PGAHum: Prior-Guided Geometry and Appearance Learning for High-Fidelity Animatable Human Reconstruction

Hao Wang
Jilin University, Changchun, China
whao22@mails.jlu.edu.cn

Qingshan Xu*
Nanyang Technological University, Singapore
qingshan.xu@ntu.edu.sg

Hongyuan Chen
Jilin University, Changchun, China
chenhy5521@mails.jlu.edu.cn

Tieru Wu
Jilin University, Changchun, China
wutr@jlu.edu.cn

Rui Ma*
Jilin University, Changchun, China
Engineering Research Center of Knowledge-Driven Human-Machine Intelligence, MOE, China
ruim@jlu.edu.cn

Abstract

Recent techniques on implicit geometry representation learning and neural rendering have shown promising results for 3D clothed human reconstruction from sparse video inputs. However, it is still challenging to reconstruct detailed surface geometry and even more difficult to synthesize photorealistic novel views with animated human poses. In this work, we introduce PGAHum, a prior-guided geometry and appearance learning framework for high-fidelity animatable human reconstruction. We thoroughly exploit 3D human priors in three key modules of PGAHum to achieve high-quality geometry reconstruction with intricate details and photorealistic view synthesis on unseen poses. First, a prior-based implicit geometry representation of 3D human, which contains a delta SDF predicted by a tri-plane network and a base SDF derived from the prior SMPL model, is proposed to model the surface details and the body shape in a disentangled manner. Second, we introduce a novel prior-guided sampling strategy that fully leverages the prior information of the human pose and body to sample the query points within or near the body surface. By avoiding unnecessary learning in the empty 3D space, the neural rendering can recover more appearance details. Last, we propose a novel iterative backward deformation strategy to progressively find the correspondence for the query point in observation space. A skinning weights prediction model is learned based on the prior provided by the SMPL model to achieve the iterative backward LBS deformation. Extensive quantitative and qualitative comparisons on various datasets are conducted and the results demonstrate the superiority of our framework. Ablation studies also verify the effectiveness of each scheme for geometry and appearance learning. Code will be released.

Keywords: ***Avatar, Geometry, Appearance, Prior-Guided.***

1. Introduction

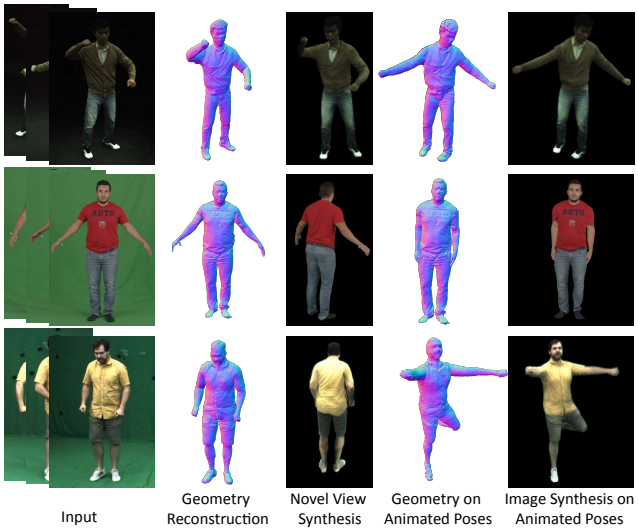

Figure 1. Given sparse input videos, our PGAHum can reconstruct high-fidelity animatable avatar with fine-grained geometry and appearance details on various datasets, e.g., ZJU-Mocap [36] (top), PeopleSnapshot [1] (middle) and MonoCap [35] (bottom).

* Corresponding authors.

The digitization of the human body is crucial for various applications such as gaming, film, mixed reality, remote interaction and the metaverse. In the industries, high-fidelity human body reconstruction typically requires acquiring data by multi-camera systems in well-equipped studios and building pre-captured templates with the assistance from skilled artists. These requirements prohibit the conventional applications for consumers, such as personalized avatars used in AR/VR, body measurements, virtual try-on, etc.

Human body reconstruction has been a popular and important research topic in recent years. Recent methods can reconstruct the clothed 3D human body from sparse videos, making the 3D human avatar acquisition more convenient and flexible. One series of clothing-aware body reconstruction methods use explicit mesh or fixed-resolution truncated signed distance fields (TSDFs) to represent the geometric shape of humans, while the textures are represented by vertex colors or UV maps. Based on existing statistical human models [25, 32, 30, 15], these methods [26, 33, 16] initialize 3D human avatar in some geometry representations of fixed-resolution, and then learn geometry offsets or displacements to enhance surface details. While the spatiotemporal consistency of 3D human avatar can be obtained by building their human representation on top of models such as SMPL [25], the capability to express details is limited since the geometry offset learning is bound to a fixed number of geometric vertices. Moreover, the appearance of the reconstructed human is not realistic enough.

With the success of neural implicit representations [31, 27, 28], recent reconstruction methods combine neural implicit representations with neural rendering techniques and achieve promising results in geometry or appearance learning of 3D human. Some methods [12, 52, 10, 14] aim to learn implicit representations of the human body which are capable of handling various topologies, adapting to different characters and clothing, as well as supporting animations with different poses. Meanwhile, a series of neural rendering methods [18, 23, 13, 24, 51, 47, 46, 49] mainly focus on synthesizing photorealistic novel viewpoint images, while not explicitly performing optimization for the geometry. However, high-quality dense multi-view data may still be needed as the supervision for the neural rendering process. On the other hand, there exist several works [36, 42, 34, 35] simultaneously performs human geometry reconstruction and appearance learning. While the reconstruction results capture the overall body shape, these methods struggle with finer geometry details, such as clothing wrinkles. The main reason is the inherent non-rigid motion characteristics of humans and clothing may cause challenges for the correspondence searching during the implicit representation learning. As the result of the sub-optimal geometry learning, the quality of view synthesis may also be affected to some extent.

In this work, we aim for high-fidelity animatable human reconstruction, i.e., high-quality geometry reconstruction with intricate human body and clothing details, as well as photorealistic image synthesis on novel views and unseen poses. To this end, we propose PGAHum, a novel framework which extensively utilizes 3D human priors to reconstruct high-fidelity animatable clothed humans from sparse videos. Specifically, we first define a *prior-based implicit geometry representation* to effectively learn the underlying body shape and surface details in a disentangled manner. On top of the base signed distance field (SDF) derived from the SMPL model fitted to the subject, a tri-plane network is learned to predict a delta SDF layer for encoding the surface details. With the strong prior from the SMPL model, the network can focus on learning the fine-grained geometry details rather than the overall body shape. Second, to further confine the geometry and appearance learning on the fine details of the human body, a *prior-guided sampling* scheme which fully leverages the prior information of the human pose and body is proposed. Comparing to existing methods [28, 41] which mainly use the stratified sampling to sample query points in the full space, we first compute the ray-body intersection and only sample the points within or near the body surface, so that unnecessary sampling in empty 3D space can be avoided and the learned neural radiance filed (NeRF) can recover more appearance details. Last, we propose an *iterative backward deformation* strategy to warp query points in the observation space to the canonical space in a progressive manner, so that both the geometry and appearance of the unified human model can be better optimized. Notably, we learn a backward skinning weights prediction model which takes the transformed query points to generate skinning weights for the iterative backward Linear Blend Skinning (LBS) deformation. Comparing to the forward deformation [42], our iterative backward deformation does not need to solve the root-finding problem to find the correspondances for the query points. Also, it can effectively finds the appropriate correspondences through multiple backward deformation, alleviating the errors for the one-step backward deformation. Compared to analytical root-finding methods, such as Newton's method, our iterative neural update does not rely on explicit Jacobian computation or well-conditioned derivatives. In practice, analytical solvers can be sensitive to initialization and local nonlinearity, and may become unstable or fail to converge when the deformation field is complex or noisy. Compared to one-step inverse networks, which attempt to directly predict the inverse deformation in a single forward pass, our method allows progressive refinement of the solution. This iterative process enables the network to correct intermediate errors and handle harder cases that are difficult to invert accurately in one step, which we empirically find to be beneficial, espe-

cially for surface reconstruction. These results suggest that the iterative backward deformation strikes a better balance between expressiveness and robustness in practice.

By combining above schemes, our PGAHum takes a solid step further to high-fidelity animatable human reconstruction, obtaining avatar with appealing geometry and appearance details for novel pose or view synthesis (see Figure 1). Extensive quantitative and qualitative comparisons are conducted and the results demonstrate the superiority of our framework. Ablation studies also verify the effectiveness of each scheme for geometry and appearance learning. In summary, our contributions are as follows:

- We propose PGAHum, a novel framework which extensively exploits the 3D human priors for high-fidelity animatable human reconstruction. Our results show more fine-grained geometry and appearances details than existing SOTA methods.
- We disentangle the 3D clothed human with a prior-based implicit geometry representation. Such fully implicit representation not only supports disentangled surface detail modeling, but also fits well for following NeRF learning.
- We leverage the priors of the human model and propose a novel prior-guided sampling strategy to avoid the unnecessary learning in empty 3D space, so that the neural rendering can recover more appearance details.
- We develop an iterative backward deformation strategy which reduces the computation cost for the forward deformation and the error for one-step backward deformation. To achieve the iterative backward LBS deformation, a skinning weights prediction model is learned based on the prior provided by the SMPL model.

2. Related Works

Geometry-conditioned human reconstruction. Based on existing statistical models [25, 32, 30, 15], some reconstruction methods [3, 29, 19] learn human geometry from image inputs. To represent clothed human bodies, some methods [26, 33, 2] learn geometry offsets on top of the base geometry model. However, these representations mainly support compact clothing types and the reconstructed geometry is often coarse. For higher-fidelity reconstructions, some approaches [48, 7, 8] use character-specific templates to assist in pose tracking and performance capture. However, these methods rely on explicit geometry templates or fixed-resolution representations, limiting fine-grained geometry and appearance reconstruction.

Implicit representation learning for human reconstruction. The PIFu series [38, 39, 10] utilize pixel-aligned feature fusion to predict high-precision depths for estimating the implicit function of human body. PaMIR [52] combines the features derived from a parametric body model and the image to learn an implicit function for human body. ARCH [12] and ARCH++ [9] combine pixel-aligned features from semantic-aware geometric encoders and appearance encoders to learn a joint-space human body model. These methods mainly train networks to predict implicit values with the enhancement of appearance features. Though certain level of geometry details can be recovered, they cannot learn fine-grained geometry details by only learning one global implicit representation. In contrast, our method uses a neural implicit representation to learn fine-grained details based on prior geometry. Note that, unlike approaches that learn high-frequency details in static scenes [44], our method targets dynamic scenes. Moreover, our base geometry relies on strong priors from SMPL rather than learning a neural implicit field from scratch. This allows our geometry detail layer to focus on capturing high-frequency details, without the need to first learn the overall shape as in [44].

Neural rendering for human novel view synthesis. Neural rendering [28] has been explored by recent methods for synthesizing novel views of human [16, 18, 46, 49, 5, 17]. The key of these methods is to learn a deformation field or model to warp or transform the query point in the observation space to the canonical space, allowing the radiance field to be optimized across different poses. However, these methods mainly focus on the appearance model learning and synthesizing realistic novel view images, and overlook the geometry model learning. Without the guidance of human geometry, the images synthesized at unseen poses may contain artifacts or lack details.

Geometry and appearance learning for human reconstruction. Instead of only learning geometry or appearance with NeRF, some recent methods simultaneously learn both of them for human reconstruction. Neural Body [36] anchors a set of structured latent codes to the vertices of the SMPL mesh to learn the implicit geometry and radiance fields from sparse videos. A-NeRF [40] estimates the 3D skeleton structure of a human body and performs skeleton-relative encoding for geometry and appearance learning. The skeleton pose has also been used to define a pose-driven deformation field for optimizing geometry and color [34, 35]. ARAH [42] integrates ray-surface intersection and correspondence search to identify transformed query points in the canonical space, followed by SDF-based volume rendering to predict SDF and color values. Recently, some methods [11, 37, 22, 45] have leveraged 3DGS [20] to reconstruct human appearance from videos. While these methods can synthesize impressive novel view images, they struggle to recover accurate geometry due to the sparse and unstructured point cloud representation. DEGAS focuses on modeling detailed full-body Gaussian avatars with an emphasis on appearance representation, whereas our work

primarily targets accurate geometric reconstruction. GAvatar is a generative approach for animatable 3D Gaussian avatars and is not specifically designed for human reconstruction, making direct comparison less appropriate. TaoAvatar and MoGA are more closely related to our problem setting and were proposed recently.

In contrast, we extensively use human priors to enhance both geometry representation and neural rendering strategies for more fine-grained geometry and appearance learning.

3. Method

Our method aims to learn high-fidelity animatable human avatars from videos. The pipeline is shown in Figure 2. Specifically, we first define a prior-based implicit geometry representation based on the SMPL model. It models the global body shape and local surface details in a fully implicit and disentangled manner (Sec. 3.1). To learn this representation, given an input view with human pose, we design a prior-guided sampling scheme to sample 3D spatial points in observation space (Sec. 3.2). By leveraging the SMPL model, our sampling focuses more on spatial points within or near the human body. Then, we present an iterative backward deformation strategy to progressively warp the sampled points to the canonical space (Sec. 3.3). Finally, the warped points in the canonical space are used for volume rendering to compose the final rendered images (Sec. 3.4). By imposing a series of loss functions on the rendered images and the geometry representation (Sec. 3.5), our method learns a personalized animatable avatar that exhibits fine-grained surface geometry details and can be photorealisticlly rendered from different viewpoints and under novel poses.

3.1. Prior-Based Implicit Representation

Some methods [42, 14] have represent and learn the overall geometry of human avatars using one global SDF. However, we observe that these methods cannot reconstruct fine geometry details. Hence, we propose a prior-based implicit geometry representation to effectively learn the underling body shape and surface details in a disentangled manner.

Specifically, our proposed representation $\mathscr{S}$ is defined in the canonical space with the star-pose. It consists of base geometry prior field $\mathscr{S}_{base}$ and geometry detail layer $\mathscr{S}_{delta}$, which effectively combines the advantages of the global body consistency derived from the prior SMPL model and the local detail modeling by the tri-plane representation [4]. The base geometry prior field $\mathscr{S}_{base}$ is represented by the SDF volume derived from the SMPL model fitted to each subject. Then, we use a neural network F_{ϕ_s} to learn the delta SDF $\mathscr{S}_{delta}$ for $\mathscr{S}_{base}$, where ϕ_s represents learnable parameters. More concretely, the F_{ϕ_s} contains a tri-plane representation $T = (T_{xy}, T_{yz}, T_{xz})$ and a shallow Multi-Layer Perceptron (MLP), where T_{xy}, T_{yz}, and T_{xz} are three learnable feature planes that are orthogonal to each other and form a 3D cube of size L^3 centered at $(0,0,0)$. To learn the delta SDF $S_{delta}(\mathbf{x})$ at position $\mathbf{x}$, we project $\mathbf{x}$ onto each of the three planes and query the corresponding features on each plane by bilinear interpolation. By concatenating these features and passing them into the MLP, we obtain $S_{delta}(\mathbf{x})$. Finally, we combine both the S_{base} and S_{delta} to yield a complete SDF for any position $\mathbf{x}$ in canonical space as:

$$\mathscr{S}(\mathbf{x}) = \mathscr{S}_{base}(\mathbf{x}) + \mathscr{S}_{delta}(\mathbf{x}). \tag{1}$$

Based on this representation, when querying the SDF value for a point $\mathbf{x}_{cnl}$ in canonical space, we first sample the $\mathscr{S}_{base}(\mathbf{x}_{cnl})$ from SDF volume, and then predict the corresponding delta SDF value $\mathscr{S}_{delta}(\mathbf{x}_{cnl})$ via F_{ϕ_s} to compute the final SDF value. Thanks to the strong SDF prior from $\mathscr{S}_{base}$, the network can focus on learning fine-grained geometry details on clothed human rather than the overall body shape. In addition to the SDF value, a feature vector μ, also output by F_{ϕ_s}, is passed to a color branch to predict view-dependent RGB values. The color branch is represented by a shallow MLP, denoted as F_{ϕ_c}.

3.2. Prior-Guided Sampling

Following [41, 42], we leverage the volume rendering technique to render images and supervise the SDF learning. In the vanilla volume rendering process [28, 41], a stratified sampling strategy is used to sample spatial points on a ray, which introduces many unnecessary spatial point queries, hindering the efficiency and effectiveness of volume rendering on learning details. Considering the simple human body topology, accessible poses and easy-to-use SMPL model, we propose a prior-guided sampling strategy, which fully leverages the prior information of human poses and shapes.

Specifically, for a ray emitted from the camera center, we determine whether it will intersect with the human body shape which is represented by the SMPL model estimated at current viewpoint. We compute all intersection points with the human body and keep the depth value z along the ray. Since the camera center must be outside the human model, there must be an even number of intersection points, corresponding to a set of depth values $\{z_0, z_1, \ldots, z_{2n-1}\}$ in an ascending order, where the number of intersection points is $2n$. Each pair of intersection points with depth of (z_{2i}, z_{2i+1}) forms an intersection interval with length l_i, where $i = 0, 1, \ldots, n-1$.

To make our sampled points within and near the body surface, we extend the intersection interval to $(z_{2i} - 0.1l_i, z_{2i+1} + 0.1l_i)$. According to the length of each intersection interval, the sampling number of each interval is computed as $Num_i = \frac{l_i}{\sum_{i=0}^{n-1} l_i} * N_{spl}$, where N_{spl} denotes

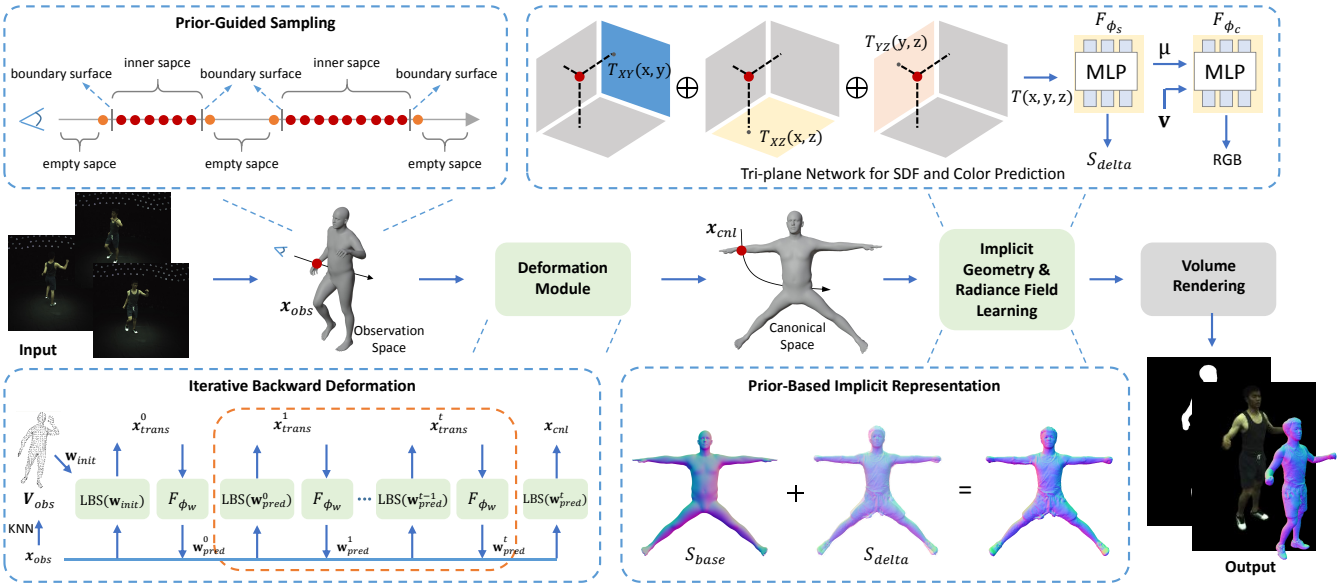

Figure 2. Overview of our pipeline. Given an input view with estimated human pose, we first utilize prior-guided sampling to sample points inside and around the prior SMPL human body. For a sampled point $\mathbf{x}_{obs}$, we deform it to the corresponding point $\mathbf{x}_{cnl}$ in a canonical space through the iterative backward deformation. We then learn a prior-based implicit geometry representation which combines the prior SDF volume $\mathcal{S}_{base}$ derived from SMPL with $\mathcal{S}_{delta}$ predicted by a tri-plane network F_{ϕ_s} for modeling the human body with surface details. Finally, volume rendering is performed to render images, normal maps and subject mask for the loss computation.

Algorithm 1 Iterative Backward Deformation

Input: $\mathbf{x}_{obs}, \mathbf{V}_{obs}, \mathbf{B}, \tilde{\mathbf{W}}, F_{\phi_w}$

Initialize: $m \leftarrow 0$

Find the K nearest neighbor vertices to point $\mathbf{x}_{obs}$ in $\mathbf{V}_{obs}$

Compute $\mathbf{w}_{init}$ from $\tilde{\mathbf{W}}$ based on Equation 3

$skinning\_weights \leftarrow \mathbf{w}_{init}$

repeat

  $\mathbf{x}^m_{trans} \leftarrow LBS(\mathbf{x}_{obs}, skinning\_weights, \mathbf{B}^{-1})$

  $skinning\_weights \leftarrow F_{\phi_w}(\mathbf{x}^m_{trans})$

  $m \leftarrow m+1$

until $m \geq t$

$\mathbf{x}_{cnl} \leftarrow LBS(\mathbf{x}_{obs}, skinning\_weights, \mathbf{B}^{-1})$

Output: Deformed point $\mathbf{x}_{cnl}$ in canonical space

the total number of sampling points required along a ray. Within each interval, Num_i points will be uniformly sampled. If the ray does not intersect with the human body, we will uniformly sample points in the whole sampling span.

3.3. Iterative Backward Deformation

It is important to calculate accurate correspondences between the observation space and canonical space such that the appearance information in the observation space can be used to optimize the geometry and appearance representation in the canonical space. Backward deformation refers to transforming points from observation space to the canonical space, while forward deformation refers to the inverse progress. We follow [17, 35] and adopt the backward deformation for more efficient deformation model learning. However, the backward deformation in one step is hard to guarantee accuracy. To address this issue, inspired by iterative root finding algorithm of forward deformation in [42, 14], we propose an iterative backward deformation module so that the points in the observation space are progressively transformed to the appropriate positions in the canonical space through multiple iterations. In this way, the non-rigid and pose-dependent deformation is also naturally handled. This module will learn a skinning model F_{ϕ_w}, a 4-layer MLP with trainable parameters ϕ_w, in canonical space to predict skinning weights for LBS deformation.

Specifically, we first find the K nearest neighbor vertices in the mesh $\mathbf{V}_{obs}$ derived from the posed SMPL for each sampled point, and calculate the initial skinning weight for each sampled point as:

$$\mathbf{w}_{init} = \sum_{k=1}^{K} \tilde{\mathbf{W}}(k) \cdot \frac{1}{\beta_k}, \quad \beta_k = \frac{d(k)}{\sum_{k=1}^{K} d(k)}, \tag{2}$$

where the $\tilde{\mathbf{W}}(k)$ denotes the prior skinning weights of the k-th vertex, and $d(k)$ denotes the distance between the current sampled point and the k-th nearest vertex. Afterwards, we deform the sampled point $\mathbf{x}_{obs}$ to canonical space by back-

ward LBS deformation:

$$\begin{aligned}\mathbf{x}^0_{trans} &= LBS(\mathbf{x}_{obs}, \mathbf{w}_{init}, \mathbf{B}^{-1}) \\ &= (\sum_{i=1}^{J} \mathbf{w}_{init}(i) \cdot \mathbf{B}_i^{-1}) \mathbf{x}_{obs},\end{aligned} \tag{3}$$

where $\mathbf{x}^0_{trans}$ denotes the point in canonical space deformed from $\mathbf{x}_{obs}$, $\mathbf{B} = \{\mathbf{B}_i\}^{24}$ denotes the bone-based transformation matrix which deforms from the canonical pose to the observation pose, and $\mathbf{w}_{init}(i)$ is the weight corresponding to bone $\mathbf{B}_i$.

We then query skinning model F_{ϕ_w} with $\mathbf{x}^0_{trans}$ to get predicted skinning weights $\mathbf{w}^0_{pred} = F_{\phi_w}(\mathbf{x}^0_{trans})$ for $\mathbf{x}_{obs}$, and transform $\mathbf{x}_{obs}$ to $\mathbf{x}^1_{trans}$ using Eq. (3) by $LBS(\mathbf{x}_{obs}, \mathbf{w}^0_{pred}, \mathbf{B}^{-1})$. However, the $\mathbf{x}^1_{trans}$ maybe not be the accurate correspondence of $\mathbf{x}_{obs}$ since the $\mathbf{w}_{init}$ is computed from the prior SMPL model which only has an estimated base geometry. Hence, we additionally iterative t times to get a more appropriate corresponding point in canonical space. Finally, we obtain the corresponding point $\mathbf{x}_{cnl}$ in canonical space for point $\mathbf{x}_{obs}$.

For a more convenient understanding, the whole iterative backward deformation process is summarized in Algorithm 1. We set $K = 10$ and $t = 3$ in our current implementation.

3.4. Animatable Human Volume Rendering

For the volume rendering, we follow [43] and incorporate the SDF values during the radiance field learning process. Specifically, with the corresponding canonical points $\{\mathbf{x}^i_{cnl}\}^{N_{spl}}$ for sampled points $\{\mathbf{x}^i_{obs}\}^{N_{spl}}$ along the ray, we query $\mathscr{S}_{base}$ and $\mathscr{S}_{delta}$ to compute final SDF values by Eq. (1). Besides, we query radiance field F_{ϕ_c} to get radiance (both radiance field and SDF are defined in canonical space). Finally, we accumulate the queried radiance $\{c_i\}^{N_{spl}}$ along the ray to get pixel color C:

$$C = \sum_{i=1}^{N_{spl}} \prod_{j<i} (1-\alpha_j) c_i, \quad \alpha_i = 1 - \exp(-\sigma_i \delta_i), \tag{4}$$

$$\sigma_i = s * (\Phi_s(\mathscr{S}(\mathbf{x}^i_{cnl})) - 1) \nabla \mathscr{S}(\mathbf{x}^i_{cnl}) \cdot \mathbf{v}, \tag{5}$$

where $\delta_i = ||\mathbf{x}^{i+1}_{obs} - \mathbf{x}^i_{obs}||_1$ is the distance between two adjacent sampled points, $\Phi_s(\cdot)$ is the sigmoid function, $s \in \mathbb{R}$ is a learnable scale parameter, and $\mathbf{v}$ is the view direction.

To further ensure the volume rendering to focus on the foreground region, we also synthesize a foreground mask based on the human shape represented by the SDF field. Different from previous works [35] which use the learned global SDF to obtain the mask, we directly use the prior base SDF for mask rendering to enhance the stability and facilitate the convergence of network learning. In addition, we slightly inflate the mask with a certain distance to enable the learning of surface details. Formally, the foreground mask is obtained by:

$$M(\mathbf{x}) = \begin{cases} 0 & \text{if } \mathscr{S}_{base}(\mathbf{x}) - \tau > 0 \\ 1 & \text{if } \mathscr{S}_{base}(\mathbf{x}) - \tau \leq 0 \end{cases}, \tag{6}$$

where τ (empirically set to 0.05) is a distance threshold for considering a point that is outside from the shape surface represented by $\mathscr{S}_{base}$ as foreground.

3.5. Optimization

The final loss, including several photometric losses in observation space and multiple regularizers in canonical space, is defined as:

$$\begin{aligned}\mathscr{L}_{total} =& \lambda_1 \mathscr{L}_{rgb} + \lambda_2 \mathscr{L}_{lpips} + \lambda_3 \mathscr{L}_{nssim} + \\ & \lambda_4 \mathscr{L}_{eikonal} + \lambda_5 \mathscr{L}_{skinning} + \lambda_6 \mathscr{L}_{mask}.\end{aligned} \tag{7}$$

$\mathscr{L}_{rgb}$ is reconstruction loss by comparing the ground truth color $\tilde{C}(r)$ and rendered color. $C(r)$:

$$\mathscr{L}_{rgb} = \sum_{r \in \mathscr{R}} ||\tilde{C}(r) - C(r)||_1, \tag{8}$$

where $\mathscr{R}$ denotes the set of rays. $\mathscr{L}_{nssim}$ is similarity loss derived from Structural Similarity Index Measure (SSIM) and defined as:

$$\mathscr{L}_{nssim} = 1 - SSIM(C(r), \tilde{C}(r)). \tag{9}$$

$\mathscr{L}_{skinning}$ is l1 loss for regularizing predicted skinning weights and prior ground truth.

$$\mathscr{L}_{skinning} = ||\mathbf{w}^t_{pred} - \mathbf{w}_{init}||_1, \tag{10}$$

where $\mathbf{w}_{init}$ is sampled from $\tilde{\mathbf{W}}$ by K nearest neighbors, and $\mathbf{w}^t_{pred}$ is skinning weights queried from F_{ϕ_w} at the iteration step t. $\mathscr{L}_{eikonal}$ is Eikonal regularization [6], and $\mathscr{L}_{mask}$ is used to supervise the SDF with mask [35]. The coefficients λ_1 to λ_6 are the weights of each loss function. We set $\lambda_1 = 10$, $\lambda_2 = \lambda_3 = \lambda_5 = \lambda_6 = 1$, $\lambda_4 = 0.1$. After the first 50K iterations, λ_5 is set to 0.

4. Experiments

4.1. Experimental Settings

Datasets. Following [42], we use the **ZJU-MoCap** [36] dataset as our primary testbed and maintain the same training/test splits as them. We evaluate our method on three tasks: novel view synthesis on training poses (NVS), unseen pose synthesis (Unseen), and geometry reconstruction (Recon). We also conduct unseen pose synthesis experiments on the **PeopleSnapshot** [1] dataset, which contains monocular video of human subjects rotating in front of a camera. We follow the evaluation protocol in [17].

Table 1. Quantitative results for novel view synthesis and unseen pose synthesis tasks. We compare PSNR (↑), LPIPS* = LIPIPS $\times 10$ (↓) metrics on ZJU-MoCap dataset for novel view synthesis and unseen pose synthesis tasks, denoted as NVS and Unseen respectively. We bold the values with the **best** metric value and underline the second-best ones.

NVS/Unseen	377		387		386		393		394	
Method	PSNR	LPIPS*	PSNR	LPIPS*	PSNR	LPIPS*	PSNR	LPIPS*	PSNR	LPIPS*
Ani-NeRF	24.2/22.6	1.24/1.53	25.4/23.1	1.31/1.45	25.6/25.5	1.99/1.87	26.1/23.8	1.51/1.55	27.5/24.1	1.42/1.71
ARAH	27.8/25.5	0.71/0.93	**27.0**/24.2	0.79/0.99	**29.2**/27.0	1.05/1.27	**27.7**/24.4	0.93/1.04	28.9/25.2	0.84/1.11
InstantNVR	26.1/24.0	0.94/1.20	24.5/23.5	1.29/1.41	28.3/26.9	1.24/1.55	25.9/23.8	1.17/1.32	26.8/24.3	1.13/1.38
GoMAvatar	26.1/**25.9**	0.65/0.71	25.6/**25.5**	0.69/**0.68**	27.6/27.7	1.02/1.01	25.4/**25.5**	0.82/**0.74**	26.9/**26.0**	0.75/**0.80**
Ours	**27.9**/25.6	**0.47/0.70**	26.9/25.0	**0.66**/0.73	**29.2/27.8**	**0.83/0.92**	**27.7**/25.4	**0.74**/0.79	**29.0/26.0**	**0.63**/0.86

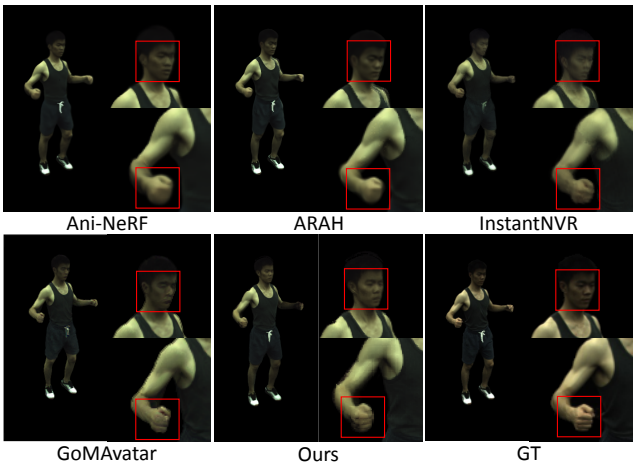

Figure 3. Qualitative results on ZJU-MoCap dataset for novel view synthesis on training poses.

Table 2. Generalization to unseen poses on PeopleSnapshot.

Methods	male-3-casual		male-4-casual	
Method	PSNR	LPIPS*	PSNR	LPIPS*
NB	24.94	0.326	24.71	0.423
InstantAvatar	29.65	**0.192**	27.97	0.346
GoMAvatar	**33.91**	0.226	**30.56**	0.344
Ours	30.47	0.270	28.86	**0.281**
Methods	female-3-casual		female-4-casual	
Method	PSNR	LPIPS*	PSNR	LPIPS*
NB	23.87	0.361	24.37	0.382
InstantAvatar	27.90	**0.249**	28.92	**0.180**
GoMAvatar	29.48	0.359	23.21	0.790
Ours	**30.86**	0.358	**32.49**	0.218

Metrics. To assess the synthesized images in the NVS/Unseen tasks, we use the PSNR and LPIPS [50] metrics to compare rendered and corresponding ground-truth (GT) images. For geometry reconstruction, we use Chamfer Distance (CD) to evaluate our method and baselines on the training poses. As it is difficult to obtain GT geometry for dynamic humans, we follow ARAH [42] to generate pseudo-GT geometry for ZJU-MoCap by Pet-NeuS [43].

Implementation details. Our method is implemented with the PyTorch framework. The Adam [21] is adopted for the training. The learning rate starts from $5e^{-4}$ and decays exponentially to $5e^{-5}$ along the optimization. The training is conducted on 4 A40 GPUs. The overall training and inference time are similar to ARAH. Please refer to the supplementary for more details.

4.2. Comparisons

Novel view synthesis. In Table 1 and Figure 3, we report the results of our method on novel view synthesis on ZJU-MoCap [36] dataset. In Figure 3, it can be observed that our method produces synthetic images with clearer human hands and heads, to the extent that even the number of fingers can be discerned. This is very difficult for other methods. As shown in Table 1, our approach demonstrates the best or comparable performance in terms of PSNR and LPIPS metrics in most cases compared to state-of-the-art methods.

Generalization to unseen poses. On the ZJU-MoCap dataset, we typically train using the first 300 frames, while the frames after the 300th are used as unseen poses for testing generalization ability. In Table 1, it can be observed that our method exhibits superior generalization performance on unseen poses. This can be attributed to the use of prior knowledge of the human body to achieve accurate deformation (Sec. 3.3) and sampling (Sec. 3.2). In addition, the results in Table 2 demonstrate that our method achieves promising generalization performance on the PeopleSnapshot [2] dataset as well. It should be noted for unseen poses, the results of GoMAvatar [45] and ours are comparable, while our method can consistently produce competing results since we have explicitly learned the human geometry based on prior guidance. In contrast, GoMAvatar may still produce inferior results for certain cases, e.g., the female-4-casual. Please refer to the supplementary for more qualitative results.

Geometry reconstruction. In Figures 4, we qualitatively compare our method with others [42, 45] on ZJU-MoCap dataset. It can be observed that our method the reconstructs geometry with fine surface details due to the design of prior-based implicit representation (Sec. 3.1). In Table

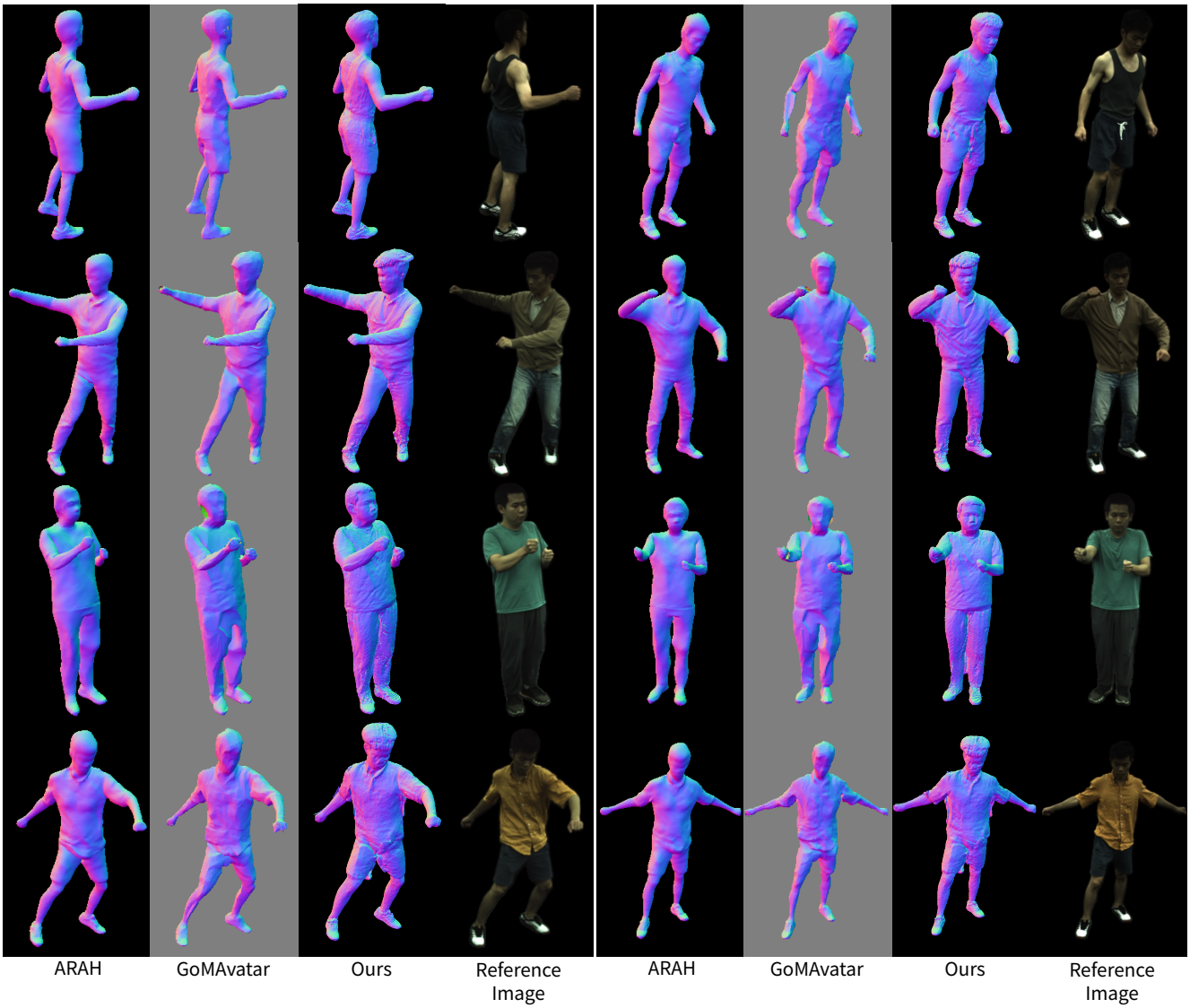

Figure 4. Qualitative results on ZJU-MoCap dataset for geometry reconstruction.

Table 3. Quantative comparison for geometry reconstruction in training poses. We compare Chamfer Distance (↓) metric on ZJU-MoCap dataset.

Subject	377	386	387	393	394	mean
NB	1.4417	1.3705	1.0814	1.4898	1.2034	1.3174
ARAH	0.6846	0.2032	0.3168	0.8284	1.0341	0.6134
GoMAvatar	0.6561	0.1791	0.3397	0.8688	1.0256	0.6138
Ours	**0.6366**	**0.1758**	**0.3051**	**0.8009**	**1.0048**	**0.5863**

3, we present a quantitative comparison of our method with NB [36], ARAH [42] and GoMAvatar [45] on the reconstructed geometries from the ZJU-MoCap dataset. It can be observed that our method consistently achieves the best results in terms of the Chamfer Distance (CD).

4.3. Ablation Study

We conduct ablation studies on subject 377 of ZJU-MoCap dataset to assess our proposed modules. In Table 4, clear improvements can be observed when the components are sequentially added. The last line shows the percentage increased between the full model and the baseline.

Prior-based implicit representation. Our proposed method employs a combination of prior base geometry field $\mathscr{S}_{base}$ and geometry detail layer $\mathscr{S}_{delta}$ to represent an overall human SDF. To validate its effectiveness, we conduct experiments by removing $\mathscr{S}_{base}$ module and solely utilizing a tri-plane network to learn the overall SDF. As shown in Table 4, when $\mathscr{S}_{base}$ is dropped (row A+B w.r.t. A+B+C), there is a certain degree of performance decline in PSNR and LPIPS for both novel view synthesis and generalization

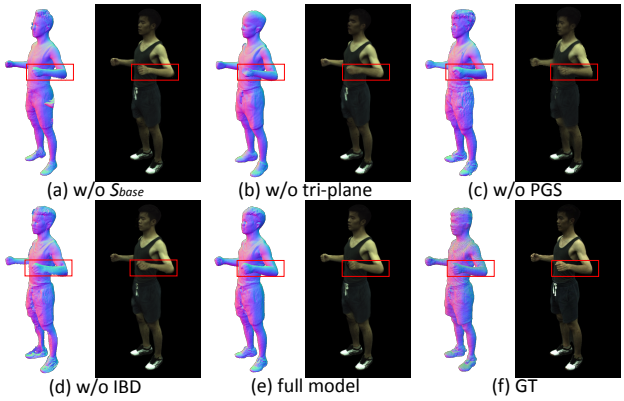

Figure 5. Qualitative results for ablation study.

Table 4. Ablation study. A: baseline+iterative backward deformation, B: tri-plane, C: $\mathscr{S}_{base}$, D: prior-guided sampling, full model: A+B+C+D. Improvement is w.r.t. baseline.

Component	NVS		Unseen		Recon
Metric	PSNR	LPIPS	PSNR	LPIPS	CD
Baseline	25.92	0.080	25.23	0.078	0.7001
A	26.58	0.071	25.42	0.070	0.6912
A+B	27.05	0.065	25.62	0.063	0.6819
A+B+C	27.10	0.055	25.54	0.062	0.6570
Full Model	**27.95**	**0.047**	**25.84**	**0.062**	**0.6366**
Improvement	7.832%	41.250%	2.418%	20.513%	9.070%

to unseen poses. In Figure 5(a), we can also observe that without this module, some ghosting artifacts appear at the end of the left elbow and on the pants in the synthesized images. We attribute this to the lack of underlying models to support the neural network in learning sufficiently robust and fine-grained representations for certain areas which are textureless or near the boundary.

In our method, we use a tri-plane representation to model $\mathscr{S}_{delta}$. Similarly, we conduct an ablation experiment by replacing it with a naive MLP network. In the row A of Table 4 and Figure 5(b), it can be seen that, due to the powerful expressive capability of the tri-plane representation, our method is able to achieve improved geometry reconstruction and novel view synthesis.

Prior-guided sampling. To validate the effectiveness of our proposed prior-guided sampling strategy, we conduct experiments by replacing this module with naive uniform sampling. As observed in Table 4, upon removing this module D, there is a certain degree of performance decline in PSNR and LPIPS for both NVS and Unseen tasks. In Figure 5(c), we notice obvious blurring in the arms and facial regions of the synthesized images. This is because the uniform sampling distracts the model's focus on learning the representative regions of the human body and introduces unnecessary points in empty space, which may lead to degraded performance on geometry and appearance learning.

Table 5. Analysis of the Number of Iterations of Iterative Backward Deformation. LIPIS*=LIPIPS $\times$ 10.

Times	1		2		3		4		5	
Subjects	PSNR	LIPIS*	PSNR	LIPIS*	PSNR	LIPIS*	PSNR	LIPIS*	PSNR	LIPIS*
377	27.92	0.44	28.04	0.42	**28.05**	**0.41**	**28.05**	**0.41**	28.01	0.42
386	29.36	0.79	29.34	0.78	**29.37**	**0.77**	29.34	0.78	29.36	0.78
387	26.88	0.62	26.96	0.62	**27.04**	**0.61**	27.01	**0.61**	27.03	0.60
393	27.58	**0.76**	27.62	0.77	**27.66**	0.77	27.62	0.77	27.64	**0.76**
394	28.91	0.66	28.97	0.65	**29.03**	**0.64**	29.0	0.65	**29.03**	0.65

Iterative backward deformation. To investigate the effectiveness of our proposed iterative backward deformation, we conduct an ablation experiment by replacing this module with one-step backward skinning deformation module. As observed in the row baseline (w.r.t. A) of Table 4, without this module, there is a decrease of performance in PSNR and LPIPS for all tasks. In Figure 5(d), we notice blurry patches and noise in the hand and pants regions of the subject. These results verify the iterative backward deformation module can effectively transform points from the observation space to the canonical space. By performing multiple deformations, the transformed points can gradually converge to the actual corresponding canonical points, alleviating the errors by performing the single-step deformation. We also added experiments and analysis about the number of iterations of Iterative Backward Deformation module, as shown in the Table 5. We will include this table in the supplementary materials to help understand. As shown in Table 5, after three iterations, we find more accurate corresponding points between the observation and canonical space, and the metrics improve. Then, they gradually converge without further enhancement.

5. Conclusion

We introduce PGAHum, a novel framework that takes a solid step further to high-fidelity animatable human reconstruction from multi-view or monocular videos. To achieve fine-grained geometry and appearance learning, we effectively use human priors in three novel modules. First, the prior-based implicit geometry representation combines the advantages of the global body consistency derived from the prior SMPL model and the powerful local detail modeling by the tri-plane representation, allowing the network to focus on learning fine surface details of clothed human. Second, the prior-guided sampling leverages prior human pose and shape information to constrain sampling around human body, which encourages the volume rendering to learn more appearance details. Last, the iterative backward deformation warps the query points using the skinning weight model learned based on the initial SMPL weights, ensuring efficient and accurate space transformation to facilitate the optimization in the canonical space. The experimental results demonstrate that our method achieves superior fine-grained geometry and appearance reconstruction compared to the current state-of-the-art methods.

Limitation and future work. Despite our method performing well on geometric reconstruction, novel view synthesis, and generalization to unseen poses, there are still some limitations. 1) Similar to [35], our method does not perform well on reconstruction of loose clothing. This is because our sampling strategy collects sampled points around the human body as much as possible, which inevitably misses sampling some points that should be on loose clothing. We have tried inflating the prior SMPL model, but experimental results show that the effect is not obvious. Although our method does not perform well in reconstructing very loose clothing such as skirts and dresses, as shown in Figure 6 Column 8 of Supplementary, it can effectively reconstruct characters wearing moderately loose clothing. And reconstructing loose clothing such as skirts and dresses will be our future work. 2) Our method requires a relatively long training time of about 50 hours, which brings an expensive cost for rapid iteration to optimize the framework design. We think there are two reasons for this: one is that the sampling method we implemented is not computationally efficient enough, and the second is that the complex tri-plane representation we adopted brings great computational complexity. We acknowledge that NeRF-based rendering is generally slower than recent 3D Gaussian Splatting (3DGS) methods in terms of pure rendering speed. Our primary goal, however, is not real-time rendering but high-quality geometry and appearance reconstruction under complex deformations. In our framework, NeRF is chosen for its flexibility in representing view-dependent appearance and fine geometric details. Moreover, the proposed prior-guided sampling significantly reduces the number of evaluated samples, improving practical efficiency. 3) Similar to [28], our method is trained per scene, which takes a lot of time to produce a dynamic human body model. How to generalize this method to loose clothing while improving training efficiency will be explored in our future work.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (No. 62572212), Science and Technology Development Plan of Jilin Province (No. 20260203049SF) and the Fundamental Research Funds for the Central Universities.

References

[1] Alldieck, T.; Magnor, M.; Xu, W.; Theobalt, C.; and Pons-Moll, G. 2018a. Detailed human avatars from monocular video. In *2018 International Conference on 3D Vision (3DV)*, 98–109. IEEE. 1, 6

[2] Alldieck, T.; Magnor, M.; Xu, W.; Theobalt, C.; and Pons-Moll, G. 2018b. Video based reconstruction of 3d people models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 8387–8397. 3, 7

[3] Bogo, F.; Kanazawa, A.; Lassner, C.; Gehler, P.; Romero, J.; and Black, M. J. 2016. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, 561–578. Springer. 3

[4] Chan, E. R.; Lin, C. Z.; Chan, M. A.; Nagano, K.; Pan, B.; De Mello, S.; Gallo, O.; Guibas, L. J.; Tremblay, J.; Khamis, S.; et al. 2022. Efficient geometry-aware 3d generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 16123–16133. 4

[5] Geng, C.; Peng, S.; Xu, Z.; Bao, H.; and Zhou, X. 2023. Learning Neural Volumetric Representations of Dynamic Humans in Minutes. In *CVPR*. 3

[6] Gropp, A.; Yariv, L.; Haim, N.; Atzmon, M.; and Lipman, Y. 2020. Implicit Geometric Regularization for Learning Shapes. In *Proceedings of Machine Learning and Systems 2020*, 3569–3579. 6

[7] Habermann, M.; Xu, W.; Zollhoefer, M.; Pons-Moll, G.; and Theobalt, C. 2019. Livecap: Real-time human performance capture from monocular video. *ACM Transactions On Graphics (TOG)*, 38(2): 1–17. 3

[8] Habermann, M.; Xu, W.; Zollhofer, M.; Pons-Moll, G.; and Theobalt, C. 2020. Deepcap: Monocular human performance capture using weak supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5052–5063. 3

[9] He, T.; Xu, Y.; Saito, S.; Soatto, S.; and Tung, T. 2021. Arch++: Animation-ready clothed human reconstruction revisited. In *Proceedings of the IEEE/CVF international conference on computer vision*, 11046–11056. 3

[10] Hong, Y.; Zhang, J.; Jiang, B.; Guo, Y.; Liu, L.; and Bao, H. 2021. Stereopifu: Depth aware clothed human digitization via stereo vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 535–545. 2, 3

[11] Hu, S.; and Liu, Z. 2023. GauHuman: Articulated Gaussian Splatting from Monocular Human Videos. *arXiv preprint arXiv:*. 3

[12] Huang, Z.; Xu, Y.; Lassner, C.; Li, H.; and Tung, T. 2020. Arch: Animatable reconstruction of clothed humans. In *Proceedings of the IEEE/CVF Conference*

on Computer Vision and Pattern Recognition, 3093–3102. 2, 3

[13] Jayasundara, V.; Agrawal, A.; Heron, N.; Shrivastava, A.; and Davis, L. S. 2023. FlexNeRF: Photorealistic free-viewpoint rendering of moving humans from sparse views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21118–21127. 2

[14] Jiang, B.; Hong, Y.; Bao, H.; and Zhang, J. 2022a. SelfRecon: Self Reconstruction Your Digital Avatar from Monocular Video. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2, 4, 5

[15] Jiang, B.; Zhang, J.; Cai, J.; and Zheng, J. 2020a. Disentangled human body embedding based on deep hierarchical neural network. *IEEE transactions on visualization and computer graphics*, 26(8): 2560–2575. 2, 3

[16] Jiang, B.; Zhang, J.; Hong, Y.; Luo, J.; Liu, L.; and Bao, H. 2020b. Bcnet: Learning body and cloth shape from a single image. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16*, 18–35. Springer. 2, 3

[17] Jiang, T.; Chen, X.; Song, J.; and Hilliges, O. 2023. Instantavatar: Learning avatars from monocular video in 60 seconds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16922–16932. 3, 5, 6

[18] Jiang, W.; Yi, K. M.; Samei, G.; Tuzel, O.; and Ranjan, A. 2022b. Neuman: Neural human radiance field from a single video. In *European Conference on Computer Vision*, 402–418. Springer. 2, 3

[19] Kanazawa, A.; Black, M. J.; Jacobs, D. W.; and Malik, J. 2018. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7122–7131. 3

[20] Kerbl, B.; Kopanas, G.; Leimkühler, T.; and Drettakis, G. 2023. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Transactions on Graphics*, 42(4). 3

[21] Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*. 7

[22] Kocabas, M.; Chang, R.; Gabriel, J.; Tuzel, O.; and Ranjan, A. 2023. HUGS: Human Gaussian Splats. 3

[23] Li, Z.; Zheng, Z.; Liu, Y.; Zhou, B.; and Liu, Y. 2023. PoseVocab: Learning Joint-structured Pose Embeddings for Human Avatar Modeling. *arXiv preprint arXiv:2304.13006*. 2

[24] Liao, T.; Zhang, X.; Xiu, Y.; Yi, H.; Liu, X.; Qi, G.-J.; Zhang, Y.; Wang, X.; Zhu, X.; and Lei, Z. 2023. High-Fidelity Clothed Avatar Reconstruction from a Single Image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8662–8672. 2

[25] Loper, M.; Mahmood, N.; Romero, J.; Pons-Moll, G.; and Black, M. J. 2015. SMPL: A Skinned Multi-Person Linear Model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6): 248:1–248:16. 2, 3

[26] Ma, Q.; Yang, J.; Ranjan, A.; Pujades, S.; Pons-Moll, G.; Tang, S.; and Black, M. J. 2020. Learning to dress 3d people in generative clothing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6469–6478. 2, 3

[27] Mescheder, L.; Oechsle, M.; Niemeyer, M.; Nowozin, S.; and Geiger, A. 2019. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4460–4470. 2

[28] Mildenhall, B.; Srinivasan, P. P.; Tancik, M.; Barron, J. T.; Ramamoorthi, R.; and Ng, R. 2020. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. In *ECCV*. 2, 3, 4, 10

[29] Omran, M.; Lassner, C.; Pons-Moll, G.; Gehler, P.; and Schiele, B. 2018. Neural body fitting: Unifying deep learning and model based human pose and shape estimation. In *2018 international conference on 3D vision (3DV)*, 484–494. IEEE. 3

[30] Osman, A. A.; Bolkart, T.; and Black, M. J. 2020. Star: Sparse trained articulated human body regressor. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, 598–613. Springer. 2, 3

[31] Park, J. J.; Florence, P.; Straub, J.; Newcombe, R.; and Lovegrove, S. 2019. DeepSDF: Learning Continuous Signed Distance Functions for Shape Representation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2

[32] Pavlakos, G.; Choutas, V.; Ghorbani, N.; Bolkart, T.; Osman, A. A.; Tzionas, D.; and Black, M. J. 2019. Expressive body capture: 3d hands, face, and body from

a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10975–10985. 2, 3

[33] Pavlakos, G.; Zhu, L.; Zhou, X.; and Daniilidis, K. 2018. Learning to estimate 3D human pose and shape from a single color image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 459–468. 2, 3

[34] Peng, S.; Dong, J.; Wang, Q.; Zhang, S.; Shuai, Q.; Zhou, X.; and Bao, H. 2021a. Animatable neural radiance fields for modeling dynamic human bodies. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 14314–14323. 2, 3

[35] Peng, S.; Xu, Z.; Dong, J.; Wang, Q.; Zhang, S.; Shuai, Q.; Bao, H.; and Zhou, X. 2024. Animatable Implicit Neural Representations for Creating Realistic Avatars from Videos. *TPAMI*. 1, 2, 3, 5, 6, 10

[36] Peng, S.; Zhang, Y.; Xu, Y.; Wang, Q.; Shuai, Q.; Bao, H.; and Zhou, X. 2021b. Neural Body: Implicit Neural Representations with Structured Latent Codes for Novel View Synthesis of Dynamic Humans. In *CVPR*. 1, 2, 3, 6, 7, 8

[37] Qian, Z.; Wang, S.; Mihajlovic, M.; Geiger, A.; and Tang, S. 2024. 3DGS-Avatar: Animatable Avatars via Deformable 3D Gaussian Splatting. 3

[38] Saito, S.; Huang, Z.; Natsume, R.; Morishima, S.; Kanazawa, A.; and Li, H. 2019. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2304–2314. 3

[39] Saito, S.; Simon, T.; Saragih, J.; and Joo, H. 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. 3

[40] Su, S.-Y.; Yu, F.; Zollhöfer, M.; and Rhodin, H. 2021. A-nerf: Articulated neural radiance fields for learning human shape, appearance, and pose. *Advances in Neural Information Processing Systems*, 34: 12278–12291. 3

[41] Wang, P.; Liu, L.; Liu, Y.; Theobalt, C.; Komura, T.; and Wang, W. 2021. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *arXiv preprint arXiv:2106.10689*. 2, 4

[42] Wang, S.; Schwarz, K.; Geiger, A.; and Tang, S. 2022. ARAH: Animatable Volume Rendering of Articulated Human SDFs. In *European Conference on Computer Vision*. 2, 3, 4, 5, 6, 7, 8

[43] Wang, Y.; Skorokhodov, I.; and Wonka, P. 2023. Petneus: Positional encoding tri-planes for neural surfaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12598–12607. 6, 7

[44] Wang, Y.; Skorokhodov, I.; and Wonka, P. 2024. HF-NeuS: improved surface reconstruction using high-frequency details. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22. Red Hook, NY, USA: Curran Associates Inc. ISBN 9781713871088. 3

[45] Wen, J.; Zhao, X.; Ren, Z.; Schwing, A.; and Wang, S. 2024. GoMAvatar: Efficient Animatable Human Modeling from Monocular Video Using Gaussians-on-Mesh. In *CVPR*. 3, 7, 8

[46] Weng, C.-Y.; Curless, B.; Srinivasan, P. P.; Barron, J. T.; and Kemelmacher-Shlizerman, I. 2022. Humannerf: Free-viewpoint rendering of moving people from monocular video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern Recognition*, 16210–16220. 2, 3

[47] Xu, H.; Alldieck, T.; and Sminchisescu, C. 2021. H-nerf: Neural radiance fields for rendering and temporal reconstruction of humans in motion. *Advances in Neural Information Processing Systems*, 34: 14955–14966. 2

[48] Xu, W.; Chatterjee, A.; Zollhöfer, M.; Rhodin, H.; Mehta, D.; Seidel, H.-P.; and Theobalt, C. 2018. Monoperfcap: Human performance capture from monocular video. *ACM Transactions on Graphics (ToG)*, 37(2): 1–15. 3

[49] Yu, Z.; Cheng, W.; Liu, X.; Wu, W.; and Lin, K.-Y. 2023. MonoHuman: Animatable Human Neural Field from Monocular Video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16943–16953. 2, 3

[50] Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 586–595. 7

[51] Zheng, Z.; Huang, H.; Yu, T.; Zhang, H.; Guo, Y.; and Liu, Y. 2022. Structured local radiance fields for human avatar modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15893–15903. 2

[52] Zheng, Z.; Yu, T.; Liu, Y.; and Dai, Q. 2021. Pamir: Parametric model-conditioned implicit representation for image-based human reconstruction. *IEEE transactions on pattern analysis and machine intelligence*, 44(6): 3170–3184. 2, 3