HumanRAM: Feed-forward Human Reconstruction and Animation Model using Transformers

ZHIYUAN YU*†, Hong Kong University of Science and Technology, China
ZHE LI*, Huawei, China
HUJUN BAO, State Key Laboratory of CAD&CG, Zhejiang University, China
CAN YANG‡, Hong Kong University of Science and Technology, China
XIAOWEI ZHOU‡, State Key Laboratory of CAD&CG, Zhejiang University, China

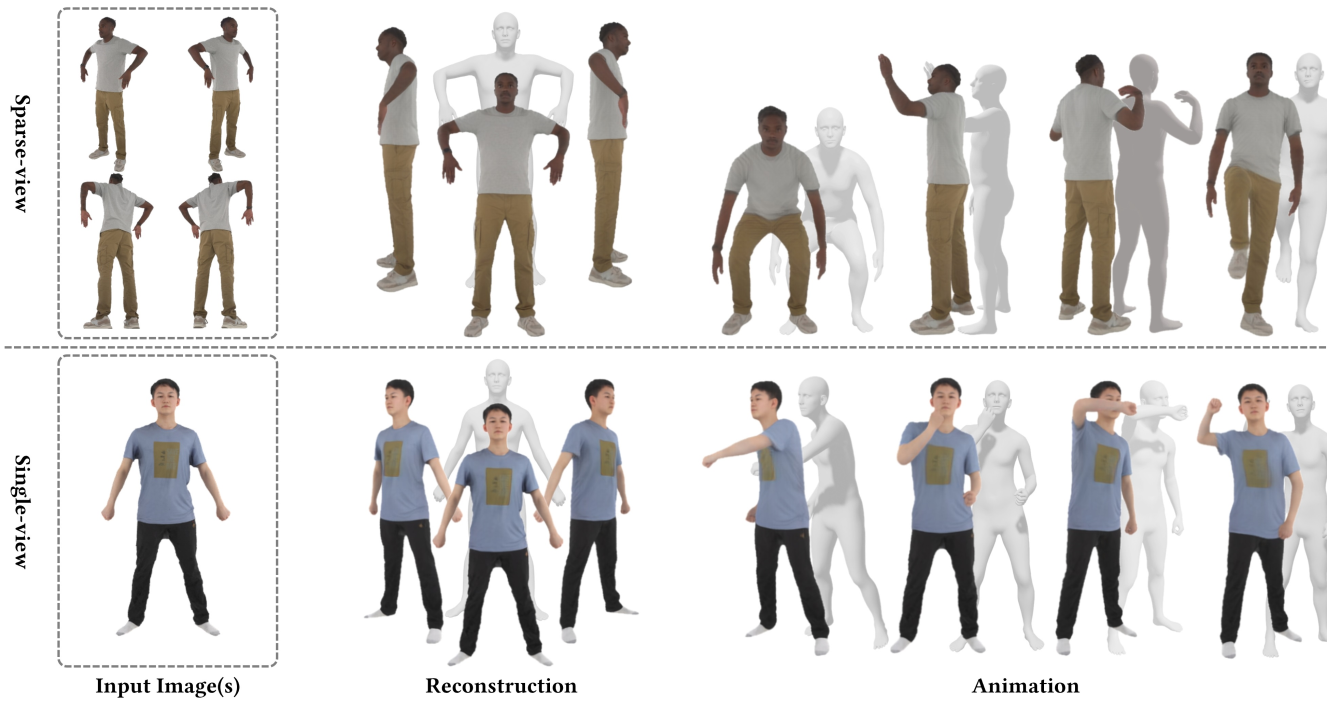


Fig. 1. **We propose HumanRAM, a novel approach for feed-forward novel view synthesis (reconstruction) and novel pose synthesis (animation) from sparse/single-view human image(s). The animation poses are from ActorsHQ [Işık et al. 2023] and AMASS [Mahmood et al. 2019].**

3D human reconstruction and animation are long-standing topics in computer graphics and vision. However, existing methods typically rely on sophisticated dense-view capture and/or time-consuming per-subject optimization procedures. To address these limitations, we propose HumanRAM, a novel feed-forward approach for generalizable human reconstruction and animation from monocular or sparse human images. Our approach integrates human reconstruction and animation into a unified framework by introducing explicit pose conditions, parameterized by a shared SMPL-X neural texture, into transformer-based large reconstruction models (LRM). Given monocular or sparse input images with associated camera parameters and SMPL-X poses, our model employs scalable transformers and a DPT-based decoder to synthesize realistic human renderings under novel viewpoints and novel poses. By leveraging the explicit pose conditions, our model simultaneously enables high-quality human reconstruction and high-fidelity pose-controlled animation. Experiments show that HumanRAM significantly

*The first two authors contributed equally to this work.
†Work done during an internship at Huawei.
‡Corresponding authors.

Authors' addresses: Zhiyuan Yu, Hong Kong University of Science and Technology, Mathematics, Hong Kong, China, zyuaq@ust.hk; Zhe Li, Huawei, Hangzhou, China, lizhe_thu@126.com; Hujun Bao, State Key Laboratory of CAD&CG, Zhejiang University, Hangzhou, China, bao@cad.zju.edu.cn; Can Yang, Hong Kong University of Science and Technology, Mathematics, Hong Kong, China, macyang@ust.hk; Xiaowei Zhou, State Key Laboratory of CAD&CG, Zhejiang University, Hangzhou, China, xwzhou@zju.edu.cn.

SIGGRAPH Conference Papers '25, August 10–14, 2025, Vancouver, BC, Canada

ACM ISBN 979-8-4007-1540-2/2025/08
https://doi.org/10.1145/3721238.3730605

surpasses previous methods in terms of reconstruction accuracy, animation fidelity, and generalization performance on real-world datasets. Video results are available at https://zju3dv.github.io/humanram/.

CCS Concepts: • **Computing methodologies → Rendering**; **Animation**.

Additional Key Words and Phrases: Human reconstruction, human animation, neural rendering

ACM Reference Format:
Zhiyuan Yu, Zhe Li, Hujun Bao, Can Yang, and Xiaowei Zhou. 2025. HumanRAM: Feed-forward Human Reconstruction and Animation Model using Transformers. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers (SIGGRAPH Conference Papers '25), August 10–14, 2025, Vancouver, BC, Canada.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3721238.3730605

1 INTRODUCTION

Reconstruction and animation are two core topics in human-centric 3D vision and graphics. Although high-end dense-view capture systems and modeling technologies [Collet et al. 2015; Guo et al. 2019; Işık et al. 2023; Li et al. 2024b] achieve high-quality 3D human reconstruction and animation, the complicated hardware and time-consuming per-subject optimization limit their broader applications.

This bottleneck leads to a trend of sparse/single-view reconstruction, which employs generalizable feed-forward networks to predict 3D humans directly from limited inputs. Pioneering works like PIFu [Saito et al. 2019, 2020] proposed to reconstruct human geometry via occupancy fields, but paid less attention to photo-realistic rendering demanded by real applications. Recently, with significant advances in differentiable rendering [Kerbl et al. 2023; Laine et al. 2020; Mildenhall et al. 2020] and neural rendering [Tewari et al. 2020; Thies et al. 2019], researchers start using scalable transformers [Vaswani 2017] to predict 3D representations or novel-view renderings from input images. These approaches originate from the Large Reconstruction Model (LRM) [Hong et al. 2023]. After being exposed to a large amount of 3D data (e.g., Objaverse [Deitke et al. 2023]), LRM and its follow-ups [Jin et al. 2024; Wei et al. 2024; Zhang et al. 2024a] learn to predict 3D models or images in a single forward pass. While LRMs achieve feed-forward 3D reconstruction, they struggle with fine-grained details in human geometry under complex poses and cloth deformations. Moreover, existing frameworks focus on static reconstruction, ignoring dynamic animations that are essential for interactive applications.

To address these challenges, we propose Human Reconstruction and Animation Model (HumanRAM), a novel framework that integrates human reconstruction and animation into a unified feed-forward model. We leverage Large View Synthesis Model (LVSM) [Jin et al. 2024] as the foundational architecture, which implicitly learns 3D structures and directly regresses novel-view renderings. Previous explicit 3D representations usually require precise geometry for high-quality output. However, geometric constraints are insufficient under sparse observations. By harnessing the implicit nature of LVSM, we overcome this limitation and improve the model's performance and generalization capacity.

Original LVSM maps input images & cameras, as well as the target camera, into patch tokens and regresses the target-view image using transformers. To endow LVSM with the animation ability and improve the reconstruction quality for humans, we introduce SMPL-X [Pavlakos et al. 2019], a parametric human mesh model that provides strong pose and geometry priors as additional input tokens. Given calibrated multi-view human images, SMPL-X can be estimated using off-the-shelf tools [Sun et al. 2024b; Zhang et al. 2021]. To tokenize the SMPL-X prior, we introduce rasterization of a shared neural texture map bound to the SMPL-X mesh across input and target views. This process yields pose images that spatially align with the RGB and camera tokens used in LVSM. These pose images serve as a strong geometrical and semantical guide for the transformer's attention mechanism, thereby enabling a more realistic novel-view synthesis and pose-controlled animation.

The key insights of our method are: 1) The rasterized pose images establish shared embedding space between input and target views, providing explicit correspondences for the self-attention layers of transformers to reassemble the target view, thus producing higher-fidelity reconstruction. 2) The pose images enable LVSM to match appearance across diverse poses through the shared neural texture map, thereby achieving realistic animation. Moreover, we propose a DPT-based [Ranftl et al. 2021] decoder to facilitate the information exchange among neighboring patches and intermediate transformer features, effectively suppressing checkerboard artifacts prevalent in linear decoders.

Overall, the synergy of design choices enables high-quality human reconstruction and photo-realistic human animation from sparse/single image(s), as shown in Fig. 1 and Fig. 4.

2 RELATED WORK

2.1 Generalizable Human Reconstruction

Human reconstruction has been widely explored over the past few decades. Traditional methods reconstruct human geometry and texture from dense-view images [Bradley et al. 2008; Collet et al. 2015; Guo et al. 2019; Liu et al. 2009; Starck and Hilton 2007; Vlasic et al. 2009; Wu et al. 2011]. With advancements in differentiable 3D representations like implicit functions [Chabra et al. 2020; Mescheder et al. 2019; Park et al. 2019; Peng et al. 2020], neural radiance fields (NeRF) [Mildenhall et al. 2020], and 3D Gaussian splatting (3DGS) [Kerbl et al. 2023], researchers tend to learn data-driven feed-forward models. Methods like BodyNet [Varol et al. 2018], DeepHuman [Zheng et al. 2019], and Tang et al. [2023] regress volumetric outputs from image(s) but face resolution limits from GPU memory constraints. PIFu [Saito et al. 2019] and its successors [Cao et al. 2023; Saito et al. 2020; Xiu et al. 2023, 2022; Yang et al. 2024c; Yu et al. 2021; Zhang et al. 2024c,d; Zheng et al. 2021a,b] address this by learning pixel-aligned implicit functions for human geometry recovery.

In recent years, many works have developed generalizable models based on NeRF or 3DGS for human novel view synthesis. Similar to PIFu, NeRF-based approaches [Chen et al. 2024a, 2023; Hu et al. 2023b; Kwon et al. 2021; Lin et al. 2022; Mihajlovic et al. 2022; Raj et al. 2021; Shao et al. 2022; Sun et al. 2024a; Zhou et al. 2024a] extract pixel-aligned features and learn image-conditioned radiance fields. In contrast, 3DGS-based methods explicitly parameterize 3D gaussians in pixel space [Dong et al. 2024; Hu et al. 2024b; Tu et al. 2024; Zheng et al. 2024; Zhou et al. 2024b], UV space [Kwon et al. 2024] or tokens [Prospero et al. 2024]. More recently, 3D AIGC has made remarkable progress [Liu et al. 2024b, 2023; Poole et al. 2023;

Voleti et al. 2024], leading researchers to model the reconstruction as an image-conditioned generation task [AlBahar et al. 2023; Cao et al. 2024; Chen et al. 2024b; Gao et al. 2024; He et al. 2024; Huang et al. 2024; I Ho et al. 2024; Kolotouros et al. 2024; Li et al. 2024a; Liu et al. 2024c; Sengupta et al. 2024; Weng et al. 2023; Xiu et al. 2024; Xu et al. 2023b; Yang et al. 2024a]. Although diffusion models enhance texture hallucination for occluded regions, their iterative refinement process incurs higher computational costs than feed-forward models.

2.2 Large Reconstruction Model

Large Reconstruction Model (LRM) was first proposed by Hong et al. [2023], which learns a generalizable NeRF [Mildenhall et al. 2020] from a single image. Subsequent works [Tang et al. 2025; Wang et al. 2023b; Wei et al. 2024; Weng et al. 2023; Xie et al. 2024; Xu et al. 2024b, 2023a; Zhang et al. 2024a] explore LRM in various downstream tasks. For instance, PF-LRM [Wang et al. 2023b] learns from unposed images. LRM-Zero [Xie et al. 2024] and MegaSynth [Jiang et al. 2024] train LRM on synthetic data and successfully generalize to real data. DMV3D [Xu et al. 2023a] applies LRM as a diffusion denoiser to improve generation view consistency. Some researchers extend the representation of LRM from NeRF to mesh [Wei et al. 2024], 3DGS [Liang et al. 2024b; Shen et al. 2024; Tang et al. 2025; Xu et al. 2024b; Yi et al. 2024; Zhang et al. 2024a; Ziwen et al. 2024] and 2DGS [Chen et al. 2024c]. More recently, LVSM [Jin et al. 2024] synthesizes novel views using pure transformers. Despite these advances, existing LRM variants mainly focus on object/scene reconstruction, ignoring human-centric applications. In contrast, our method specializes in human reconstruction and animation.

2.3 Human Animation

Human animation aims to generate novel-pose images given one or more input images. Previous works are categorized into 2D and 3D animation. 2D animation formulates the task as signal-driven image generation [Chan et al. 2019; Liu et al. 2019; Ren et al. 2020; Siarohin et al. 2019a,b, 2021; Yu et al. 2023b; Zhang et al. 2022; Zhao and Zhang 2022]. Recently, diffusion-based methods [Hu et al. 2023a; Ma et al. 2023; Men et al. 2024; Shao et al. 2024a; Wang et al. 2023a; Xu et al. 2024c; Zhang et al. 2024b; Zhu et al. 2024] have gained huge attention for their powerful generation capabilities, but they suffer from time-consuming generation due to the denoising process.

3D methods typically optimize person-specific avatars from single or multi-view videos using various 3D representations (e.g., point clouds [Su et al. 2023], mesh [Bagautdinov et al. 2021; Chen et al. 2024d], implicit field [Jiang et al. 2022b; Li et al. 2023; Peng et al. 2024; Wang et al. 2022; Xu et al. 2024a; Zhang et al. 2023], NeRF [Jiang et al. 2022a,c; Li et al. 2022; Liu et al. 2021, 2024a; Peng et al. 2021a; Xiao et al. 2024; Xu et al. 2022; Yu et al. 2023a] and 3DGS [Hu et al. 2024a; Kocabas et al. 2024; Lei et al. 2024; Li et al. 2024b; Lin et al. 2024; Moon et al. 2024; Shao et al. 2024b; Wen et al. 2024; Zielonka et al. 2025]). The avatars are then animated using linear blend skinning (LBS). However, the optimization process is time-intensive and can fail with very sparse inputs. To generalize, researchers use learned priors [Chatziagapi et al. 2024; Hsuan-I Ho and Hilliges 2023; Mu et al. 2023] or feed-forward models [Gao et al. 2023, 2022; He et al. 2021; Huang et al. 2020; Kwon et al. 2023; Shin et al. 2025]. Our method bridges human prior and Large Reconstruction Model, leading to more realistic animation.

3 METHOD

3.1 Preliminary: LVSM

Large View Synthesis Model (LVSM) [Jin et al. 2024] is a recent method for neural rendering without using any explicit 3D representations. This method inputs multi-view images and camera parameters and outputs target-view renderings through encoder-decoder or decoder-only transformers. Specifically, given N images with their corresponding camera poses parameterized by Plücker ray embeddings [Plücker 1865], denoted as $\{\mathbf{I}_i \in \mathbb{R}^{H\times W\times 3}, \mathbf{P}_i \in \mathbb{R}^{H\times W\times 6} | i = 1, ..., N\}$, LVSM first maps them into patch tokens $\mathbf{x}_{ij} \in \mathbb{R}^d$ with a linear layer (d is the token dimension):

$$\mathbf{x}_{ij} = \text{Linear}_{\text{inp}}([\mathbf{I}_{ij}, \mathbf{P}_{ij}]), \tag{1}$$

where $\mathbf{I}_{ij} \in \mathbb{R}^{3p^2}$ and $\mathbf{P}_{ij} \in \mathbb{R}^{6p^2}$ mean the j-th $p\times p$ patch of $\mathbf{I}_i$ and $\mathbf{P}_i$, and $[\cdot,\cdot]$ means concatenation. The target-view pose is also represented as Plücker ray embedding $\mathbf{P}^t \in \mathbb{R}^{H\times W\times 6}$ and mapped to patch tokens $\mathbf{q}_j \in \mathbb{R}^d$ with another linear layer:

$$\mathbf{q}_j = \text{Linear}_{\text{tar}}(\mathbf{P}^t_j). \tag{2}$$

Given input and target tokens, decoder-only LVSM synthesizes target-view tokens $\mathbf{y}_j \in \mathbb{R}^d$ through transformers $\mathcal{T}$:

$$\mathbf{x}'_i, ..., \mathbf{x}'_{l_x}, \mathbf{y}_1, ..., \mathbf{y}_{l_q} = \mathcal{T}(\mathbf{x}_i, ..., \mathbf{x}_{l_x}, \mathbf{q}_1, ..., \mathbf{q}_{l_q}), \tag{3}$$

where l_x and l_q mean the number of input and target tokens. Finally, LVSM regresses the RGB values of each target patch from output tokens with a linear layer followed by a Sigmoid function:

$$\hat{\mathbf{I}}^t_j = \text{Sigmoid}(\text{Linear}_{\text{out}}(\mathbf{y}_j)) \in \mathbb{R}^{3p^2}. \tag{4}$$

The predicted RGB values are unpatchfied to 2D space to form the final target image. In this paper, we incorporate dedicated designs into LVSM for human reconstruction and animation.

3.2 Overview

Given sparse-view images of a character, we aim to synthesize the character under novel views and poses, i.e., to perform feed-forward human reconstruction and animation. As a state-of-the-art feed-forward large reconstruction model (LRM), LVSM [Jin et al. 2024] is introduced as a foundational architecture of our method. To endow LRM with the animation ability, we introduce pose tokens parameterized by a neural texture [Thies et al. 2019] bound with SMPL-X [Pavlakos et al. 2019] into LVSM. Specifically, as illustrated in Fig. 2, we render the learnable SMPL-X neural texture to the sparse input views, resulting in N feature maps, which we refer to as pose images. The input RGB images, their corresponding Plücker embeddings, and the pose images are concatenated and then patchfied as *input tokens*. Given the target view and the target human poses to be synthesized, we similarly concatenate and patchify them as *target tokens*. The input and target tokens are fed into a transformer model, and the *output tokens* are regressed to produce the synthesized human image under the target view and pose.

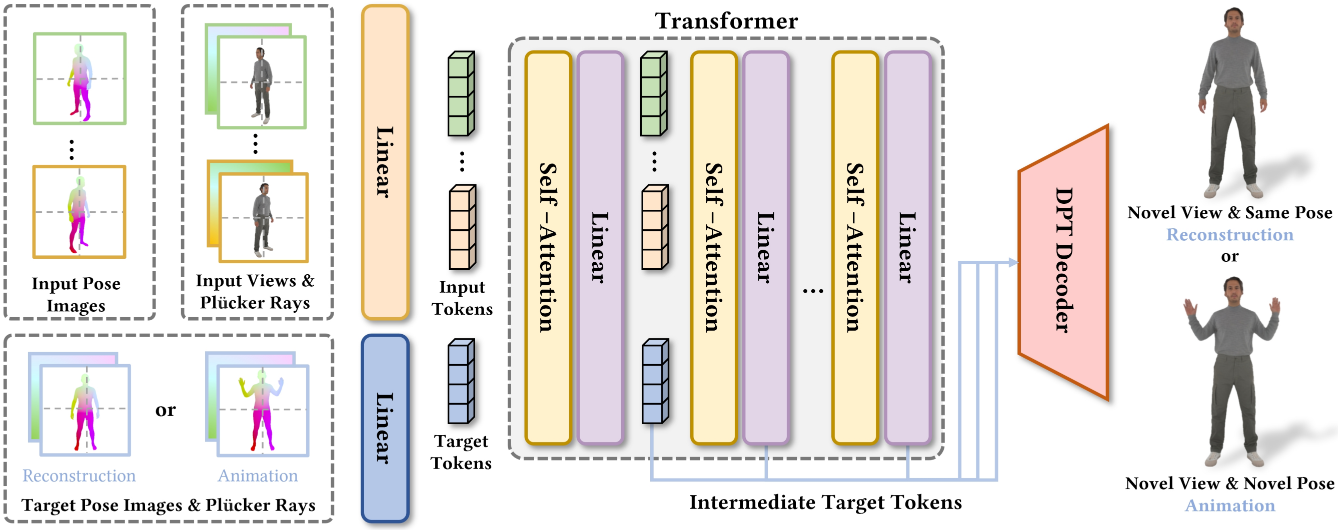


Fig. 2. **Pipeline of HumanRAM.** HumanRAM adopts transformers for human reconstruction and animation from sparse view images in a feed-forward manner. We first patchify and project spare-view RGB images and their corresponding Plücker rays and pose images into input tokens through a linear layer. The pose images are acquired by rasterizing the SMPL-X neural texture onto the input views. Similarly, given the target novel view under the same or another novel pose, the target tokens are created from the target Plücker rays and pose images through another linear layer. Then both input tokens and target tokens are fed into transformer blocks. Finally, a DPT-based decoder regresses the intermediate target tokens to a high-fidelity human image under the target view and target pose. Overall, HumanRAM realizes feed-forward reconstruction and animation by controlling the target views and target poses at the input end.

3.3 Pose-conditioned Reconstruction and Animation

Since LVSM discards explicit 3D representation, we cannot directly use the SMPL-X [Pavlakos et al. 2019] model as a geometry prior or proxy for pose-conditioned reconstruction and animation, as done in previous works [Huang et al. 2020; Kwon et al. 2024; Taubner et al. 2024; Xiu et al. 2023; Zheng et al. 2021b]. Inspired by neural texture [Deng et al. 2024; Thies et al. 2019; Yoon et al. 2022], we render the SMPL-X mesh with a neural texture onto multi-view 2D image planes, generating multi-view pose conditions. These pose conditions serve not only as a strong geometry prior for the novel view synthesis but also as an enabler for the animation ability.

SMPL-X Neural Texture. We adopt tri-planes [Chan et al. 2021] to represent the neural texture for its effectiveness and compactness. As illustrated in Fig. 3, the neural texture is defined as learnable feature tri-planes within a canonical space, determined by SMPL-X with canonical pose and mean shape. We denote the canonical SMPL-X vertices as $\mathbf{V}_{\text{cano}} \in \mathbb{R}^{N_V \times 3}$ and feature tri-planes as $\mathbf{T} \in \mathbb{R}^{3\times H'\times W'\times C}$, where H' and W' are the resolution of each plane, and C is the feature dimension. For each position $\mathbf{v} \in \mathbb{R}^3$ on the canonical SMPL-X surface, its corresponding neural texture is the concatenation of sampled features on each plane:

$$\begin{aligned}\mathbf{F}(\mathbf{v};\mathbf{T}) = [&\text{BLerp}(\mathbf{v}^{xy};\mathbf{T}^{xy}),\\ &\text{BLerp}(\mathbf{v}^{xz};\mathbf{T}^{xz}), \text{BLerp}(\mathbf{v}^{yz};\mathbf{T}^{yz})] \in \mathbb{R}^{3C},\end{aligned} \tag{5}$$

where $\text{BLerp}(\cdot)$ is the bilinear interpolation function on the feature plane given 2D query coordinates. The SMPL-X neural texture is shared across all the identities, providing guidance for pose-conditioned reconstruction and animation.

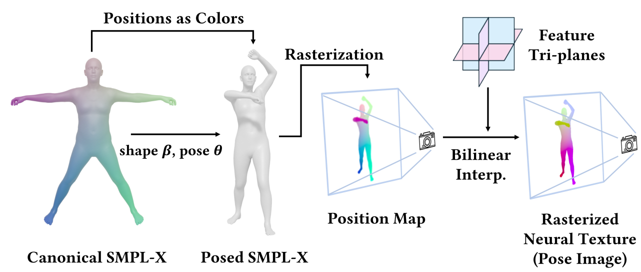


Fig. 3. **Illustration of the process of neural texture rasterization.** We first render position maps with canonical SMPL-X as vertex colors, and then the position maps are used to sample triplane-based neural texture.

Input Tokens. Given calibrated multi-view input images of a character, we obtain the registered SMPL-X mesh using multi-view motion capture like [Zhang et al. 2021]. The registered SMPL-X serves as a geometry proxy. We then bind its vertex attributes with the canonical positions $\mathbf{V}_{\text{cano}}$ and rasterize it to the input views, producing N position maps $\{\mathbf{V}_i \in \mathbb{R}^{H\times W\times 3}\}$, where each pixel corresponds to a canonical position. The rendered position maps are then used to sample the neural texture using Eq. 5. Consequently, we rasterize the SMPL-X neural texture onto the input views, obtaining N pose images $\{\mathbf{F}_i \in \mathbb{R}^{H\times W\times 3C}\}$. Then, similar to Eq. 1, we concatenate RGB images, Plücker embeddings, and pose images along the channel dimension, and then patchify them as "input tokens":

$$\mathbf{x}_{ij} = \text{Linear}_{\text{inp}}([\mathbf{I}_{ij}, \mathbf{P}_{ij}, \mathbf{F}_{ij}]). \tag{6}$$

Target Tokens for Reconstruction. Given a new target viewpoint, we rasterize the neural texture onto it using the registered SMPL-X model to obtain a novel view pose image $\mathbf{F}^t$. Similar to Eq. 2, we concatenate the Plücker embeddings and pose image of the target view, and then patchify them as "target tokens" for reconstruction:

$$\mathbf{q}_j^{\text{recon}} = \text{Linear}_{\text{tar}}([\mathbf{P}_j^t, \mathbf{F}_j^t]). \tag{7}$$

Target Tokens for Animation. Since the rasterized neural texture provides rich human pose information, it is natural to explore our model's potential animation ability for both novel view and pose synthesis. Specifically, given a novel target pose $\hat{\theta}$, we first transform it into a posed SMPL-X model. We then rasterize the neural texture onto a novel view using the posed SMPL-X, obtaining the pose image $\hat{\mathbf{F}}^t$ representing both the novel view and the novel pose. Following Eq. 7, we acquire "target tokens" $\{\mathbf{q}_j^{\text{ani}}\}$ for animation.

Output Tokens. The input and target tokens are subsequently fed to a decoder-only transformer $\mathcal{T}$, composed of a series of self-attention layers, producing a sequence of "output tokens". These output tokens and intermediate features are decoded as an image under the target view and target pose using a DPT-based decoder [Ranftl et al. 2021] (Sec. 3.4).

Discussion on Pose Conditions. We discuss the impact of the neural texture-based pose images on both reconstruction and animation.

- From the perspective of novel view synthesis, i.e., reconstruction, LVSM [Jin et al. 2024] learns cross-view matching from RGB and camera pose information to reassemble a novel view image using the attention mechanism. Our method introduces additional SMPL-X neural texture into the matching process, providing more explicit correspondences for higher-quality view synthesis, as demonstrated in Fig. 5, compared to vanilla LVSM.
- On the other end of the spectrum, the pose condition enables texture matching across different poses with a shared neural texture, achieving novel pose synthesis. To the best of our knowledge, HumanRAM is the first to endow Large Reconstruction Model with the animation ability.

3.4 DPT-based Decoder

Original LVSM [Jin et al. 2024] uses a linear layer to decode tokens into RGB values directly. We empirically find that such a simple decoder yields patch-like artifacts for humans, especially in regions suffering severe self-occlusions or containing thin structures, as shown in Fig. 8. We hypothesize that such artifacts are attributed to the lack of information exchange between neighboring patch tokens when decoding. Inspired by the dense prediction transformers (DPT) used in various vision transformer models [Oquab et al. 2023; Ranftl et al. 2021; Wang et al. 2024; Yang et al. 2024b], we replace the linear layer with stacks of residual CNN layers, similar to DPT heads, enhancing the local information fusion. Therefore, the final synthesized image $\hat{\mathbf{I}}^t$ is formed using a DPT-based decoder:

$$\hat{\mathbf{I}}^t = \text{Sigmoid}(\text{DPT}(\{\mathbf{y}^i | i = 3, 6, 9, 12\}), \tag{8}$$

where $\mathbf{y}^i$ denotes the intermediate tokens of the i-th layer.

3.5 Loss Functions

Given the predicted target-view images $\{\hat{\mathbf{I}}_i \in \mathbb{R}^{H\times W\times 3} | i = 1, ..., M\}$, we optimize HumanRAM using the following objective:

$$\mathcal{L} = \frac{1}{M}\Sigma_{i=1}^{M}(\mathcal{L}_{\text{MSE}}(\hat{\mathbf{I}}_i, \mathbf{I}_i) + \lambda \cdot \mathcal{L}_{\text{Perc}}(\hat{\mathbf{I}}_i, \mathbf{I}_i)), \tag{9}$$

where $\mathcal{L}_{MSE}$ denotes the mean squared error and $\mathcal{L}_{\text{Perc}}$ denotes the perceptual loss [Chen and Koltun 2017], computing L_1 difference between the extracted features from the VGG-19 network Φ [Simonyan and Zisserman 2014]. λ is the loss weight of $\mathcal{L}_{\text{Perc}}$ and set to 1.0 in our experiments.

4 EXPERIMENTS

As shown in Fig. 1 and Fig. 4, our method can create realistic human reconstruction and animation from single and sparse images since our transformer-based architecture is flexible to the image token number. Video results can be found in the Supp. video.

4.1 Settings

The implementation and training details are presented in the Supp. document.

Datasets. We conduct experiments on four public datasets: *THuman2.1* [Yu et al. 2021], *Human4DiT* [Shao et al. 2024a], *ZJUMoCap* [Peng et al. 2021b], and *ActorsHQ* [Işık et al. 2023] for training and evaluation. THuman2.1 and Human4DiT comprise thousands of high-quality 3D human scans, texture maps, and SMPL-X [Pavlakos et al. 2019] fittings. We use the training set of 2300 scans from THuman2.1. The training scans are categorized according to human identities, enabling the model to learn animation across different poses of the same identity. Each training scan is normalized into a $[-1, 1]^3$ bounding box and rendered to 60-view images at a resolution of 512 via Cycles [Community 2018]. The cameras are randomly sampled with an altitude of $[-45°, 45°]$ and a radius of $[2.0, 3.0]$. ZJUMoCap and ActorsHQ are human avatar datasets that provide multi-view human videos and SMPL(-X)s. We convert the SMPL [Loper et al. 2015] parameters into SMPL-X for ZJUMoCap and use the SMPL-X provided by Li et al. [2024b] for ActorsHQ. All images are resized to 512×512 for aligning the input resolution of networks.

Baselines. We compare with generalizable human reconstruction methods GPS-Gaussian [Zheng et al. 2024] and GHG [Kwon et al. 2024]. We also compare against LRM-like methods LaRa [Chen et al. 2024c] and LVSM [Jin et al. 2024]. For animation, we compare with generalizable human avatar methods NNA [Gao et al. 2023] and SHERF [Hu et al. 2023b], as well as a personalized avatar method 3DGS-Avatar [Qian et al. 2024].

Metrics. We utilize the Peak Signal-to-Noise Ratio (PSNR), Structure Similarity Index Measure (SSIM) [Wang et al. 2004], and Learned Perceptual Image Patch Similarity (LPIPS) [Zhang et al. 2018] as metrics to assess the results quantitatively and qualitatively. PSNR and SSIM are evaluated on mask-cropped images, while LPIPS is computed on full-size images.

Fig. 4. **Qualitative results on ActorsHQ [Işık et al. 2023] and THuman2.1 [Yu et al. 2021].** The top two rows show the reconstruction and animation results from multi-view inputs, while the bottom two rows show the results from single-view input. The driving poses for animation are from ActorsHQ [Işık et al. 2023] and AMASS [Mahmood et al. 2019].

Table 1. **Quantitative comparison of reconstruction on THuman2.1 [Yu et al. 2021] and Human4DiT [Shao et al. 2024a]**. We report PSNR, SSIM, and LPIPS to evaluate the reconstruction quality. All methods are trained or finetuned on THuman2.1 for fair comparison.

Metrics	THuman2.1					Human4DiT				
	Ours	LVSM	GPS-Gaussian	GHG	LaRa	Ours	LVSM	GPS-Gaussian	GHG	LaRa
PSNR↑	**30.34**	28.24	22.11	21.88	23.71	**26.35**	25.56	20.87	19.47	22.91
SSIM↑	**0.9535**	0.9396	0.9007	0.8780	0.8913	**0.9373**	0.9247	0.8953	0.8539	0.8900
LPIPS↓	**0.0184**	0.0226	0.0421	0.0517	0.0679	**0.0211**	0.0248	0.0419	0.0586	0.0663

4.2 Comparison on Reconstruction

We compare HumanRAM with baselines on synthetic and real-world datasets. For synthetic dataset, we randomly select 200 scans from Thuman2.1 [Yu et al. 2021] and Human4DiT [Shao et al. 2024a] as the test set. We input 4 uniform views for all methods except GPS-Gaussian [Zheng et al. 2024], which requires 5 equal-height images for reasonable stereo rectification. The qualitative results are shown in Fig. 5. LaRa [Chen et al. 2024c] demonstrates blurry results due to its low-resolution volume representation, limiting its ability to model complicated geometries and textures. GPS-Gaussian is inclined to generate incomplete results because its stereo matching may fail when input views are sparse. GHG [Kwon et al. 2024] applies multi-scaffold SMPL-X [Pavlakos et al. 2019] mesh as the geometry proxy, which cannot handle loose cloth and tends to produce artifacts if severe self-occlusion occurs. LVSM [Jin et al. 2024] fails to synthesize fine-grained structures like hands and faces due to the

Fig. 5. **Qualitative comparisons for reconstruction on THuman2.1 [Yu et al. 2021] and Human4DiT [Shao et al. 2024a]**. We input 4 multi-view images of unseen subjects, and our method achieves a more faithful rendering compared to other reconstruction methods. The first four rows are from THuman2.1 and the last two rows are from Human4DiT. The red boxes indicate the improvements of our method over LVSM [Jin et al. 2024].

lack of human priors. Tab. 1 reports the numerical comparison on reconstruction. Overall, our method significantly outperforms previous methods both qualitatively and quantitatively. To demonstrate the generalization ability of HumanRAM, we conduct experiments on ActorsHQ [Işık et al. 2023] and in-the-wild images. For ActorsHQ, we select 5 uniform cameras as input and sample 100 frames per subject for evaluation. The results are shown in Tab. 2 and Fig. 6. All previous methods fail to generate reasonable results on real-captured data. In contrast, our proposed SMPL-X neural texture provides transformer blocks with coarse correspondences for cross-view matching, leading to better generalization. For in-the-wild images, we present the qualitative results of HumanRAM in the Supp. document.

4.3 Comparison on Animation

We compare HumanRAM with generalizable approaches including NNA [Gao et al. 2023] and SHERF [Hu et al. 2023b], as well as a personalized approach 3DGS-Avatar [Qian et al. 2024], on ZJU-MoCap [Peng et al. 2021b]. We train HumanRAM and SHERF on THuman2.1 [Yu et al. 2021] and ZJUMoCap [Peng et al. 2021b]. We use official weights for NNA since its training code has not been released. For 3DGS-Avatar, we train it on input views and animate it with novel poses. The evaluation is conducted on 100 randomly selected frames for each test subject. The multi-view animation results are shown in Tab. 3 and Fig. 7. The single-view results are shown in Tab. 4 and the Supp. document. 3DGS-Avatar requires a lengthy video to learn per-subject pose-dependent deformation. However, when the data size is limited (single frame in our experiments), it is prone to overfitting the input images, leading to severe artifacts in novel views and poses. NNA and SHERF learn a generalizable canonical avatar from the input image(s) and deform it to a novel pose using LBS wrapping. Compared to 3DGS-Avatar, these generalizable methods achieve better animation results owing to the data-driven prior learning. However, their canonical representation suffers from blurred textures and overfitting. Besides, LBS wrapping tends to produce unnatural deformation in the underarm region. Conversely, HumanRAM returns more realistic results in terms of quality and quantity thanks to the human structure prior learned through our dedicated designs. We further present in-the-wild animation results in the Supp. document to demonstrate the generalization capacity of HumanRAM.

Table 2. **Quantitative comparison of reconstruction on ActorsHQ [Işık et al. 2023]**. All methods are evaluated directly on ActorsHQ without training or finetuning.

Metrics	Ours	LVSM	GHG	LaRa
PSNR↑	**25.47**	20.25	18.01	19.98
SSIM↑	**0.9088**	0.8023	0.7922	0.8177
LPIPS↓	**0.0350**	0.0724	0.0880	0.0945

4.4 Ablation Study

Core components. We conduct ablation studies to evaluate the impact of our core components, i.e., *Pose Image* and *DPT-based Decoder*.

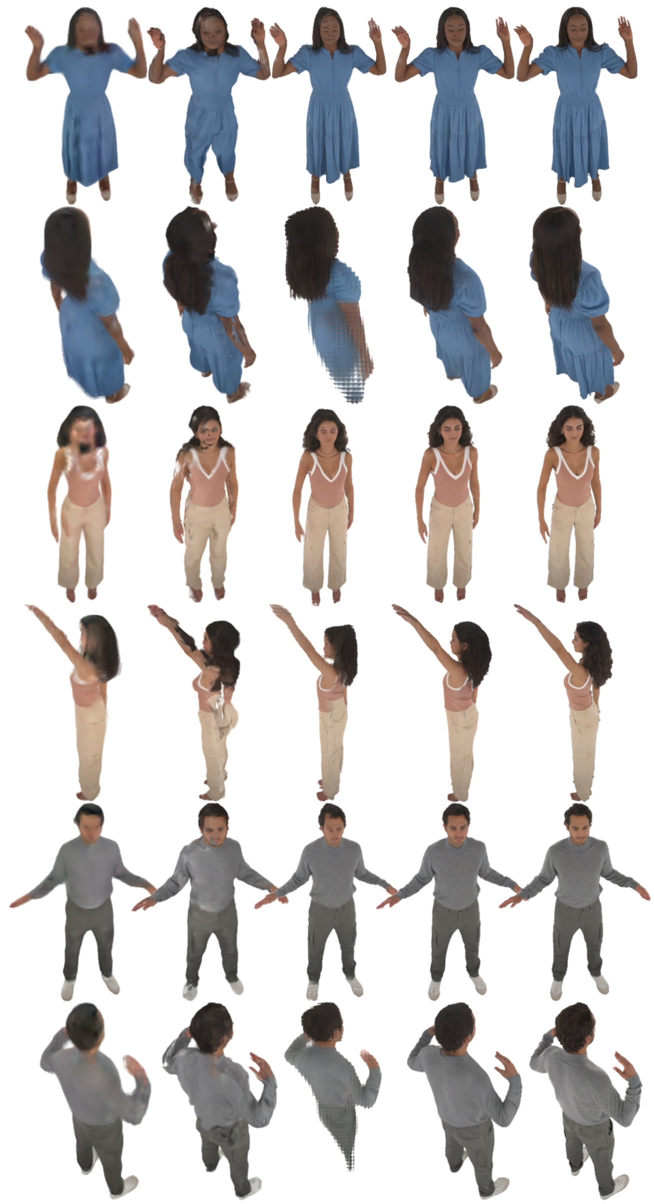


Fig. 6. **Qualitative comparisons for reconstruction on ActorsHQ [Işık et al. 2023]**. We input 5 multi-view images, and our method achieves a more faithful rendering compared to other state-of-the-art generalizable reconstruction methods.

The experiments are evaluated on the THuman2.1 [Yu et al. 2021] dataset as shown in Tab. 5 and Fig. 8. "Position" means replacing the pose image with 3-dim position maps, and this ablation shows that such replacement decreases the performance, indicating the superiority of learnable neural texture. "Linear" means replacing the DPT-based decoder with a linear layer used in vanilla LVSM. Experiments show that the skip-connection and convolution operations

Fig. 7. **Qualitative comparisons for multi-view animation on ZJUMoCap [Peng et al. 2021b]**. We input 4 multi-view images of the unseen subject, and our method achieves a more photo-realistic rendering compared to other methods.

in DPT are helpful in integrating information from multiple scales and neighboring patches, thus eliminating the patch-like artifacts and improving the overall visual quality.

Number of Views. We evaluate the impact of view numbers on THuman2.1 [Yu et al. 2021]. The model is directly evaluated on different numbers of input views without finetuning. Tab. 5 shows the rendering quality increases with more input views, which aligns with the performance pattern reported in LVSM [Jin et al. 2024].

Table 3. **Quantitative comparison of multi-view animation on ZJU-MoCap [Peng et al. 2021b]**. Metrics are computed on unseen subjects using the same crop manner as NNA [Gao et al. 2023].

Method	PSNR↑	SSIM↑	LPIPS↓
NNA	21.29	0.9369	0.0530
3DGS-Avatar	18.50	0.8367	0.0499
Ours	**23.40**	**0.9529**	**0.0252**

Table 4. **Quantitative comparison of single-view animation on ZJU-MoCap [Peng et al. 2021b]**. Metrics are computed on unseen poses and unseen subjects following SHERF [Hu et al. 2023b].

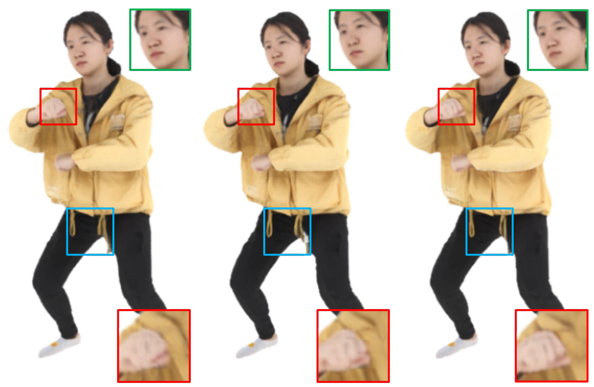
	Method	PSNR↑	SSIM↑	LPIPS↓
Unseen Poses	SHERF	18.56	0.8760	0.0501
	3DGS-Avatar	17.28	0.8243	0.0778
	Ours	**21.07**	**0.9152**	**0.0234**
Unseen Subjects	SHERF	17.80	0.8768	0.0536
	3DGS-Avatar	17.97	0.8481	0.0687
	Ours	**20.63**	**0.9184**	**0.0250**

Table 5. **Ablation study on THuman2.1 [Yu et al. 2021]**. We report PSNR, SSIM, and LPIPS to evaluate the contribution of proposed components and the impact of different input views.

Method	PSNR↑	SSIM↑	LPIPS↓
Position + DPT	29.32	0.9443	0.0197
Pose Image + Linear	30.07	0.9526	0.0186
Ours (Pose Image + DPT, 4 views)	**30.34**	**0.9535**	**0.0184**
Ours (1 view)	21.69	0.8834	0.0479
Ours (2 views)	25.01	0.9097	0.0344
Ours (8 views)	**32.34**	**0.9663**	**0.0150**

5 DISCUSSION

Conclusion. We propose HumanRAM, a novel generalizable feed-forward model for human reconstruction and animation. We integrate human reconstruction and animation into a unified framework by introducing pose conditions into large reconstruction models. We introduce a shared SMPL-X neural texture and rasterize it onto input and target views to associate correspondences across different views and poses, enabling higher-quality reconstruction and realistic animation. Overall, our method outperforms other state-of-the-art methods in terms of novel view and pose synthesis, both qualitatively and quantitatively.



Position + DPT   Pose Image + Linear   Ours

Fig. 8. **Qualitative comparisons for ablations on proposed core components.** Compared with positions, pose image helps capture detailed structures (indicated by red and blue boxes). Furthermore, the DPT-based decoder helps reduce the patch-like artifacts in regions with severe self-occlusion (red box), thin structures (blue box), and face (green box).

Limitation. Our method cannot handle high-resolution image inputs since the token number increases quadratically with the image resolution. One possible solution is to transfer the inputs and outputs from the high-resolution RGB space to the compressed low-resolution latent space, like WonderLand [Liang et al. 2024a] and HumanSplat [Pan et al. 2024].

ACKNOWLEDGMENTS

This work was partially supported by the following grants: National Key R&D Program of China (No. 2024YFB2809105), NSFC (No. U24B20154, No. 62172364), Zhejiang Provincial Natural Science Foundation of China (No. LR25F020003), Information Technology Center and State Key Lab of CAD&CG, Zhejiang University, and Research Grants Council of the Hong Kong Special Administrative Region, China (Project Reference Number: AoE/E-601/24-N).

REFERENCES

Badour AlBahar, Shunsuke Saito, Hung-Yu Tseng, Changil Kim, Johannes Kopf, and Jia-Bin Huang. 2023. Single-image 3d human digitization with shape-guided diffusion. In *SIGGRAPH Asia 2023 Conference Papers*. 1–11.

Timur Bagautdinov, Chenglei Wu, Tomas Simon, Fabian Prada, Takaaki Shiratori, Shih-En Wei, Weipeng Xu, Yaser Sheikh, and Jason Saragih. 2021. Driving-signal aware full-body avatars. *ACM Transactions on Graphics (TOG)* 40, 4 (2021), 1–17.

Derek Bradley, Tiberiu Popa, Alla Sheffer, Wolfgang Heidrich, and Tamy Boubekeur. 2008. Markerless garment capture. In *ACM SIGGRAPH*. 1–9.

Yukang Cao, Yan-Pei Cao, Kai Han, Ying Shan, and Kwan-Yee K Wong. 2024. Dreamavatar: Text-and-shape guided 3d human avatar generation via diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 958–968.

Yukang Cao, Kai Han, and Kwan-Yee K. Wong. 2023. SeSDF: Self-evolved Signed Distance Field for Implicit 3D Clothed Human Reconstruction. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Rohan Chabra, Jan E Lenssen, Eddy Ilg, Tanner Schmidt, Julian Straub, Steven Lovegrove, and Richard Newcombe. 2020. Deep local shapes: Learning local sdf priors for detailed 3d reconstruction. In *ECCV*. Springer, 608–625.

Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A Efros. 2019. Everybody Dance Now. In *IEEE International Conference on Computer Vision (ICCV)*.

Eric R. Chan, Connor Z. Lin, Matthew A. Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas Guibas, Jonathan Tremblay, Sameh Khamis, Tero Karras, and Gordon Wetzstein. 2021. Efficient Geometry-aware 3D Generative Adversarial Networks. In *arXiv*.

Aggelina Chatziagapi, Grigorios G. Chrysos, and Dimitris Samaras. 2024. MIGS: Multi-Identity Gaussian Splatting via Tensor Decomposition. In *ECCV*.

Anpei Chen, Haofei Xu, Stefano Esposito, Siyu Tang, and Andreas Geiger. 2024c. LaRa: Efficient Large-Baseline Radiance Fields. In *European Conference on Computer Vision (ECCV)*.

Jinnan Chen, Chen Li, and Gim Hee Lee. 2024a. DiHuR: Diffusion-Guided Generalizable Human Reconstruction. *arXiv preprint arXiv:2411.11903* (2024).

Jinnan Chen, Chen Li, Jianfeng Zhang, Lingting Zhu, Buzhen Huang, Hanlin Chen, and Gim Hee Lee. 2024b. Generalizable Human Gaussians from Single-View Image. *arXiv preprint* (2024).

Jianchuan Chen, Wentao Yi, Liqian Ma, Xu Jia, and Huchuan Lu. 2023. GM-NeRF: Learning Generalizable Model-based Neural Radiance Fields from Multi-view Images. In *CVPR*.

Qifeng Chen and Vladlen Koltun. 2017. Photographic image synthesis with cascaded refinement networks. In *Proceedings of the IEEE international conference on computer vision*. 1511–1520.

Yushuo Chen, Zerong Zheng, Zhe Li, Chao Xu, and Yebin Liu. 2024d. MeshAvatar: Learning High-quality Triangular Human Avatars from Multi-view Videos. arXiv:2407.08414 [cs.CV] https://arxiv.org/abs/2407.08414

Alvaro Collet, Ming Chuang, Pat Sweeney, Don Gillett, Dennis Evseev, David Calabrese, Hugues Hoppe, Adam Kirk, and Steve Sullivan. 2015. High-quality streamable free-viewpoint video. *ACM Transactions on Graphics (ToG)* 34, 4 (2015), 1–13.

Blender Online Community. 2018. *Blender - a 3D modelling and rendering package*. Blender Foundation, Stichting Blender Foundation, Amsterdam. http://www.blender.org

Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. 2023. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13142–13153.

Xiang Deng, Zerong Zheng, Yuxiang Zhang, Jingxiang Sun, Chao Xu, Xiaodong Yang, Lizhen Wang, and Yebin Liu. 2024. RAM-Avatar: Real-time Photo-Realistic Avatar from Monocular Videos with Full-body Control. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1996–2007.

Zheng Dong, Ke Xu, Yaoan Gao, Hujun Bao, Weiwei Xu, and Rynson WH Lau. 2024. Gaussian Surfel Splatting for Live Human Performance Capture. *ACM Transactions on Graphics (TOG)* 43, 6 (2024), 1–17.

Qingzhe Gao, Yiming Wang, Libin Liu, Lingjie Liu, Christian Theobalt, and Baoquan Chen. 2023. Neural novel actor: Learning a generalized animatable neural representation for human actors. *IEEE Transactions on Visualization and Computer Graphics* (2023).

Xiangjun Gao, Xiaoyu Li, Chaopeng Zhang, Qi Zhang, Yanpei Cao, Ying Shan, and Long Quan. 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 (CVPR)*. 10084–10094.

Xiangjun Gao, Jiaolong Yang, Jongyoo Kim, Sida Peng, Zicheng Liu, and Xin Tong. 2022. MPS-NeRF: Generalizable 3D Human Rendering From Multiview Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2022), 1–12. https://doi.org/10.1109/TPAMI.2022.3205910

Kaiwen Guo, Peter Lincoln, Philip Davidson, Jay Busch, Xueming Yu, Matt Whalen, Geoff Harvey, Sergio Orts-Escolano, Rohit Pandey, Jason Dourgarian, et al. 2019. The relightables: Volumetric performance capture of humans with realistic relighting. *ACM Transactions on Graphics (TOG)* 38, 6 (2019), 1–19.

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

Xu He, Xiaoyu Li, Di Kang, Jiangnan Ye, Chaopeng Zhang, Liyang Chen, Xiangjun Gao, Han Zhang, Zhiyong Wu, and Haolin Zhuang. 2024. MagicMan: Generative Novel View Synthesis of Humans with 3D-Aware Diffusion and Iterative Refinement. arXiv:2408.14211 [cs.CV]

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

Jie Song Hsuan-I Ho, Lixin Xue and Otmar Hilliges. 2023. Learning Locally Editable Virtual Humans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Hezhen Hu, Zhiwen Fan, Tianhao Wu, Yihan Xi, Seoyoung Lee, Georgios Pavlakos, and Zhangyang Wang. 2024a. Expressive Gaussian Human Avatars from Monocular RGB Video. In *NeurIPS*.

Li Hu, Xin Gao, Peng Zhang, Ke Sun, Bang Zhang, and Liefeng Bo. 2023a. Animate Anyone: Consistent and Controllable Image-to-Video Synthesis for Character Animation. *arXiv preprint arXiv:2311.17117* (2023).

Shoukang Hu, Fangzhou Hong, Liang Pan, Haiyi Mei, Lei Yang, and Ziwei Liu. 2023b. Sherf: Generalizable human nerf from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 9352–9364.

Yingdong Hu, Zhening Liu, Jiawei Shao, Zehong Lin, and Jun Zhang. 2024b. EVA-Gaussian: 3D Gaussian-based Real-time Human Novel View Synthesis under Diverse

Camera Settings. *arXiv preprint arXiv:2410.01425* (2024).

Xin Huang, Ruizhi Shao, Qi Zhang, Hongwen Zhang, Ying Feng, Yebin Liu, and Qing Wang. 2024. Humannorm: Learning normal diffusion model for high-quality and realistic 3d human generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4568–4577.

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

Hsuan I Ho, Jie Song, and Otmar Hilliges. 2024. SiTH: Single-view Textured Human Reconstruction with Image-Conditioned Diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 538–549.

Mustafa Işık, Martin Rünz, Markos Georgopoulos, Taras Khakhulin, Jonathan Starck, Lourdes Agapito, and Matthias Nießner. 2023. HumanRF: High-Fidelity Neural Radiance Fields for Humans in Motion. *ACM Transactions on Graphics (TOG)* 42, 4 (2023), 1–12. https://doi.org/10.1145/3592415

Boyi Jiang, Yang Hong, Hujun Bao, and Juyong Zhang. 2022b. SelfRecon: Self Reconstruction Your Digital Avatar from Monocular Video. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Hanwen Jiang, Zexiang Xu, Desai Xie, Ziwen Chen, Haian Jin, Fujun Luan, Zhixin Shu, Kai Zhang, Sai Bi, Xin Sun, Jiuxiang Gu, Qixing Huang, Georgios Pavlakos, and Hao Tan. 2024. MegaSynth: Scaling Up 3D Scene Reconstruction with Synthesized Data. (2024).

Tianjian Jiang, Xu Chen, Jie Song, and Otmar Hilliges. 2022a. InstantAvatar: Learning Avatars from Monocular Video in 60 Seconds. *arXiv* (2022).

Wei Jiang, Kwang Moo Yi, Golnoosh Samei, Oncel Tuzel, and Anurag Ranjan. 2022c. NeuMan: Neural Human Radiance Field from a Single Video. In *Proceedings of the European conference on computer vision (ECCV)*.

Haian Jin, Hanwen Jiang, Hao Tan, Kai Zhang, Sai Bi, Tianyuan Zhang, Fujun Luan, Noah Snavely, and Zexiang Xu. 2024. Lvsm: A large view synthesis model with minimal 3d inductive bias. *arXiv preprint arXiv:2410.17242* (2024).

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 2023. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics* 42, 4 (2023), 139–1.

Muhammed Kocabas, Jen-Hao Rick Chang, James Gabriel, Oncel Tuzel, and Anurag Ranjan. 2024. HUGS: Human Gaussian Splatting. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. https://arxiv.org/abs/2311.17910

Nikos Kolotouros, Thiemo Alldieck, Andrei Zanfir, Eduard Bazavan, Mihai Fieraru, and Cristian Sminchisescu. 2024. Dreamhuman: Animatable 3d avatars from text. *Advances in Neural Information Processing Systems* 36 (2024).

Youngjoong Kwon, Baole Fang, Yixing Lu, Haoye Dong, Cheng Zhang, Francisco Vicente Carrasco, Albert Mosella-Montoro, Jianjin Xu, Shingo Takagi, Daeil Kim, et al. 2024. Generalizable Human Gaussians for Sparse View Synthesis. *European Conference on Computer Vision* (2024).

Youngjoong Kwon, Dahun Kim, Duygu Ceylan, and Henry Fuchs. 2021. Neural human performer: Learning generalizable radiance fields for human performance rendering. *Advances in Neural Information Processing Systems* 34 (2021), 24741–24752.

Youngjoong Kwon, Dahun Kim, Duygu Ceylan, and Henry Fuchs. 2023. Neural image-based avatars: Generalizable radiance fields for human avatar modeling. *arXiv preprint arXiv:2304.04897* (2023).

Samuli Laine, Janne Hellsten, Tero Karras, Yeongho Seol, Jaakko Lehtinen, and Timo Aila. 2020. Modular Primitives for High-Performance Differentiable Rendering. *ACM Transactions on Graphics* 39, 6 (2020).

Jiahui Lei, Yufu Wang, Georgios Pavlakos, Lingjie Liu, and Kostas Daniilidis. 2024. Gart: Gaussian articulated template models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 19876–19887.

Ruilong Li, Julian Tanke, Minh Vo, Michael Zollhofer, Jurgen Gall, Angjoo Kanazawa, and Christoph Lassner. 2022. TAVA: Template-free animatable volumetric actors. *European Conference on Computer Vision (ECCV)*.

Yuanzhen Li, Fei Luo, and Chunxia Xiao. 2024a. Diffusion-FOF: Single-View Clothed Human Reconstruction via Diffusion-Based Fourier Occupancy Field. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 9525–9534.

Zhe Li, Zerong Zheng, Yuxiao Liu, Boyao Zhou, and Yebin Liu. 2023. PoseVocab: Learning Joint-structured Pose Embeddings for Human Avatar Modeling. In *ACM SIGGRAPH Conference Proceedings*.

Zhe Li, Zerong Zheng, Lizhen Wang, and Yebin Liu. 2024b. Animatable Gaussians: Learning Pose-dependent Gaussian Maps for High-fidelity Human Avatar Modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Hanwen Liang, Junli Cao, Vidit Goel, Guocheng Qian, Sergei Korolev, Demetri Terzopoulos, Konstantinos Plataniotis, Sergey Tulyakov, and Jian Ren. 2024a. Wonderland: Navigating 3D Scenes from a Single Image. *arXiv preprint arXiv:2412.12091* (2024).

Hanxue Liang, Jiawei Ren, Ashkan Mirzaei, Antonio Torralba, Ziwei Liu, Igor Gilitschenski, Sanja Fidler, Cengiz Oztireli, Huan Ling, Zan Gojcic, et al. 2024b. Feed-Forward Bullet-Time Reconstruction of Dynamic Scenes from Monocular Videos. *arXiv preprint arXiv:2412.03526* (2024).

Haotong Lin, Sida Peng, Zhen Xu, Yunzhi Yan, Qing Shuai, Hujun Bao, and Xiaowei Zhou. 2022. Efficient Neural Radiance Fields for Interactive Free-viewpoint Video. In *SIGGRAPH Asia Conference Proceedings*.

Siyou Lin, Zhe Li, Zhaoqi Su, Zerong Zheng, Hongwen Zhang, and Yebin Liu. 2024. LayGA: Layered Gaussian Avatars for Animatable Clothing Transfer. In *SIGGRAPH Conference Papers*.

Lingjie Liu, Marc Habermann, Viktor Rudnev, Kripasindhu Sarkar, Jiatao Gu, and Christian Theobalt. 2021. Neural Actor: Neural Free-view Synthesis of Human Actors with Pose Control. *ACM Trans. Graph.(ACM SIGGRAPH Asia)* (2021).

Minghua Liu, Chao Xu, Haian Jin, Linghao Chen, Mukund Varma T, Zexiang Xu, and Hao Su. 2024b. One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization. *Advances in Neural Information Processing Systems* 36 (2024).

Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. 2023. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF international conference on computer vision*. 9298–9309.

Wen Liu, Wenhan Luo Lin Ma Zhixin Piao, Min Jie, , and Shenghua Gao. 2019. Liquid Warping GAN: A Unified Framework for Human Motion Imitation, Appearance Transfer and Novel View Synthesis. In *The IEEE International Conference on Computer Vision (ICCV)*.

Xian Liu, Xiaohang Zhan, Jiaxiang Tang, Ying Shan, Gang Zeng, Dahua Lin, Xihui Liu, and Ziwei Liu. 2024c. Humangaussian: Text-driven 3d human generation with gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6646–6657.

Yebin Liu, Qionghai Dai, and Wenli Xu. 2009. A point-cloud-based multiview stereo algorithm for free-viewpoint video. *IEEE transactions on visualization and computer graphics* 16, 3 (2009), 407–418.

Yuxiao Liu, Zhe Li, Yebin Liu, and Haoqian Wang. 2024a. TexVocab: Texture Vocabulary-conditioned Human Avatars. *arXiv preprint arXiv:2404.00524* (2024).

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. 2015. SMPL: A Skinned Multi-Person Linear Model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)* 34, 6 (Oct. 2015), 248:1–248:16.

Yue Ma, Yingqing He, Xiaodong Cun, Xintao Wang, Ying Shan, Xiu Li, and Qifeng Chen. 2023. Follow Your Pose: Pose-Guided Text-to-Video Generation using Pose-Free Videos. *arXiv preprint arXiv:2304.01186* (2023).

Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. 2019. AMASS: Archive of Motion Capture as Surface Shapes. In *International Conference on Computer Vision*. 5442–5451.

Yifang Men, Yuan Yao, Miaomiao Cui, and Liefeng Bo. 2024. MIMO: Controllable Character Video Synthesis with Spatial Decomposed Modeling. *arXiv preprint arXiv:2409.16160*.

Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. 2019. Occupancy networks: Learning 3d reconstruction in function space. In *CVPR*. 4460–4470.

Marko Mihajlovic, Aayush Bansal, Michael Zollhoefer, Siyu Tang, and Shunsuke Saito. 2022. KeypointNeRF: Generalizing image-based volumetric avatars using relative spatial encoding of keypoints. In *European conference on computer vision*. Springer, 179–197.

B Mildenhall, PP Srinivasan, M Tancik, JT Barron, R Ramamoorthi, and R Ng. 2020. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*.

Gyeongsik Moon, Takaaki Shiratori, and Shunsuke Saito. 2024. Expressive Whole-Body 3D Gaussian Avatar. In *ECCV*.

Jiteng Mu, Shen Sang, Nuno Vasconcelos, and Xiaolong Wang. 2023. ActorsNeRF: Animatable Few-shot Human Rendering with Generalizable NeRFs. (2023), 18391–18401.

Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. 2023. DINOv2: Learning Robust Visual Features without Supervision.

Panwang Pan, Zhuo Su, Chenguo Lin, Zhen Fan, Yongjie Zhang, Zeming Li, Tingting Shen, Yadong Mu, and Yebin Liu. 2024. HumanSplat: Generalizable Single-Image Human Gaussian Splatting with Structure Priors. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. 2019. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 165–174.

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. 2019. Expressive Body Capture: 3D Hands, Face, and Body from a Single Image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 10975–10985.

Sida Peng, Junting Dong, Qianqian Wang, Shangzhan Zhang, Qing Shuai, Xiaowei Zhou, and Hujun Bao. 2021a. Animatable Neural Radiance Fields for Modeling Dynamic Human Bodies. In *ICCV*.

Songyou Peng, Michael Niemeyer, Lars Mescheder, Marc Pollefeys, and Andreas Geiger. 2020. Convolutional occupancy networks. In *ECCV*. Springer, 523–540.

Sida Peng, Zhen Xu, Junting Dong, Qianqian Wang, Shangzhan Zhang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. 2024. Animatable Implicit Neural Representations for Creating Realistic Avatars from Videos. *TPAMI* (2024).

Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. 2021b. Neural Body: Implicit Neural Representations with Structured Latent Codes for Novel View Synthesis of Dynamic Humans. In *CVPR*.

Julius Plücker. 1865. Xvii. on a new geometry of space. *Philosophical Transactions of the Royal Society of London* 155 (1865), 725–791.

Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. 2023. Dreamfusion: Text-to-3d using 2d diffusion. In *ICLR*.

Lorenza Prospero, Abdullah Hamdi, Joao F. Henriques, and Christian Rupprecht. 2024. GST: Precise 3D Human Body from a Single Image with Gaussian Splatting Transformers. arXiv:2409.04196 [cs.CV] https://arxiv.org/abs/2409.04196

Zhiyin Qian, Shaofei Wang, Marko Mihajlovic, Andreas Geiger, and Siyu Tang. 2024. 3DGS-Avatar: Animatable Avatars via Deformable 3D Gaussian Splatting. (2024).

Amit Raj, Michael Zollhoefer, Tomas Simon, Jason Saragih, Shunsuke Saito, James Hays, and Stephen Lombardi. 2021. Pva: Pixel-aligned volumetric avatars. *arXiv preprint arXiv:2101.02697* (2021).

René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. 2021. Vision Transformers for Dense Prediction. *ArXiv preprint* (2021).

Yurui Ren, Ge Li, Shan Liu, and Thomas H Li. 2020. Deep spatial transformation for pose-guided person image generation and animation. *IEEE Transactions on Image Processing* 29 (2020), 8622–8635.

Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. 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.

Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. 2020. 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*. 84–93.

Akash Sengupta, Thiemo Alldieck, Nikos Kolotouros, Enric Corona, Andrei Zanfir, and Cristian Sminchisescu. 2024. DiffHuman: Probabilistic Photorealistic 3D Reconstruction of Humans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 1439–1449.

Ruizhi Shao, Youxin Pang, Zerong Zheng, Jingxiang Sun, and Yebin Liu. 2024a. Human4DiT: 360-degree Human Video Generation with 4D Diffusion Transformer. *ACM Transactions on Graphics (TOG)* 43, 6 (2024).

Ruizhi Shao, Hongwen Zhang, He Zhang, Mingjia Chen, Yan-Pei Cao, Tao Yu, and Yebin Liu. 2022. Doublefield: Bridging the neural surface and radiance fields for high-fidelity human reconstruction and rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 15872–15882.

Zhijing Shao, Zhaolong Wang, Zhuang Li, Duotun Wang, Xiangru Lin, Yu Zhang, Mingming Fan, and Zeyu Wang. 2024b. SplattingAvatar: Realistic Real-Time Human Avatars with Mesh-Embedded Gaussian Splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Qiuhong Shen, Zike Wu, Xuanyu Yi, Pan Zhou, Hanwang Zhang, Shuicheng Yan, and Xinchao Wang. 2024. Gamba: Marry gaussian splatting with mamba for single view 3d reconstruction. *arXiv preprint arXiv:2403.18795* (2024).

Jisu Shin, Junmyeong Lee, Seongmin Lee, Min-Gyu Park, Ju-Mi Kang, Ju Hong Yoon, and Hae-Gon Jeon. 2025. CanonicalFusion: Generating Drivable 3D Human Avatars from Multiple Images. In *European Conference on Computer Vision*. Springer, 38–56.

Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. 2019a. Animating Arbitrary Objects via Deep Motion Transfer. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. 2019b. First Order Motion Model for Image Animation. In *Conference on Neural Information Processing Systems (NeurIPS)*.

Aliaksandr Siarohin, Oliver Woodford, Jian Ren, Menglei Chai, and Sergey Tulyakov. 2021. Motion Representations for Articulated Animation. In *CVPR*.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

Jonathan Starck and Adrian Hilton. 2007. Surface capture for performance-based animation. *IEEE computer graphics and applications* 27, 3 (2007), 21–31.

Shih-Yang Su, Timur Bagautdinov, and Helge Rhodin. 2023. Npc: Neural point characters from video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 14795–14805.

Guoxing Sun, Rishabh Dabral, Pascal Fua, Christian Theobalt, and Marc Habermann. 2024a. MetaCap: Meta-learning Priors from Multi-View Imagery for Sparse-view Human Performance Capture and Rendering. In *ECCV*.

Qingping Sun, Yanjun Wang, Ailing Zeng, Wanqi Yin, Chen Wei, Wenjia Wang, Haiyi Mei, Chi-Sing Leung, Ziwei Liu, Lei Yang, et al. 2024b. AiOS: All-in-One-Stage Expressive Human Pose and Shape Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1834–1843.

Jiaxiang Tang, Zhaoxi Chen, Xiaokang Chen, Tengfei Wang, Gang Zeng, and Ziwei Liu. 2025. Lgm: Large multi-view gaussian model for high-resolution 3d content creation. In *European Conference on Computer Vision*. Springer, 1–18.

Sicong Tang, Guangyuan Wang, Qing Ran, Lingzhi Li, Li Shen, and Ping Tan. 2023. High-resolution volumetric reconstruction for clothed humans. *ACM Transactions on Graphics* 42, 5 (2023), 1–15.

Felix Taubner, Ruihang Zhang, Mathieu Tuli, and David B. Lindell. 2024. CAP4D: Creating Animatable 4D Portrait Avatars with Morphable Multi-View Diffusion Models. (2024).

Ayush Tewari, Ohad Fried, Justus Thies, Vincent Sitzmann, Stephen Lombardi, Kalyan Sunkavalli, Ricardo Martin-Brualla, Tomas Simon, Jason Saragih, Matthias Nießner, et al. 2020. State of the art on neural rendering. In *Computer Graphics Forum*, Vol. 39. Wiley Online Library, 701–727.

Justus Thies, Michael Zollhöfer, and Matthias Nießner. 2019. Deferred neural rendering: Image synthesis using neural textures. *Acm Transactions on Graphics (TOG)* 38, 4 (2019), 1–12.

Hanzhang Tu, Ruizhi Shao, Xue Dong, Shunyuan Zheng, Hao Zhang, Lili Chen, Meili Wang, Wenyu Li, Siyan Ma, Shengping Zhang, et al. 2024. Tele-Aloha: A Telepresence System with Low-budget and High-authenticity Using Sparse RGB Cameras. In *ACM SIGGRAPH 2024 Conference Papers*. 1–12.

Gul Varol, Duygu Ceylan, Bryan Russell, Jimei Yang, Ersin Yumer, Ivan Laptev, and Cordelia Schmid. 2018. Bodynet: Volumetric inference of 3d human body shapes. In *Proceedings of the European conference on computer vision (ECCV)*. 20–36.

A Vaswani. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*.

Daniel Vlasic, Pieter Peers, Ilya Baran, Paul Debevec, Jovan Popović, Szymon Rusinkiewicz, and Wojciech Matusik. 2009. Dynamic shape capture using multi-view photometric stereo. In *ACM SIGGRAPH Asia*. 1–11.

Vikram Voleti, Chun-Han Yao, Mark Boss, Adam Letts, David Pankratz, Dmitrii Tochilkin, Christian Laforte, Robin Rombach, and Varun Jampani. 2024. SV3D: Novel Multi-view Synthesis and 3D Generation from a Single Image using Latent Video Diffusion. In *European Conference on Computer Vision (ECCV)*.

Peng Wang, Hao Tan, Sai Bi, Yinghao Xu, Fujun Luan, Kalyan Sunkavalli, Wenping Wang, Zexiang Xu, and Kai Zhang. 2023b. Pf-lrm: Pose-free large reconstruction model for joint pose and shape prediction. *arXiv preprint arXiv:2311.12024* (2023).

Ruicheng Wang, Sicheng Xu, Cassie Dai, Jianfeng Xiang, Yu Deng, Xin Tong, and Jiaolong Yang. 2024. MoGe: Unlocking Accurate Monocular Geometry Estimation for Open-Domain Images with Optimal Training Supervision. arXiv:2410.19115 [cs.CV] https://arxiv.org/abs/2410.19115

Shaofei Wang, Katja Schwarz, Andreas Geiger, and Siyu Tang. 2022. ARAH: Animatable Volume Rendering of Articulated Human SDFs. In *European Conference on Computer Vision*.

Tan Wang, Linjie Li, Kevin Lin, Yuanhao Zhai, Chung-Ching Lin, Zhengyuan Yang, Hanwang Zhang, Zicheng Liu, and Lijuan Wang. 2023a. Disco: Disentangled control for realistic human dance generation. *arXiv preprint arXiv:2307.00040* (2023).

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing* 13, 4 (2004), 600–612.

Xinyue Wei, Kai Zhang, Sai Bi, Hao Tan, Fujun Luan, Valentin Deschaintre, Kalyan Sunkavalli, Hao Su, and Zexiang Xu. 2024. Meshlrm: Large reconstruction model for high-quality mesh. *arXiv preprint arXiv:2404.12385* (2024).

Jing Wen, Xiaoming Zhao, Zhongzheng Ren, Alex Schwing, and Shenlong Wang. 2024. GoMAvatar: Efficient Animatable Human Modeling from Monocular Video Using Gaussians-on-Mesh. In *CVPR*.

Zhenzhen Weng, Jingyuan Liu, Hao Tan, Zhan Xu, Yang Zhou, Serena Yeung-Levy, and Jimei Yang. 2023. Template-Free Single-View 3D Human Digitalization with Diffusion-Guided LRM. *Preprint* (2023).

Chenglei Wu, Kiran Varanasi, Yebin Liu, Hans-Peter Seidel, and Christian Theobalt. 2011. Shading-based dynamic shape refinement from multi-view video under general illumination. In *2011 International Conference on Computer Vision*. IEEE, 1108–1115.

Junjin Xiao, Qing Zhang, Zhan Xu, and Wei-Shi Zheng. 2024. NECA: Neural Customizable Human Avatar. In *CVPR*.

Desai Xie, Sai Bi, Zhixin Shu, Kai Zhang, Zexiang Xu, Yi Zhou, Sören Pirk, Arie Kaufman, Xin Sun, and Hao Tan. 2024. LRM-Zero: Training Large Reconstruction Models with Synthesized Data. *arXiv preprint arXiv:2406.09371* (2024).

Yuliang Xiu, Jinlong Yang, Xu Cao, Dimitrios Tzionas, and Michael J. Black. 2023. ECON: Explicit Clothed humans Optimized via Normal integration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

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

Yuliang Xiu, Yufei Ye, Zhen Liu, Dimitris Tzionas, and Michael J. Black. 2024. PuzzleAvatar: Assembling 3D Avatars from Personal Albums. *ACM Trans. Graph.* 43, 6, Article 283 (Nov. 2024), 15 pages. https://doi.org/10.1145/3687771

Tianhan Xu, Yasuhiro Fujita, and Eiichi Matsumoto. 2022. Surface-Aligned Neural Radiance Fields for Controllable 3D Human Synthesis. In *CVPR*.

Yinghao Xu, Zifan Shi, Wang Yifan, Hansheng Chen, Ceyuan Yang, Sida Peng, Yujun Shen, and Gordon Wetzstein. 2024b. Grm: Large gaussian reconstruction model for efficient 3d reconstruction and generation. *arXiv preprint arXiv:2403.14621* (2024).

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

Yuanyou Xu, Zongxin Yang, and Yi Yang. 2023b. Seeavatar: Photorealistic text-to-3d avatar generation with constrained geometry and appearance. *arXiv preprint arXiv:2312.08889* (2023).

Zhen Xu, Sida Peng, Chen Geng, Linzhan Mou, Zihan Yan, Jiaming Sun, Hujun Bao, and Xiaowei Zhou. 2024a. Relightable and Animatable Neural Avatar from Sparse-View Video. In *CVPR*.

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

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. 2024b. Depth Anything V2. *arXiv:2406.09414* (2024).

Xihe Yang, Xingyu Chen, Daiheng Gao, Shaohui Wang, Xiaoguang Han, and Baoyuan Wang. 2024a. HAVE-FUN: Human Avatar Reconstruction from Few-Shot Unconstrained Images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 742–752.

Yifan Yang, Dong Liu, Shuhai Zhang, Zeshuai Deng, Zixiong Huang, and Mingkui Tan. 2024c. HiLo: Detailed and Robust 3D Clothed Human Reconstruction with High-and Low-Frequency Information of Parametric Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 10671–10681.

Xuanyu Yi, Zike Wu, Qiuhong Shen, Qingshan Xu, Pan Zhou, Joo-Hwee Lim, Shuicheng Yan, Xinchao Wang, and Hanwang Zhang. 2024. MVGamba: Unify 3D Content Generation as State Space Sequence Modeling. *arXiv preprint arXiv:2406.06367* (2024).

Jae Shin Yoon, Duygu Ceylan, Tuanfeng Y Wang, Jingwan Lu, Jimei Yang, Zhixin Shu, and Hyun Soo Park. 2022. Learning motion-dependent appearance for high-fidelity rendering of dynamic humans from a single camera. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 3407–3417.

Tao Yu, Zerong Zheng, Kaiwen Guo, Pengpeng Liu, Qionghai Dai, and Yebin Liu. 2021. Function4d: Real-time human volumetric capture from very sparse consumer rgbd sensors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 5746–5756.

Wing-Yin Yu, Lai-Man Po, Ray CC Cheung, Yuzhi Zhao, Yu Xue, and Kun Li. 2023b. Bidirectionally deformable motion modulation for video-based human pose transfer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 7502–7512.

Zhengming Yu, Wei Cheng, Xian Liu, Wayne Wu, and Kwan-Yee Lin. 2023a. MonoHuman: Animatable Human Neural Field from Monocular Video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 16943–16953.

Kai Zhang, Sai Bi, Hao Tan, Yuanbo Xiangli, Nanxuan Zhao, Kalyan Sunkavalli, and Zexiang Xu. 2024a. GS-LRM: Large Reconstruction Model for 3D Gaussian Splatting. *European Conference on Computer Vision* (2024).

Pengze Zhang, Lingxiao Yang, Jian-Huang Lai, and Xiaohua Xie. 2022. Exploring Dual-Task Correlation for Pose Guided Person Image Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 7713–7722.

Ruiqi Zhang, Jie Chen, and Qiang Wang. 2023. Explicifying neural implicit fields for efficient dynamic human avatar modeling via a neural explicit surface. In *Proceedings of the 31st ACM International Conference on Multimedia*. 1955–1963.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *CVPR*.

Yuang Zhang, Jiaxi Gu, Li-Wen Wang, Han Wang, Junqi Cheng, Yuefeng Zhu, and Fangyuan Zou. 2024b. MimicMotion: High-Quality Human Motion Video Generation with Confidence-aware Pose Guidance. *arXiv preprint arXiv:2406.19680* (2024).

Yuxiang Zhang, Zhe Li, Liang An, Mengcheng Li, Tao Yu, and Yebin Liu. 2021. Lightweight multi-person total motion capture using sparse multi-view cameras. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 5560–5569.

Zechuan Zhang, Li Sun, Zongxin Yang, Ling Chen, and Yi Yang. 2024c. Global-correlated 3d-decoupling transformer for clothed avatar reconstruction. *Advances in Neural Information Processing Systems* 36 (2024).

Zechuan Zhang, Zongxin Yang, and Yi Yang. 2024d. 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 (CVPR)*. 9936–9947.

Jian Zhao and Hui Zhang. 2022. Thin-plate spline motion model for image animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3657–3666.

Shunyuan Zheng, Boyao Zhou, Ruizhi Shao, Boning Liu, Shengping Zhang, Liqiang Nie, and Yebin Liu. 2024. Gps-gaussian: Generalizable pixel-wise 3d gaussian splatting for real-time human novel view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 19680–19690.

Yang Zheng, Ruizhi Shao, Yuxiang Zhang, Tao Yu, Zerong Zheng, Qionghai Dai, and Yebin Liu. 2021a. Deepmulticap: Performance capture of multiple characters using sparse multiview cameras. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 6239–6249.

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

Zerong Zheng, Tao Yu, Yixuan Wei, Qionghai Dai, and Yebin Liu. 2019. Deephuman: 3d human reconstruction from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 7739–7749.

Boyao Zhou, Shunyuan Zheng, Hanzhang Tu, Ruizhi Shao, Boning Liu, Shengping Zhang, Liqiang Nie, and Yebin Liu. 2024b. GPS-Gaussian+: Generalizable Pixel-wise 3D Gaussian Splatting for Real-Time Human-Scene Rendering from Sparse Views. *arXiv preprint arXiv:2411.11363* (2024).

Tiansong Zhou, Jing Huang, Tao Yu, Ruizhi Shao, and Kun Li. 2024a. HDhuman: High-Quality Human Novel-View Rendering From Sparse Views. *IEEE Transactions on Visualization and Computer Graphics* 30, 8 (2024), 5328–5338.

Shenhao Zhu, Junming Leo Chen, Zuozhuo Dai, Yinghui Xu, Xun Cao, Yao Yao, Hao Zhu, and Siyu Zhu. 2024. Champ: Controllable and Consistent Human Image Animation with 3D Parametric Guidance. arXiv:2403.14781 [cs.CV]

Wojciech Zielonka, Timur Bagautdinov, Shunsuke Saito, Michael Zollhöfer, Justus Thies, and Javier Romero. 2025. Drivable 3D Gaussian Avatars. (March 2025).

Chen Ziwen, Hao Tan, Kai Zhang, Sai Bi, Fujun Luan, Yicong Hong, Li Fuxin, and Zexiang Xu. 2024. Long-LRM: Long-sequence Large Reconstruction Model for Wide-coverage Gaussian Splats. *arXiv preprint 2410.12781* (2024).