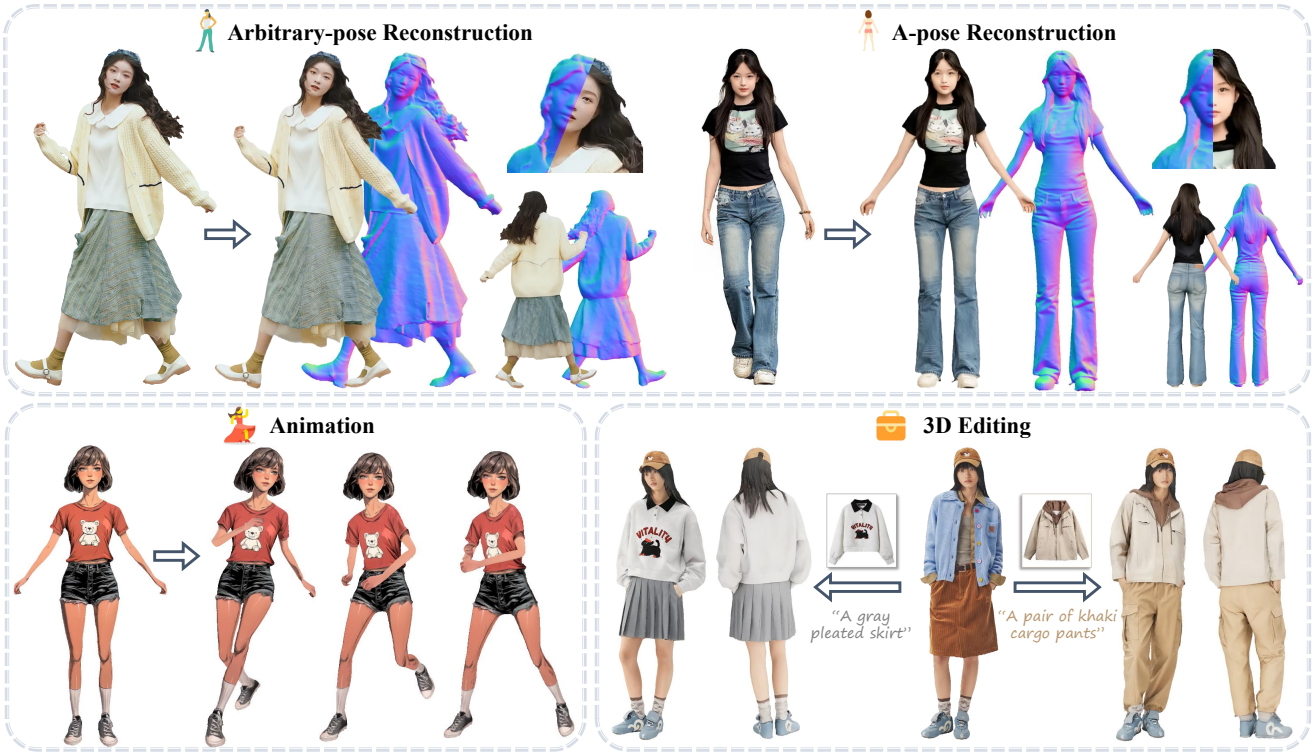


Figure 1. We present FISHuman, a method for generating fine-grained 3D humans from single in-the-wild images. Top: FISHuman enables high-fidelity and ultra-photorealistic textured human reconstruction under arbitrary and standard poses. Bottom: The generated avatars showcase practical applications in direct animation and 3D editing. Please zoom in for better inspection.

Abstract

Single-image 3D human reconstruction holds significant promise due to its convenience and high demand in various applications. Previous methods have garnered tremendous progress by employing 2D multi-view diffusion models to generate auxiliary views as reconstruction priors, but they struggle with 3D inconsistencies and limited generalization capabilities. In this paper, we present FISHuman, which aims to generate fine-grained, high-fidelity, and content-wise diverse 3D humans from a single-view input, providing production-ready 3D assets. We propose an elaborately designed workflow that reconstructs dynamic 3D meshes from multi-view inconsistent guidance. Specifically, we adapt a dual-stream transformer-based video diffusion model to generate cross-modally aligned multi-view RGB and normal sequences. We find that naively employing static 3D reconstruction can lead to geometric distortions and texture blurriness, due to the lack of 3D awareness within the generated frames. To address this, we introduce a novel 4D remeshing module that explicitly disentangles the learning of the globally shared canonical mesh and transient variations by tracking per-vertex deformations under different viewpoints. The topological consistency of the deformed meshes inherently enables the opti-

*Corresponding author. E-mail: lianzhouhui@pku.edu.cn

mization of a unified UV representation that effectively integrates appearance attributes across frames. Both qualitative and quantitative experimental results demonstrate the superiority of our method over prior works in terms of appearance realism, geometric fineness, and generalization diversity. We also showcase the applicability of our reconstructed avatars for downstream applications including animation and 3D editing.

1. Introduction

Realistic 3D avatar modeling plays a pivotal role in applications including film production, gaming, and VR/AR industries. Reconstructing 3D digital humans from a single in-the-wild image offers a promising path to scale up the generation of human assets in a customer-friendly manner. However, lifting a monocular image to 3D is inherently ill-posed due to severe self-occlusions and the substantial variability of the human body in appearance, pose, and clothing.

To tackle the challenge of generating high-quality 3D human assets from a single image, early methods [47, 48, 65, 66, 88] reconstruct 3D humans by leveraging pixel-aligned implicit features or explicit body priors (e.g., SMPL [37]). Despite their efficiency, these approaches suffer from severe artifacts in unobserved regions due to single-view constraints. Recent advances [12, 29, 86] employ 2D multi-view diffusion models to synthesize a fixed set of auxiliary views for 3D reconstruction. However, constrained by the model architecture and memory overhead, these methods are limited to sparse view synthesis, leaving self-occluded regions inadequately covered. Moreover, the diffusion-generated multi-view renderings lack explicit 3D constraints, resulting in distorted reconstructions and blurred textures that are incompatible with downstream applications, which require fine-grained 3D assets. While some methods [11, 42, 71] adopt diffusion or transformer-based 3D generative models [15, 54] to enforce cross-view consistency, their generation quality remains limited by the resolution constraints and the distribution of training data, hindering robust generalization across diverse character styles.

To address the aforementioned challenges, we propose ***FISHuman***, a novel pipeline that reconstructs fine-grained 3D humans from a single-view input, and produces 3D assets ready for extraction and downstream applications, as illustrated in Fig. 1. Our key insight is to bridge the gap between the single image and explicit 3D representation with a 3D-aware dual-stream transformer-based video generator. This model generates an arbitrary number of coherent RGB and normal images with spatial awareness and multimodal consistency, by inducing 3D-aware specialization within each modality and enforcing structural alignment via cross-modal attention. This provides dense geometric and textural priors for subsequent reconstruction, enabling the generation of intrinsically consistent 3D entities.

Although our 3D-aware video generator significantly improves 3D consistency across multi-view renderings, it still struggles to generate physically precise results without rule-based modeling, especially for drifting cloth surfaces and articulated body motions. Naive 3D reconstruction from the generated frames forces the model to average the conflicting supervisory signals, leading to distorted geometry and blurred textures. The key to coherent and fine-grained reconstruction under inconsistent video guidance is to convert the pixel-level drifting into dynamic vertex deformations, and establish accurate correspondence of the underlying geometry across all perspectives. To this end, we propose a novel 4D remeshing module, which optimizes the globally shared canonical geometry by tracking vertices across multi-view inconsistent normal maps, and updates additional per-vertex deformations to carve frame-dependent local details. The 4D remeshing module not only enables the learning of dynamic geometric changes that fit different viewpoints, but also establishes consistent correspondence between inter-frame mesh topologies. This further allows for the optimization of a unified UV representation to effectively integrate valid appearance features from the multi-view RGB sequence. In summary, our contributions are threefold:

- We present FISHuman, a novel pipeline equipped with a 3D-aware video generator and a dynamic human carver to reconstruct fine-grained and ultra-photorealistic 3D humans from single images, even under various complicated real-world scenarios. The generated avatars support animation, editing, and export as high-quality digital assets.
- We leverage a 3D-aware dual-stream video diffusion transformer to jointly synthesize auxiliary multi-view RGB and normal images as the reconstruction guidance. Through unified cross-modal diffusion with 3D-aware specialization, our model produces dense novel views that are coherent both across 3D space and dual modalities.
- We propose a novel 4D remeshing module that tracks view-dependent vertex deformations upon global canonical geometry, in cooperation with a unified UV texture learning stage. This design effectively mitigates 3D inconsistencies brought by the generated frames and significantly improves reconstruction quality.

2. Related Works

2.1. Diffusion-based Human Image Animation

With the rise of video diffusion models [2, 13, 14, 35, 58, 76], a series of works have focused on controllable human image animation [17, 18, 39, 49, 60, 70, 85, 89]. AnimateAnyone [17] adopts a UNet-based ReferenceNet that encodes texture features from the reference image. StableAn-

imator [57] employs an ID Adapter that injects enhanced face embeddings into the UNet for human identity preservation. Champ [89] incorporates the SMPL model [37] for a unified representation that encompasses both body shapes and human poses. MIMO [39] leverages decomposed spatial components for complex motion expression and 3D-aware human-scene interaction. Although these methods enable controllable human motion synthesis, their generated results often exhibit noticeable appearance variations and unrealistic dynamic deformations due to the absence of 3D physical constraints, preventing the generation of intrinsically consistent 3D entities.

2.2. Versatile 3D Asset Generation

The field of 3D object generation has witnessed vigorous development in recent years. Numerous approaches have investigated enhancements in geometric fineness [5, 30, 62, 63, 77], texture realism [3, 46, 64, 81, 90], generation efficiency [28, 67, 69], and part-level decomposition [32, 73]. A prevalent paradigm [34, 36, 50, 51, 54, 68] involves first generating multi-view consistent images via 2D diffusion models, followed by per-case optimization or feed-forward 3D reconstruction. With the advent of the Objaverse dataset [6] and the pursuit of higher efficiency and fidelity, more recent methods have extended 2D latent diffusion models to native 3D object generation, incorporating representations such as vector-sets [5, 30, 82, 83], voxels [45, 63, 77], or auto-regressive mesh generation [4, 52, 87]. Despite their impressive results in general content creation, these methods exhibit moderate fidelity in reference image restoration. This limitation becomes particularly evident in human generation, where they struggle to precisely capture individual identity and the structural layout of human figures.

2.3. Single-view Clothed Human Reconstruction

Focusing on 3D clothed human reconstruction from single images, PIFu [47] and its follow-ups [1, 10, 21, 48, 65] exploit 3D implicit functions augmented by normal maps or statistical body templates [37]. ECON [66] adopts a more explicit approach by utilizing normal integration and shape completion to handle flexible topology and challenging poses. However, these methods heavily rely on large-scale 3D human datasets and suffer from severe artifacts in regions invisible from the input view. With the rise of 2D diffusion models, recent methods either adopt score distillation sampling [19] or leverage multi-view image diffusion [12, 26, 29, 86] to harness the prior knowledge of diffusion models. SIFU [86] leverages a consistent multi-view image editing model for texture refinement of the mesh extracted from implicit fields. PSHuman [29] employs diffusion-based multi-view generation with SMPL-X [41] guidance to incorporate human anatomy priors. Nevertheless, 3D inconsistency still remains, leading to geometric defects and blurred textures during reconstruction, especially in cases with rare poses and severe self-occlusions. Some methods [11, 42, 71] propose using generated multi-views as input to a diffusion or transformer-based 3D generator for natively 3D consistent modeling, but they struggle with low-resolution outputs and limited generalization abilities. To improve appearance realism, more recent advances [7, 20, 43, 44, 59, 91] represent 3D humans as a set of Gaussian kernels [23]. Despite their photorealistic rendering with high efficiency, these approaches fail to support high-quality explicit 3D asset export due to the unstructured nature of 3D Gaussian Splatting, which is incompatible with existing rendering engines.

3. Method Description

Given a single-view image I of a person, we aim to create a fine-grained 3D human represented as a textured mesh. Fig. 2 illustrates an overview of our proposed framework. We first synthesize highly aligned novel-view RGB and normal maps using a 3D-aware dual-stream video diffusion transformer, which we describe in Sec. 3.1. Then, to overcome the inherent 3D inconsistencies brought by the generated sequences, we develop the dynamic 3D human carving method, which consists of a 4D remeshing module and a unified UV representation learning stage (see Sec. 3.2), facilitating fine-grained geometry and appearance optimization under inconsistent multi-view guidance.

3.1. 3D-aware Dual-stream Video Generation

In this section, our goal is to generate RGB and normal sequences with 3D awareness and cross-modal coherence, given single input images. Our method employs a dual-stream video Diffusion Transformer (DiT) finetuned on multi-view renderings of synthetic 3D human data, where camera viewpoint variations are equivalent to rigid rotations of the human body. We define the target distribution as:

$$p(\mathcal{F}^{rgb}_{1:T}, \mathcal{F}^{norm}_{1:T} | \mathbf{c}^{rgb}_{ref}, \mathbf{c}^{norm}_{ref}, \mathbf{c}^{pose}), \tag{1}$$

where $\mathcal{F}^{rgb}_{1:T}$ and $\mathcal{F}^{norm}_{1:T}$ denote the output RGB and normal frames with length T, respectively, conditioned on the reference image $\mathbf{c}^{rgb}_{ref}$, its estimated normal map $\mathbf{c}^{norm}_{ref}$, and a 2D pose guidance $\mathbf{c}^{pose}$ extracted by DWPose [74]. The reference image $\mathbf{c}^{rgb}_{ref}$ can be optionally transformed to a canonical A-pose through pose interpolation for reduced self-occlusion and easier animation, using a pre-trained human repose model [60]. We deliberately avoid using the commonly adopted SMPL-X condition due to inaccuracies in monocular 3D pose estimation, and leverage the model's embedded 3D awareness and prior knowledge for inferring physically plausible novel-view poses.

Dual-stream diffusion transformer. Our method implements a dual-stream cross-modal transformer based on

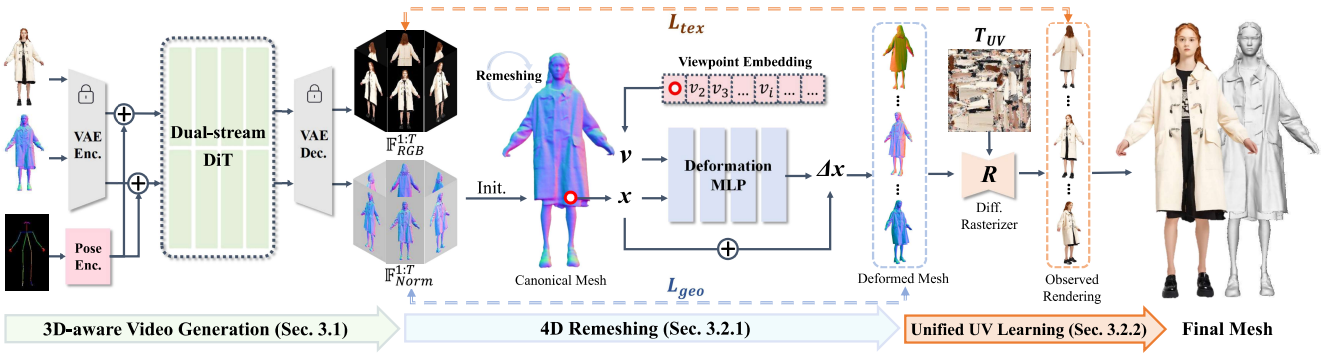


Figure 2. **Method Overview**. Given a single-view human image, FISHuman reconstructs the corresponding fine-grained 3D avatar by generating cross-modally aligned RGB and normal frames using the proposed 3D-aware dual-stream video diffusion transformer. With the generated sequences, we perform 4D Remeshing that explicitly decouples the optimization of the 3D canonical mesh and view-dependent per-vertex deformation field, effectively mitigating 3D inconsistencies of the multi-view guidance. A unified UV representation is updated on the deformed meshes to integrate valid appearance across all perspectives.

Wan2.1 image-to-video generation model [58]. The reference RGB image $\mathbf{c}_{ref}^{rgb}$ and normal map $\mathbf{c}_{ref}^{norm}$ are independently encoded by the pre-trained variational auto-encoder, then summed with the pose features $\mathbf{c}^{pose}$ as the input of the dual-stream transformer-based denoiser. As illustrated in Fig. 3(a), we redesign the transformer architecture with cross-modal specialized branches to explicitly model appearance and geometric distribution. Each modality maintains dedicated pathways in the initial and final transformer blocks for learning domain-specific modulation, while sharing the intermediate layers to facilitate cross-modal fusion for a unified generation. We incorporate LoRA [16] into the linear layers of all attention blocks, allowing the model to effectively develop multi-view 3D awareness and double-stream specialization, while maximally preserving the generation and generalization capabilities of the base model.

Cross-modal alignment. While the dual-stream architecture enables parallel generation of RGB and normal sequences, the lack of explicit cross-modal token interaction results in structural misalignment between the two modalities, which further propagates to 3D reconstruction as texture-geometry mismapping, particularly around body outlines and surface details. To address this issue, we propose replacing one out of every four shared DiT blocks with a cross-modal attention module, as demonstrated in Fig. 3 (b). In these modules, standard self-attention is performed on the concatenated features of the RGB and normal tokens, encouraging cross-domain correlation. At the cross-attention layers, the features are split back into the separate RGB and normal streams, which then interact independently with the corresponding global contexts.

This cross-modal alignment mechanism provides strongly coupled supervision signals necessary for generating intrinsically coherent 3D entities, effectively eliminating reconstruction artifacts from unaligned multi-view guidance and improving appearance quality.

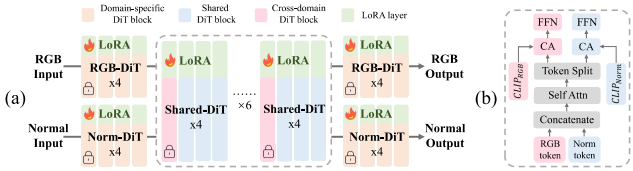


Figure 3. Illustration of model architecture. (a) Dual-stream diffusion transformer. (b) Cross-modal DiT block.

Progressive training scheme. We implement a two-stage training strategy to balance multimodal specialization and alignment. Initially, the model is trained with only modality-specific attention to build a strong foundation in modeling the distribution of each stream and establishing 3D consistency across views. In the second stage, we incorporate cross-modal attention modules for better alignment. To prevent potential quality degradation from excessive information fusion between the two branches, we randomly deactivate the cross-modal attention with a probability of 30% during training. This progressive approach allows the model to first develop high-quality unimodal representations for RGB and normal map generation with 3D awareness, and then achieve rapid convergence in cross-modal alignment during the subsequent stage.

3.2. Dynamic 3D Human Carving

While our dual-branch video model achieves joint generation of RGB and normal images with approximately accurate multi-view consistency, its training relies solely on 2D pixel-level supervision, lacking explicit 3D structural constraints. Consequently, slight color perturbations, pose jitter, or 3D spatial misalignments can result in noticeable surface artifacts under naive 3D reconstruction.

Inspired by recent advances in dynamic scene reconstruction [33, 61, 75, 84], we propose a novel 4D remeshing module that flexibly establishes vertex-level correspon-

dence across viewpoints for fine-grained geometry carving. This module decouples the learning of the static canonical mesh and view-dependent geometric variations. By tracking the coordinate offsets of canonical mesh vertices across all perspectives, we maintain a consistent topology of the deformed meshes from different viewpoints, which further enables the learning of a unified UV representation to effectively integrate appearance attributes shared across views.

3.2.1. 4D Remeshing

This module consists of two components: (1) a canonical mesh that encodes geometric structures shared across temporal viewpoints and spatial coordinates, and (2) a dynamic deformation field that captures view-dependent transient surface details. The two components are jointly optimized in a coordinated manner to reconstruct geometrically accurate meshes from inconsistent multi-view guidance.

Canonical mesh optimization. The canonical mesh is updated with continuous explicit 3D remeshing [40], which optimizes the learnable positions of mesh vertices via differentiable rendering under normal map supervision. To ensure the correctness of mesh topology and alleviate potential geometric defects, a series of surface remeshing techniques is applied, including edge splitting, collapsing, and flipping. An estimated optimal edge length is introduced for adaptive vertex density control, enabling fine-scale 3D mesh reconstruction.

View-dependent dynamic deformation. The dynamic deformation module tracks the movement of each vertex on the canonical mesh across multi-view inconsistent normal frames to align with the geometric details of each perspective. It employs an MLP network Ψ_d that takes as input the canonical vertex coordinates x_c and a viewpoint embedding, and outputs the corresponding vertex offset δx under the given view v, formulated as:

$$\delta x_i = \Psi_d(\gamma(x_c), \gamma(i)), \tag{2}$$

where $\gamma(\cdot)$ denotes the positional encoding, i is the index of viewpoint v_i ranging from 0 (front view) to T. The vertex coordinates x_c are detached before being fed into the MLP network, ensuring the deformation module focuses exclusively on learning view-dependent temporal deformations, without interfering with the optimization of the global canonical geometry.

Joint optimization. We first initialize the canonical mesh through static 3D remeshing with multi-view normals. Then, during the 4D remeshing phase, we iteratively update both the canonical vertex positions and the deformation field in a joint optimization cycle. In each iteration under view v_i, we compute the deformed vertex positions as:

$$x_d^i = x_c + \delta x_i. \tag{3}$$

By rendering the deformed mesh with differentiable rasterization [27] under the current view v_i, we obtain the resulting normal map $\hat{\mathcal{N}}_i$ and object mask $\hat{\mathcal{S}}_i$. The reconstruction supervision is enforced by computing the distance between the rendered normal $\hat{\mathcal{N}}_i$ and the generated normal frame $\mathcal{F}_i^{norm}$, as well as the distance between the rendered silhouette $\hat{\mathcal{S}}_i$ and the extracted foreground mask $\mathcal{S}_i$:

$$L_{rec} = ||\hat{\mathcal{N}}_i - \mathcal{F}_i^{norm}||_1 + ||\hat{\mathcal{S}}_i - \mathcal{S}_i||_1. \tag{4}$$

Besides, we apply the Laplacian loss on both canonical and deformed mesh surfaces for smoother shape regularization:

$$L_{lap} = \frac{1}{n}\sum_{i=1}^{n} ||\sigma_i||_2, \sigma_i = x_i - \frac{1}{|N(i)|}\sum_{j\in N(i)} x_j, \tag{5}$$

where n is the number of mesh vertices, and $N(i)$ is the set of one-ring neighbors for vertex x_i. We omit the subscripts denoting canonical and deformed vertices for simplicity. An ARAP (as-rigid-as-possible) loss [53] is also applied between pairs of deformed vertices from two randomly selected viewpoints, constraining the deformation network Ψ_d to learn near-rigid dynamics and prevent geometric distortions under single-view normal guidance:

$$L_{arap} = \sum_{i=1}^{n}\sum_{j\in N(i)} \left| ||x_i - x_j||_2 - ||x_i' - x_j'||_2 \right|, \tag{6}$$

where x and x' are the vertex positions under view v and v'. The overall training objective can be formulated as:

$$L_{geo} = L_{rec} + \lambda_{lap}L_{lap} + \lambda_{arap}L_{arap}. \tag{7}$$

The above-mentioned design enables explicit disentangled learning of the static canonical mesh and dynamic deformations. The canonical remeshing module learns globally shared spatial features, representing the minimal geometric intersection across all viewpoints; while the deformation network models view-specific offsets and fine-grained dynamics under the current guidance, and simultaneously maintains a global awareness of similar surface characteristics among different views. Once effectively learned, this module enables the inference of plausible geometric details on both visible and occluded regions.

3.2.2. Unified UV Representation

Similar to the normal branch, the RGB images also exhibit multi-view inconsistencies, primarily manifested as 3D spatial misalignment and subtle color variations. Considering that the deformed mesh for each viewpoint is obtained by applying position offsets to the canonical mesh, they inherently share the same topology. This property further enables the learning of a unified UV representation to fuse

Table 1. Quantitative comparison on arbitrary-pose reconstruction from single images. We compare our method with SOTA methods on the 2K2K and Sizer datasets. Our method surpasses all the baselines on both geometry and appearance metrics. Note that the training sets of Human3Diffusion include 2K2K data. ↑ and ↓ denote if higher or lower is better, respectively.

Method	CD (cm) ↓		P2S (cm) ↓		NC ↑		PSNR ↑		SSIM ↑		LPIPS ↓	
	2K2K	Sizer	2K2K	Sizer	2K2K	Sizer	2K2K	Sizer	2K2K	Sizer	2K2K	Sizer
SIFU [86]	2.251	1.997	2.159	2.005	0.712	0.721	21.40	18.63	0.888	0.908	0.143	0.127
SiTH [12]	1.702	2.379	1.571	2.224	0.752	0.713	21.61	17.87	0.899	0.909	0.124	0.133
Human3Diffusion [71]	1.052	1.331	1.198	1.348	0.774	0.727	23.35	19.57	0.909	0.919	0.104	0.104
PSHuman [29]	1.133	1.570	1.062	1.462	0.828	0.753	23.03	19.61	0.906	0.917	0.105	0.107
Ours	**0.817**	**1.243**	**0.778**	**1.161**	**0.858**	**0.768**	**24.49**	**20.38**	**0.917**	**0.926**	**0.086**	**0.089**

visible texture features across all generated RGB frames. We optimize the texture map $\mathcal{T}$ from a randomly initialized noise map using the differentiable rasterizer $\mathcal{R}$. Specifically, for view v_i, we utilize the deformed mesh M_i as the underlying geometry and obtain the rendered color image $\hat{C}_i = \mathcal{R}(M_i, \mathcal{T}, v_i)$ and the corresponding silhouette $\mathcal{S}_i$. This rendering, tightly aligned with the view-dependent geometric deformations, is supervised by the pixel-level loss against the corresponding RGB frame $\mathcal{F}_i^{rgb}$, formulated as:

$$L_{rgb} = w_i ||\hat{C}_i \cdot S_i - \mathcal{F}_i^{rgb} \cdot S_i||_2, \tag{8}$$

where we assign the view-specific weights w_i to emphasize primary viewpoints (e.g., front and back views). This process leverages the alignment between the normal and RGB channels, and utilizes the existing geometric deformation priors to effectively alleviate texture artifacts caused by RGB inconsistencies. Additionally, we apply a total-variation loss L_{tv} to smooth the optimized texture map. The final training loss for the texture optimization is:

$$L_{tex} = L_{rgb} + \lambda_{tv} L_{tv}, \tag{9}$$

where λ_{tv} is set to 0.5 by default.

4. Experimental Results

4.1. Implementation Details

Our 3D-aware dual-stream video generator is trained on the public 3D human dataset THuman2.1 [79], from which we filter out 1559 scans with high-quality geometric surfaces. We render 25 views at 480p resolution of both RGB and normal images evenly distributed around each 3D scan as training frames. The video model is trained on 2 A100 GPUs using the AdamW [38] optimizer with a learning rate of 1e-4. For the 4D remeshing module, we set the loss weights $\{\lambda_{lap}, \lambda_{arap}\}$ to $\{0.4, 0.03\}$. We first initialize the 3D avatar using only canonical mesh reconstruction for 300 iterations, then introduce the dynamic deformation field for joint optimization over another 200 iterations. The deformed meshes undergo UV unwrapping using Xatlas [78], followed by unified texture optimization for 500 iterations. We extract the deformed mesh corresponding to the front view as the final result. The inference of the 3D-aware video model, the optimization of 4D remeshing and unified texture representation take 5 minutes, 2 minutes, and 10 seconds on a single A6000 GPU, respectively.

4.2. Comparisons

Baselines. We compare our method with state-of-the-art approaches on two aspects: (1) arbitrary-pose 3D human reconstruction which focuses on the generalization capability to handle diverse and complicated poses, and (2) canonical-pose 3D human reconstruction that emphasizes the generation fidelity under simplified standard-pose scenarios. For arbitrary-pose reconstruction, we evaluate against mesh-based methods (SIFU [86], SiTH [12], PSHuman [29]) and a 3DGS-based approach (Human3Diffusion [71]). For canonical-pose reconstruction, we additionally compare with StdGen [11] and Hunyuan3D 2.0 [87], representing the SOTA performance in 3D avatar and versatile object generation, respectively. StdGen intrinsically supports arbitrary-pose to A-pose conversion. Other static image-based reconstruction methods are evaluated using our reposed images for fair comparison under standardized pose settings.

Evaluation dataset. For quantitative comparison, we randomly select 100 and 50 testing subjects from two public 3D human datasets, 2K2K [9] and Sizer [56], respectively. For qualitative comparison, we prepare a selection of in-the-wild images from the SHHQ dataset [8] and diverse self-collected data to test the generalization ability in real-world scenarios.

Arbitrary-pose reconstruction. *Quantitative Comparison.* We evaluate geometric quality using Chamfer Distance (CD), Point-to-Surface (P2S) distance, and Normal Consistency (NC). For appearance assessment, we employ Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), and Perceptual Similarity (LPIPS). As shown in Tab. 1, our method outperforms other baselines across all metrics, demonstrating its advantages in reconstructing high-fidelity 3D humans with accurate geometry and fine-grained appearance.

Qualitative Comparison. As shown in Fig. 4(a), SIFU

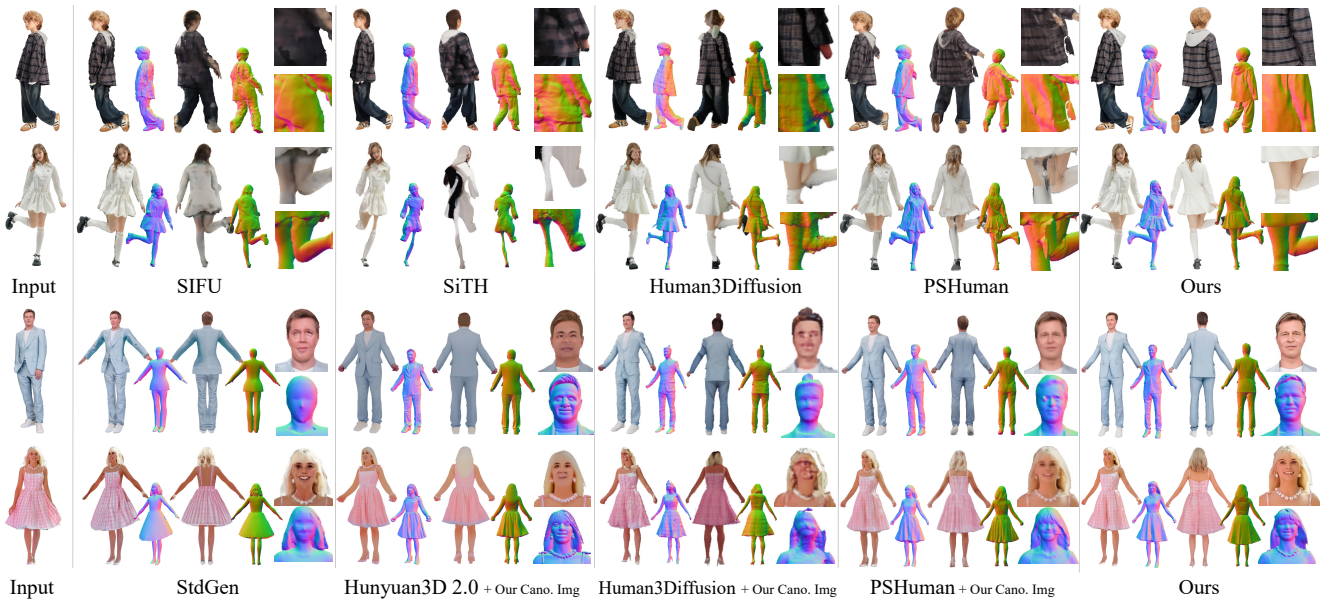


Figure 4. Visual comparison with SOTA methods on both arbitrary-pose and canonical-pose 3D human reconstruction. Given in-the-wild inputs, our method outperforms baselines in generating (1) plausible novel views with reasonable poses, and (2) fine-grained geometry and photorealistic appearance akin to real-world individuals. Please zoom in for better inspection.



Figure 5. Comparison with PSHuman on challenging cases including human-object interactions and rare back-view poses.

and SiTH generate implausible novel-view appearances and noisy surfaces. PSHuman exhibits noticeable artifacts due to its reliance on sparse and inconsistent multi-view guidance. These methods also produce unreasonable human poses resulting from inaccurate monocular SMPL estimation. Human3Diffusion suffers from low output resolution, leading to coarse renderings and distorted meshes. In contrast, our method outperforms all baselines in terms of geometric quality, pose plausibility, and texture realism.

Challenging Cases. Fig. 5 further compares our method with PSHuman on challenging scenarios involving human-object occlusion and uncommon poses. PSHuman fails to correctly interpret human-object interactions and back-view observations, resulting in geometric distortions in the chest and leg regions. In contrast, our method leverages a fine-tuned video diffusion model enriched with 3D awareness and strong prior knowledge to generate coherent novel views, providing dense and geometrically consistent guidance for 4D remeshing.

Table 2. Ablation study on cross-modal alignment and 4D remeshing. Results verify the effectiveness of the proposed modules.

Method	PSNR ↑	SSIM ↑	LPIPS ↓
w/o CMA	23.14	0.9064	0.1015
w/o 4DR	23.87	0.9142	0.0970
Full model	**24.49**	**0.9173**	**0.0858**

Canonical-pose reconstruction. We present qualitative comparisons for canonical-pose reconstruction in Fig. 4(b). StdGen and Hunyuan3D 2.0 generate smooth, high-poly meshes, but fail to preserve human identities from the reference image, hindering their generalization to real-world individuals. Human3Diffusion and PSHuman exhibit noticeable geometric defects around the face and blurred textures. On the contrary, our method yields significant improvements in both appearance realism and geometric sharpness, achieving lifelike 3D human reconstruction. More results are included in the supplementary.

4.3. Ablation Study

3D-aware dual-stream video generation. As illustrated in Fig. 6, we demonstrate the necessity of two key designs of the video generator: 3D-aware finetuning (3DFT) and cross-modal alignment (CMA). Without specialization on 3D consistent multi-view renderings, the generated frames

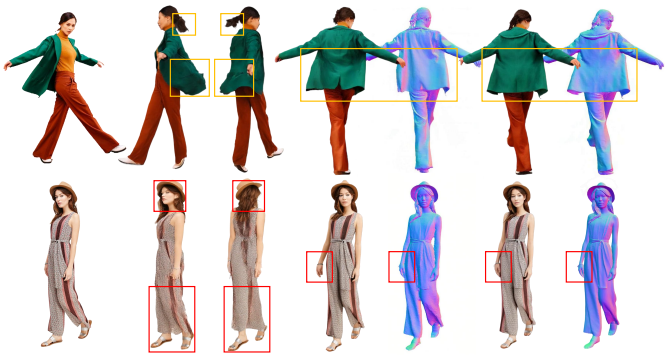


Figure 6. Ablation on 3D-aware finetuning and cross-modal alignment.

of the base model exhibit floating dynamics (e.g., hair and clothing) and the lack of structural consistency of the human anatomy. Meanwhile, the variant without explicit cross-modal alignment suffers from inconsistent outlines and misaligned details (e.g., clothing wrinkles) between the corresponding RGB and normal frames, which further lead to degraded reconstruction quality as shown in Tab. 2. Our full model achieves consistent 3D awareness and precise alignment between the two modalities in terms of both global layout and fine-scale structure. This provides spatially consistent and strongly coupled multi-view guidance, which is essential for generating intrinsically accordant 3D avatars.

Progressive training scheme (PT). As shown in Fig. 7, naively combining the training for joint dual-stream adaptation and cross-modal alignment in the video generation model results in degraded quality for both modalities, leading to noisy appearances in unseen regions and less-detailed normal maps. In contrast, our two-stage training strategy enables the model to progressively learn the distributions of both RGB and normal data, along with multimodal alignment capabilities. This approach achieves significant improvements in generating photorealistic novel views and high-frequency surface details (e.g., facial features, hairlines, and necklace geometry).

4D remeshing (4DR). During the dynamic human carving, the 4D remeshing module effectively disentangles the learning of the static canonical mesh from view-dependent geometric variations. We demonstrate its importance in Fig. 8 by comparing against a variant trained solely with static 3D remeshing. The variant produces results with severe surface cracks under exaggerated poses, failures in reconstructing complex topology, and coarse facial details, caused by 3D spatial misalignment and fine-scale inconsistencies of the multi-view guided frames. The 4D remeshing module offers a flexible approach for integrating view-dependent dynamics, enabling the generation of detailed avatars with spatially coherent geometry. In cooperation with the unified UV representation, it further enhances the overall reconstruction quality, as shown in Tab. 2.

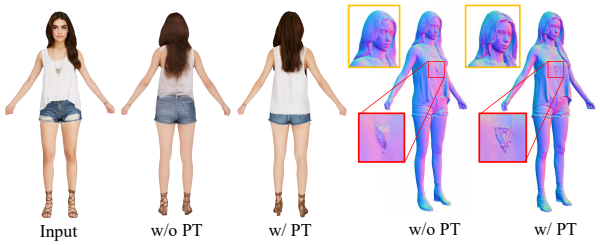


Figure 7. Ablation on the two-stage progressive training scheme.

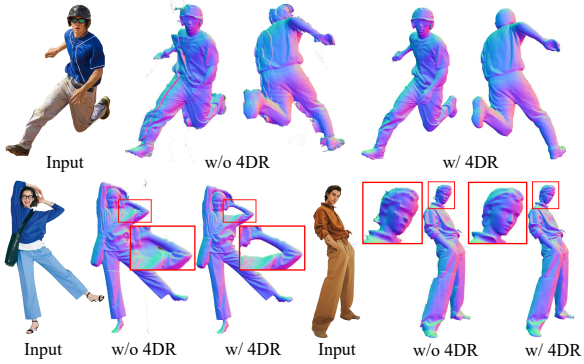


Figure 8. Ablation on the proposed 4D remeshing module.

4.4. Applications

Animation. As shown in Fig. 1, our method enables canonical-pose 3D human reconstruction from the reference image, which can be directly rigged and animated using standard pipelines, supporting effortless motion sequence generation.

3D editing. Our method facilitates intuitive 3D avatar editing through direct 2D pixel-level modifications on the reference image. Fig. 1 illustrates examples of 3D virtual try-on using the edited image with desired clothing as the model input. More results can be found in the supplementary.

5. Conclusion

In this paper, we presented FISHuman, a novel pipeline for reconstructing 3D humans with fine-grained geometry and photorealistic appearance from a single-view image. We adapted a 3D-aware dual-stream video generation model to synthesize multi-view, cross-modally aligned RGB and normal guidance. We demonstrated that naive 3D reconstruction from inconsistent frames results in surface distortion and texture blurring, and addressed this with a novel 4D remeshing module for fine-grained human mesh reconstruction. This module explicitly decouples the learning of a globally consistent canonical mesh from view-dependent per-vertex deformations. It further enables the optimization of a unified UV texture for photorealistic appearance integration. Experimental results validated the effectiveness of our method in generating lifelike 3D humans, and showcased its practical utility for animation and 3D editing.

Acknowledgments

This work was supported by National Natural Science Foundation of China (Grant No.: 62372015), Leading Projects in Key Research Fields of Language Funded by the National Language Commission, Center For Chinese Font Design and Research, Key Laboratory of Intelligent Press Media Technology, and National Engineering Research Center of New Electronic Publishing Technologies.

References

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

[2] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 2

[3] Tianshi Cao, Karsten Kreis, Sanja Fidler, Nicholas Sharp, and Kangxue Yin. Texfusion: Synthesizing 3d textures with text-guided image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4169–4181, 2023. 3

[4] Yiwen Chen, Tong He, Di Huang, Weicai Ye, Sijin Chen, Jiaxiang Tang, Xin Chen, Zhongang Cai, Lei Yang, Gang Yu, et al. Meshanything: Artist-created mesh generation with autoregressive transformers. *arXiv preprint arXiv:2406.10163*, 2024. 3

[5] Yiwen Chen, Zhihao Li, Yikai Wang, Hu Zhang, Qin Li, Chi Zhang, and Guosheng Lin. Ultra3d: Efficient and high-fidelity 3d generation with part attention. *arXiv preprint arXiv:2507.17745*, 2025. 3

[6] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13142–13153, 2023. 3

[7] Zijian Dong, Longteng Duan, Jie Song, Michael J Black, and Andreas Geiger. Moga: 3d generative avatar prior for monocular gaussian avatar reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13304–13314, 2025. 3

[8] Jianglin Fu, Shikai Li, Yuming Jiang, Kwan-Yee Lin, Chen Qian, Chen Change Loy, Wayne Wu, and Ziwei Liu. Stylegan-human: A data-centric odyssey of human generation. In *European Conference on Computer Vision*, pages 1–19. Springer, 2022. 6

[9] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12869–12879, 2023. 6

[10] Tong He, Yuanlu Xu, Shunsuke Saito, Stefano Soatto, and Tony Tung. Arch++: Animation-ready clothed human reconstruction revisited. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11046–11056, 2021. 3

[11] Yuze He, Yanning Zhou, Wang Zhao, Zhongkai Wu, Kaiwen Xiao, Wei Yang, Yong-Jin Liu, and Xiao Han. Stdgen: Semantic-decomposed 3d character generation from single images. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 26345–26355, 2025. 2, 3, 6, 13

[12] I Ho, Jie Song, Otmar Hilliges, et al. Sith: Single-view textured human reconstruction with image-conditioned diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 538–549, 2024. 2, 3, 6, 13

[13] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in neural information processing systems*, 35:8633–8646, 2022. 2

[14] 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

[15] Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. Lrm: Large reconstruction model for single image to 3d. *arXiv preprint arXiv:2311.04400*, 2023. 2

[16] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 4, 13

[17] Li Hu. 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*, pages 8153–8163, 2024. 2

[18] Li Hu, Guangyuan Wang, Zhen Shen, Xin Gao, Dechao Meng, Lian Zhuo, Peng Zhang, Bang Zhang, and Liefeng Bo. Animate anyone 2: High-fidelity character image animation with environment affordance. *arXiv preprint arXiv:2502.06145*, 2025. 2

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

[20] Yangyi Huang, Ye Yuan, Xueting Li, Jan Kautz, and Umar Iqbal. Adahuman: Animatable detailed 3d human generation with compositional multiview diffusion. *arXiv preprint arXiv:2505.24877*, 2025. 3

[21] Zeng Huang, Yuanlu Xu, Christoph Lassner, Hao Li, and Tony Tung. Arch: Animatable reconstruction of clothed humans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3093–3102, 2020. 3

[22] Adobe Inc. Mixamo. https://www.mixamo.com/. 14

[23] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time

radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023. 3, 13

[24] Rawal Khirodkar, Timur Bagautdinov, Julieta Martinez, Su Zhaoen, Austin James, Peter Selednik, Stuart Anderson, and Shunsuke Saito. Sapiens: Foundation for human vision models. In *European Conference on Computer Vision*, pages 206–228. Springer, 2024. 13, 15

[25] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 13

[26] Nikos Kolotouros, Thiemo Alldieck, Enric Corona, Eduard Gabriel Bazavan, and Cristian Sminchisescu. Instant 3d human avatar generation using image diffusion models. In *European Conference on Computer Vision*, pages 177–195. Springer, 2024. 3

[27] Samuli Laine, Janne Hellsten, Tero Karras, Yeongho Seol, Jaakko Lehtinen, and Timo Aila. Modular primitives for high-performance differentiable rendering. *ACM Transactions on Graphics (ToG)*, 39(6):1–14, 2020. 5

[28] Jiahao Li, Hao Tan, Kai Zhang, Zexiang Xu, Fujun Luan, Yinghao Xu, Yicong Hong, Kalyan Sunkavalli, Greg Shakhnarovich, and Sai Bi. Instant3d: Fast text-to-3d with sparse-view generation and large reconstruction model. *arXiv preprint arXiv:2311.06214*, 2023. 3

[29] Peng Li, Wangguandong Zheng, Yuan Liu, Tao Yu, Yangguang Li, Xingqun Qi, Xiaowei Chi, Siyu Xia, Yan-Pei Cao, Wei Xue, et al. Pshuman: Photorealistic single-image 3d human reconstruction using cross-scale multiview diffusion and explicit remeshing. *arXiv preprint arXiv:2409.10141*, 2024. 2, 3, 6, 13

[30] Weiyu Li, Jiarui Liu, Hongyu Yan, Rui Chen, Yixun Liang, Xuelin Chen, Ping Tan, and Xiaoxiao Long. Craftsman3d: High-fidelity mesh generation with 3d native generation and interactive geometry refiner. *arXiv preprint arXiv:2405.14979*, 2024. 3

[31] Yangguang Li, Zi-Xin Zou, Zexiang Liu, Dehu Wang, Yuan Liang, Zhipeng Yu, Xingchao Liu, Yuan-Chen Guo, Ding Liang, Wanli Ouyang, et al. Triposg: High-fidelity 3d shape synthesis using large-scale rectified flow models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 14

[32] Yuchen Lin, Chenguo Lin, Panwang Pan, Honglei Yan, Yiqiang Feng, Yadong Mu, and Katerina Fragkiadaki. Partcrafter: Structured 3d mesh generation via compositional latent diffusion transformers. *arXiv preprint arXiv:2506.05573*, 2025. 3

[33] Isabella Liu, Hao Su, and Xiaolong Wang. Dynamic gaussians mesh: Consistent mesh reconstruction from dynamic scenes. *arXiv preprint arXiv:2404.12379*, 2024. 4

[34] Minghua Liu, Ruoxi Shi, Linghao Chen, Zhuoyang Zhang, Chao Xu, Xinyue Wei, Hansheng Chen, Chong Zeng, Jiayuan Gu, and Hao Su. One-2-3-45++: Fast single image to 3d objects with consistent multi-view generation and 3d diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10072–10083, 2024. 3

[35] Yixin Liu, Kai Zhang, Yuan Li, Zhiling Yan, Chujie Gao, Ruoxi Chen, Zhengqing Yuan, Yue Huang, Hanchi Sun, Jianfeng Gao, et al. Sora: A review on background, technology, limitations, and opportunities of large vision models. *arXiv preprint arXiv:2402.17177*, 2024. 2

[36] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9970–9980, 2024. 3

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

[38] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[39] Yifang Men, Yuan Yao, Miaomiao Cui, and Liefeng Bo. Mimo: Controllable character video synthesis with spatial decomposed modeling. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 21181–21191, 2025. 2, 3

[40] Werner Palfinger. Continuous remeshing for inverse rendering. *Computer Animation and Virtual Worlds*, 33(5):e2101, 2022. 5, 13

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

[42] 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. *ACM Transactions on Graphics (TOG)*, 43(4): 1–13, 2024. 2, 3

[43] Lingteng Qiu, Xiaodong Gu, Peihao Li, Qi Zuo, Weichao Shen, Junfei Zhang, Kejie Qiu, Weihao Yuan, Guanying Chen, Zilong Dong, et al. Lhm: Large animatable human reconstruction model from a single image in seconds. *arXiv preprint arXiv:2503.10625*, 2025. 3, 13

[44] Lingteng Qiu, Shenhao Zhu, Qi Zuo, Xiaodong Gu, Yuan Dong, Junfei Zhang, Chao Xu, Zhe Li, Weihao Yuan, Liefeng Bo, et al. Anigs: Animatable gaussian avatar from a single image with inconsistent gaussian reconstruction. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 21148–21158, 2025. 3

[45] Xuanchi Ren, Jiahui Huang, Xiaohui Zeng, Ken Museth, Sanja Fidler, and Francis Williams. Xcube: Large-scale 3d generative modeling using sparse voxel hierarchies. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4209–4219, 2024. 3

[46] Elad Richardson, Gal Metzer, Yuval Alaluf, Raja Giryes, and Daniel Cohen-Or. Texture: Text-guided texturing of 3d shapes. In *ACM SIGGRAPH 2023 conference proceedings*, pages 1–11, 2023. 3

[47] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitiza-

tion. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2304–2314, 2019. 2, 3

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

[49] Ruizhi Shao, Youxin Pang, Zerong Zheng, Jingxiang Sun, and Yebin Liu. Human4dit: Free-view human video generation with 4d diffusion transformer. *arXiv e-prints*, pages arXiv–2405, 2024. 2

[50] Ruoxi Shi, Hansheng Chen, Zhuoyang Zhang, Minghua Liu, Chao Xu, Xinyue Wei, Linghao Chen, Chong Zeng, and Hao Su. Zero123++: a single image to consistent multi-view diffusion base model. *arXiv preprint arXiv:2310.15110*, 2023. 3

[51] Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. *arXiv preprint arXiv:2308.16512*, 2023. 3

[52] Yawar Siddiqui, Antonio Alliegro, Alexey Artemov, Tatiana Tommasi, Daniele Sirigatti, Vladislav Rosov, Angela Dai, and Matthias Nießner. Meshgpt: Generating triangle meshes with decoder-only transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19615–19625, 2024. 3

[53] Robert W Sumner, Johannes Schmid, and Mark Pauly. Embedded deformation for shape manipulation. In *ACM siggraph 2007 papers*, pages 80–es. 2007. 5

[54] 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 *European Conference on Computer Vision*, pages 1–18. Springer, 2024. 2, 3

[55] Meshy AI Team. Meshy ai, 2025. https://www.meshy.ai/. 14

[56] Garvita Tiwari, Bharat Lal Bhatnagar, Tony Tung, and Gerard Pons-Moll. Sizer: A dataset and model for parsing 3d clothing and learning size sensitive 3d clothing. In *European Conference on Computer Vision*, pages 1–18. Springer, 2020. 6

[57] Shuyuan Tu, Zhen Xing, Xintong Han, Zhi-Qi Cheng, Qi Dai, Chong Luo, and Zuxuan Wu. Stableanimator: High-quality identity-preserving human image animation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 21096–21106, 2025. 3

[58] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 2, 4, 13

[59] Boyuan Wang, Runqi Ouyang, Xiaofeng Wang, Zheng Zhu, Guosheng Zhao, Chaojun Ni, Guan Huang, Lihong Liu, and Xingang Wang. Humandreamer-x: Photorealistic single-image human avatars reconstruction via gaussian restoration. *arXiv preprint arXiv:2504.03536*, 2025. 3

[60] Xiang Wang, Shiwei Zhang, Longxiang Tang, Yingya Zhang, Changxin Gao, Yuehuan Wang, and Nong Sang. Unianimate-dit: Human image animation with large-scale video diffusion transformer. *arXiv preprint arXiv:2504.11289*, 2025. 2, 3, 14

[61] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20310–20320, 2024. 4

[62] Shuang Wu, Youtian Lin, Feihu Zhang, Yifei Zeng, Yikang Yang, Yajie Bao, Jiachen Qian, Siyu Zhu, Xun Cao, Philip Torr, et al. Direct3d-s2: Gigascale 3d generation made easy with spatial sparse attention. *arXiv preprint arXiv:2505.17412*, 2025. 3

[63] Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, and Jiaolong Yang. Structured 3d latents for scalable and versatile 3d generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 21469–21480, 2025. 3

[64] Bojun Xiong, Jialun Liu, Jiakui Hu, Chenming Wu, Jinbo Wu, Xing Liu, Chen Zhao, Errui Ding, and Zhouhui Lian. Texgaussian: Generating high-quality pbr material via octree-based 3d gaussian splatting. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 551–561, 2025. 3

[65] Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J Black. Icon: Implicit clothed humans obtained from normals. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13286–13296. IEEE, 2022. 2, 3

[66] Yuliang Xiu, Jinlong Yang, Xu Cao, Dimitrios Tzionas, and Michael J Black. Econ: Explicit clothed humans optimized via normal integration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 512–523, 2023. 2, 3

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

[68] Yinghao Xu, Hao Tan, Fujun Luan, Sai Bi, Peng Wang, Jiahao Li, Zifan Shi, Kalyan Sunkavalli, Gordon Wetzstein, Zexiang Xu, et al. Dmv3d: Denoising multi-view diffusion using 3d large reconstruction model. *arXiv preprint arXiv:2311.09217*, 2023. 3

[69] Yinghao Xu, Zifan Shi, Wang Yifan, Hansheng Chen, Ceyuan Yang, Sida Peng, Yujun Shen, and Gordon Wetzstein. Grm: Large gaussian reconstruction model for efficient 3d reconstruction and generation. In *European Conference on Computer Vision*, pages 1–20. Springer, 2024. 3

[70] Zhongcong Xu, Jianfeng Zhang, Jun Hao Liew, Hanshu Yan, Jia-Wei Liu, Chenxu Zhang, Jiashi Feng, and Mike Zheng Shou. Magicanimate: Temporally consistent human image animation using diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1481–1490, 2024. 2

[71] Yuxuan Xue, Xianghui Xie, Riccardo Marin, and Gerard Pons-Moll. Human-3diffusion: Realistic avatar creation via

explicit 3d consistent diffusion models. *Advances in Neural Information Processing Systems*, 37:99601–99645, 2024. 2, 3, 6, 13

[72] Tao Yang, Peiran Ren, Xuansong Xie, and Lei Zhang. Gan prior embedded network for blind face restoration in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 672–681, 2021. 13

[73] Yunhan Yang, Yuan-Chen Guo, Yukun Huang, Zi-Xin Zou, Zhipeng Yu, Yangguang Li, Yan-Pei Cao, and Xihui Liu. Holopart: Generative 3d part amodal segmentation. *arXiv preprint arXiv:2504.07943*, 2025. 3

[74] Zhendong Yang, Ailing Zeng, Chun Yuan, and Yu Li. Effective whole-body pose estimation with two-stages distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4210–4220, 2023. 3

[75] Ziyi Yang, Xinyu Gao, Wen Zhou, Shaohui Jiao, Yuqing Zhang, and Xiaogang Jin. Deformable 3d gaussians for high-fidelity monocular dynamic scene reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20331–20341, 2024. 4

[76] 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. 2

[77] Chongjie Ye, Yushuang Wu, Ziteng Lu, Jiahao Chang, Xiaoyang Guo, Jiaqing Zhou, Hao Zhao, and Xiaoguang Han. Hi3dgen: High-fidelity 3d geometry generation from images via normal bridging. *arXiv preprint arXiv:2503.22236*, 3:2, 2025. 3

[78] Jonathan Young. Xatlas., 2021. https://github.com/jpcy/xatlas. 6

[79] Tao Yu, Zerong Zheng, Kaiwen Guo, Pengpeng Liu, Qionghai Dai, and Yebin Liu. Function4d: Real-time human volumetric capture from very sparse consumer rgbd sensors. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR2021)*, 2021. 6

[80] Andy Zeng, Shuran Song, Matthias Nießner, Matthew Fisher, Jianxiong Xiao, and Thomas Funkhouser. 3dmatch: Learning local geometric descriptors from rgb-d reconstructions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1802–1811, 2017. 13

[81] Xianfang Zeng, Xin Chen, Zhongqi Qi, Wen Liu, Zibo Zhao, Zhibin Wang, Bin Fu, Yong Liu, and Gang Yu. Paint3d: Paint anything 3d with lighting-less texture diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4252–4262, 2024. 3

[82] Biao Zhang, Jiapeng Tang, Matthias Niessner, and Peter Wonka. 3dshape2vecset: A 3d shape representation for neural fields and generative diffusion models. *ACM Transactions On Graphics (TOG)*, 42(4):1–16, 2023. 3

[83] Longwen Zhang, Ziyu Wang, Qixuan Zhang, Qiwei Qiu, Anqi Pang, Haoran Jiang, Wei Yang, Lan Xu, and Jingyi Yu. Clay: A controllable large-scale generative model for creating high-quality 3d assets. *ACM Transactions on Graphics (TOG)*, 43(4):1–20, 2024. 3

[84] Shuai Zhang, Guanjun Wu, Zhoufeng Xie, Xinggang Wang, Bin Feng, and Wenyu Liu. Dynamic 2d gaussians: Geometrically accurate radiance fields for dynamic objects. *arXiv preprint arXiv:2409.14072*, 2024. 4

[85] Yuang Zhang, Jiaxi Gu, Li-Wen Wang, Han Wang, Junqi Cheng, Yuefeng Zhu, and Fangyuan Zou. Mimicmotion: High-quality human motion video generation with confidence-aware pose guidance. *arXiv preprint arXiv:2406.19680*, 2024. 2

[86] Zechuan Zhang, Zongxin Yang, and Yi Yang. Sifu: Side-view conditioned implicit function for real-world usable clothed human reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9936–9947, 2024. 2, 3, 6, 13

[87] Zibo Zhao, Zeqiang Lai, Qingxiang Lin, Yunfei Zhao, Haolin Liu, Shuhui Yang, Yifei Feng, Mingxin Yang, Sheng Zhang, Xianghui Yang, et al. Hunyuan3d 2.0: Scaling diffusion models for high resolution textured 3d assets generation. *arXiv preprint arXiv:2501.12202*, 2025. 3, 6, 13, 14

[88] Zerong Zheng, Tao Yu, Yebin Liu, and Qionghai Dai. Pamir: Parametric model-conditioned implicit representation for image-based human reconstruction. *IEEE transactions on pattern analysis and machine intelligence*, 44(6): 3170–3184, 2021. 2

[89] Shenhao Zhu, Junming Leo Chen, Zuozhuo Dai, Zilong Dong, Yinghui Xu, Xun Cao, Yao Yao, Hao Zhu, and Siyu Zhu. Champ: Controllable and consistent human image animation with 3d parametric guidance. In *European Conference on Computer Vision*, pages 145–162. Springer, 2024. 2, 3

[90] Shenhao Zhu, Lingteng Qiu, Xiaodong Gu, Zhengyi Zhao, Chao Xu, Yuxiao He, Zhe Li, Xiaoguang Han, Yao Yao, Xun Cao, et al. Mcmat: Multiview-consistent and physically accurate pbr material generation. *arXiv preprint arXiv:2412.14148*, 2024. 3

[91] Yiyu Zhuang, Jiaxi Lv, Hao Wen, Qing Shuai, Ailing Zeng, Hao Zhu, Shifeng Chen, Yujiu Yang, Xun Cao, and Wei Liu. Idol: Instant photorealistic 3d human creation from a single image. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 26308–26319, 2025. 3