ARAH: Animatable Volume Rendering of Articulated Human SDFs

Shaofei Wang[1](✉), Katja Schwarz[2,3], Andreas Geiger[2,3], and Siyu Tang[1]

[1] ETH Zürich, Zürich, Switzerland
shaofei.wang@inf.ethz.ch
[2] Max Planck Institute for Intelligent Systems, Tübingen, Germany
[3] University of Tübingen, Tübingen, Germany

Abstract. Combining human body models with differentiable rendering has recently enabled animatable avatars of clothed humans from sparse sets of multi-view RGB videos. While state-of-the-art approaches achieve a realistic appearance with neural radiance fields (NeRF), the inferred geometry often lacks detail due to missing geometric constraints. Further, animating avatars in out-of-distribution poses is not yet possible because the mapping from observation space to canonical space does not generalize faithfully to unseen poses. In this work, we address these shortcomings and propose a model to create animatable clothed human avatars with detailed geometry that generalize well to out-of-distribution poses. To achieve detailed geometry, we combine an articulated implicit surface representation with volume rendering. For generalization, we propose a novel joint root-finding algorithm for simultaneous ray-surface intersection search and correspondence search. Our algorithm enables efficient point sampling and accurate point canonicalization while generalizing well to unseen poses. We demonstrate that our proposed pipeline can generate clothed avatars with high-quality pose-dependent geometry and appearance from a sparse set of multi-view RGB videos. Our method achieves state-of-the-art performance on geometry and appearance reconstruction while creating animatable avatars that generalize well to out-of-distribution poses beyond the small number of training poses.

Keywords: 3D computer vision · Clothed human modeling · Cloth modeling · Neural rendering · Neural implicit functions

1 Introduction

Reconstruction and animation of clothed human avatars is a rising topic in computer vision research. It is of particular interest for various applications in AR/VR and the future metaverse. Various sensors can be used to create clothed human avatars, ranging from 4D scanners over depth sensors to simple RGB

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-19824-3_1.

S. Avidan et al. (Eds.): ECCV 2022, LNCS 13692, pp. 1–19, 2022.
https://doi.org/10.1007/978-3-031-19824-3_1

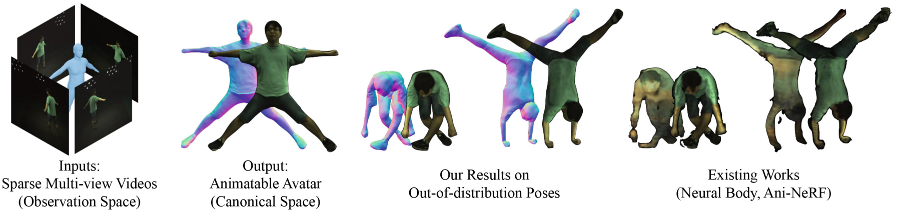


Fig. 1. Detailed Geometry and Generalization to Extreme Poses. Given sparse multi-view videos with SMPL fittings and foreground masks, our approach synthesizes animatable clothed avatars with realistic pose-dependent geometry and appearance. While existing works, *e.g.* Neural Body [56] and Ani-NeRF [54], struggle with generalizing to unseen poses, our approach enables avatars that can be animated in extreme out-of-distribution poses.

cameras. Among these data sources, RGB videos are by far the most accessible and user-friendly choice. However, they also provide the least supervision, making this setup the most challenging for the reconstruction and animation of clothed humans.

Traditional works in clothed human modeling use explicit mesh [1,2,6,7,17,18,29,33,52,63,68,78,83] or truncated signed distance fields (TSDFs) of fixed grid resolution [34,35,66,76,81] to represent the geometry of humans. Textures are often represented by vertex colors or UV-maps. With the recent success of neural implicit representations, significant progress has been made towards modeling articulated clothed humans. PIFu [60] and PIFuHD [61] are among the first works that propose to model clothed humans as continuous neural implicit functions. ARCH [24] extends this idea and develops animatable clothed human avatars from monocular images. However, this line of works does not handle dynamic pose-dependent cloth deformations. Further, they require ground-truth geometry for training. Such ground-truth data is expensive to acquire, limiting the generalization of these methods.

Another line of works removes the need for ground-truth geometry by utilizing differentiable neural rendering. These methods aim to reconstruct humans from a sparse set of multi-view videos with only image supervision. Many of them use NeRF [46] as the underlying representation and achieve impressive visual fidelity on novel view synthesis tasks. However, there are two fundamental drawbacks of these existing approaches: (1) the NeRF-based representation lacks proper geometric regularization, leading to inaccurate geometry. This is particularly detrimental in a sparse multi-view setup and often results in artifacts in the form of erroneous color blobs under novel views or poses. (2) Existing approaches condition their NeRF networks [56] or canonicalization networks [54] on inputs in observation space. Thus, they cannot generalize to unseen out-of-distribution poses.

In this work, we address these two major drawbacks of existing approaches. (1) We improve geometry by building an articulated signed-distance-field (SDF) representation for clothed human bodies to better capture the geometry of clothed humans and improve the rendering quality. (2) In order to render the

SDF, we develop an efficient joint root-finding algorithm for the conversion from observation space to canonical space. Specifically, we represent clothed human avatars as a combination of a forward linear blend skinning (LBS) network, an implicit SDF network, and a color network, all defined in canonical space and do not condition on inputs in observation space. Given these networks and camera rays in observation space, we apply our novel joint root-finding algorithm that can efficiently find the iso-surface points in observation space and their correspondences in canonical space. This enables us to perform efficient sampling on camera rays around the iso-surface. All network modules can be trained with a photometric loss in image space and regularization losses in canonical space.

We validate our approach on the ZJU-MoCap [56] and the H36M [25] dataset. Our approach generalizes well to unseen poses, enabling robust animation of clothed avatars even under out-of-distribution poses where existing works fail, as shown in Fig. 1. We achieve significant improvements over state-of-the-arts for novel pose synthesis and geometry reconstruction, while also outperforming state-of-the-arts in the novel view synthesis task on training poses. Code and data are available at https://neuralbodies.github.io/arah/.

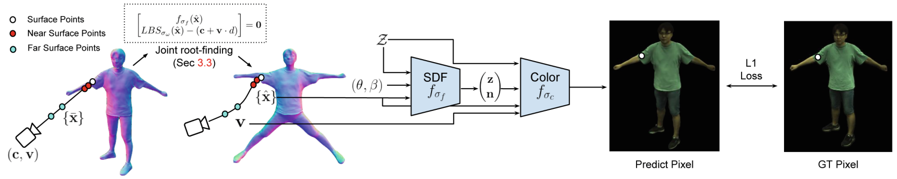
2 Related Works

Clothed Human Modeling with Explicit Representations: Many explicit mesh-based approaches represent cloth deformations as deformation layers [1,2, 6–8] added to minimally clothed parametric human body models [5,20,27,37,50, 53,75]. Such approaches enjoy compatibility with parametric human body models but have difficulties in modeling large garment deformations. Other mesh-based approaches model garments as separate meshes [17,18,29,33,52,63,68,78, 83] in order to represent more detailed and physically plausible cloth deformations. However, such methods often require accurate 3D-surface registration, synthetic 3D data or dense multi-view images for training and the garment meshes need to be pre-defined for each cloth type. More recently, point-cloud-based explicit methods [38,39,82] also showed promising results in modeling clothed humans. However, they still require explicit 3D or depth supervision for training, while our goal is to train using sparse multi-view RGB supervision alone.

Clothed Humans as Implicit Functions: Neural implicit functions [12,41,42, 51,57] have been used to model clothed humans from various sensor inputs including monocular images [21,22,24,31,59–61,65,73,86], multi-view videos [28,36, 48,54,56,74], sparse point clouds [6,13,15,70,71,87], or 3D meshes [10,11,14,44, 45,62,67]. Among the image-based methods, [4,22,24] obtain animatable reconstructions of clothed humans from a single image. However, they do not model pose-dependent cloth deformations and require ground-truth geometry for training. [28] learns generalizable NeRF models for human performance capture and only requires multi-view images as supervision. But it needs images as inputs for synthesizing novel poses. [36,48,54,56,74] take multi-view videos as inputs and do not need ground-truth geometry during training. These methods generate personalized per-subject avatars and only need 2D supervision. Our approach follows this line of work and also learns a personalized avatar for each subject.

Neural Rendering of Animatable Clothed Humans: Differentiable neural rendering has been extended to model animatable human bodies by a number of recent works [48,54,56,58,65,74]. Neural Body [56] proposes to diffuse latent per-vertex codes associated with SMPL meshes in observation space and condition NeRF [46] on such latent codes. However, the conditional inputs of Neural Body are in the observation space. Therefore, it does not generalize well to out-of-distribution poses. Several recent works [48,54,65] propose to model the radiance field in canonical space and use a pre-defined or learned backward mapping to map query points from observation space to this canonical space. A-NeRF [65] uses a deterministic backward mapping defined by piecewise rigid bone transformations. This mapping is very coarse and the model has to use a complicated bone-relative embedding to compensate for that. Ani-NeRF [54] trains a backward LBS network that does not generalize well to out-of-distribution poses, even when fine-tuned with a cycle consistency loss for its backward LBS network for each test pose. Further, all aforementioned methods utilize a volumetric radiance representation and hence suffer from noisy geometry [49,69,79,80]. In contrast to these works, we improve geometry by combining an implicit surface representation with volume rendering and improve pose generalization via iterative root-finding. H-NeRF [74] achieves large improvements in geometric reconstruction by co-training SDF and NeRF networks. However, code and models of H-NeRF are not publicly available. Furthermore, H-NeRF's canonicalization process relies on imGHUM [3] to predict an accurate signed distance in *observation space*. Therefore, imGHUM needs to be trained on a large corpus of posed human scans and it is unclear whether the learned signed distance fields generalize to out-of-distribution poses beyond the training set. In contrast, our approach does not need to be trained on any posed scans and it can generalize to extreme out-of-distribution poses.

Concurrent Works: Several concurrent works extend NeRF-based articulated models to improve novel view synthesis, geometry reconstruction, or animation quality [9,23,26,30,43,55,64,72,77,85]. [85] proposes to jointly learn forward blending weights, a canonical occupancy network, and a canonical color network using differentiable surface rendering for head-avatars. In contrast to human heads, human bodies show much more articulation. Abrupt changes in depth also occur more frequently when rendering human bodies, which is difficult to capture with surface rendering [69]. Furthermore, [85] uses the secant method to find surface points. For each secant step, this needs to solve a root-finding problem from scratch. Instead, we use volume rendering of SDFs and formulate the surface-finding task of articulated SDFs as a joint root-finding problem that only needs to be solved once per ray. We remark that [26] proposes to formulate surface-finding and correspondence search as a joint root-finding problem to tackle geometry reconstruction from photometric and mask losses. However, they use pre-defined skinning fields and surface rendering. They also require estimated normals from PIFuHD [61] while our approach achieves detailed geometry reconstructions without such supervision.


Surface Points
Near Surface Points
Far Surface Points
$\begin{bmatrix} f_{\sigma_f}(\hat{\mathbf{x}}) \\ LBS_{\sigma_\omega}(\hat{\mathbf{x}}) - (\mathbf{c} + \mathbf{v} \cdot d) \end{bmatrix} = \mathbf{0}$
Joint root-finding
(Sec 3.3)
$\{\bar{\mathbf{x}}\}$
$(\mathbf{c}, \mathbf{v})$
$\{\hat{\mathbf{x}}\}$
$\mathbf{v}$
$\mathcal{Z}$
(θ, β)
SDF f_{σ_f}
$\begin{pmatrix} \mathbf{z} \\ \mathbf{n} \end{pmatrix}$
Color f_{σ_c}
L1 Loss
Predict Pixel
GT Pixel

(a) Root-finding and point sampling (b) Canonicalization of sampled points (c) SDF-based volume rendering (d) Photometric loss

Fig. 2. Overview of Our Pipeline. (a) Given a ray $(\mathbf{c}, \mathbf{v})$ with camera center $\mathbf{c}$ and ray direction $\mathbf{v}$ in observation space, we jointly search for its intersection with the SDF iso-surface and the correspondence of the intersection point via a novel joint root-finding algorithm (Sect. 3.3). We then sample near/far surface points $\{\bar{\mathbf{x}}\}$. (b) The sampled points are mapped into canonical space as $\{\hat{\mathbf{x}}\}$ via root-finding. (c) In canonical space, we run an SDF-based volume rendering with canonicalized points $\{\hat{\mathbf{x}}\}$, local body poses and shape (θ, β), an SDF network feature $\mathbf{z}$, surface normals $\mathbf{n}$, and a per-frame latent code $\mathcal{Z}$ to predict the corresponding pixel value of the input ray (Sect. 3.4). (d) All network modules, including the forward LBS network LBS_{σ_ω}, the canonical SDF network f_{σ_f}, and the canonical color network f_{σ_c}, are trained end-to-end with a photometric loss in image space and regularization losses in canonical space (Sect. 3.5).

3 Method

Our pipeline is illustrated in Fig. 2. Our model consists of a forward linear blend skinning (LBS) network (Sect. 3.1), a canonical SDF network, and a canonical color network (Sect. 3.2). When rendering a specific pixel of the image in observation space, we first find the intersection of the corresponding camera ray and the observation-space SDF iso-surface. Since we model a canonical SDF and a forward LBS, we propose a novel joint root-finding algorithm that can simultaneously search for the ray-surface intersection and the canonical correspondence of the intersection point (Sect. 3.3). Such a formulation does not condition the networks on observations in observation space. Consequently, it can generalize to unseen poses. Once the ray-surface intersection is found, we sample near/far surface points on the camera ray and find their canonical correspondences via forward LBS root-finding. The canonicalized points are used for volume rendering to compose the final RGB value at the pixel (Sect. 3.4). The predicted pixel color is then compared to the observation using a photometric loss (Sect. 3.5). The model is trained end-to-end using the photometric loss and regularization losses. The learned networks represent a personalized animatable avatar that can robustly synthesize new geometries and appearances under novel poses (Sect. 4.1).

3.1 Neural Linear Blend Skinning

Traditional parametric human body models [5,20,37,50,53,75] often use linear blend skinning (LBS) to deform a template model according to rigid bone transformations and skinning weights. We follow the notations of [71] to describe LBS. Given a set of N points in canonical space, $\hat{\mathbf{X}} = \{\hat{\mathbf{x}}^{(i)}\}_{i=1}^{N}$, LBS takes a set of rigid bone transformations $\{\mathbf{B}_b\}_{b=1}^{24}$ as inputs, each $\mathbf{B}_b$ being a 4×4 rotation-translation matrix. We use 23 local transformations and one global transformation with an underlying SMPL [37] model. For a 3D point $\hat{\mathbf{x}}^{(i)} \in \hat{\mathbf{X}}$[1], a skinning weight vector is defined as $\mathbf{w}^{(i)} \in [0,1]^{24}, \text{s.t.} \sum_{b=1}^{24} \mathbf{w}_b^{(i)} = 1$. This vector indicates the affinity of the point $\hat{\mathbf{x}}^{(i)}$ to each of the bone transformations $\{\mathbf{B}_b\}_{b=1}^{24}$. Following recent works [11,45,62,71], we use a neural network $f_{\sigma_\omega}(\cdot) : \mathbb{R}^3 \mapsto [0,1]^{24}$ with parameters σ_ω to predict the skinning weights of any point in space. The set of transformed points $\bar{\mathbf{X}} = \{\bar{\mathbf{x}}^{(i)}\}_{i=1}^{N}$ is related to $\hat{\mathbf{X}}$ via:

$$\begin{aligned} \bar{\mathbf{x}}^{(i)} &= LBS_{\sigma_\omega}\left(\hat{\mathbf{x}}^{(i)}, \{\mathbf{B}_b\}\right), \quad \forall i = 1, \ldots, N \\ \Longleftrightarrow \bar{\mathbf{x}}^{(i)} &= \left(\sum_{b=1}^{24} f_{\sigma_\omega}(\hat{\mathbf{x}}^{(i)})_b \mathbf{B}_b\right) \hat{\mathbf{x}}^{(i)}, \quad \forall i = 1, \ldots, N \end{aligned} \tag{1}$$

where Eq. (1) is referred to as the forward LBS function. The process of applying Eq. (1) to all points in $\hat{\mathbf{X}}$ is often referred to as *forward skinning*. For brevity, for the remainder of the paper, we drop $\{\mathbf{B}_b\}$ from the LBS function and write $LBS_{\sigma_\omega}(\hat{\mathbf{x}}^{(i)}, \{\mathbf{B}_b\})$ as $LBS_{\sigma_\omega}(\hat{\mathbf{x}}^{(i)})$.

3.2 Canonical SDF and Color Networks

We model an articulated human as a neural SDF $f_{\sigma_f}(\hat{\mathbf{x}}, \theta, \beta, \mathcal{Z})$ with parameters σ_f in canonical space, where $\hat{\mathbf{x}}$ denotes the canonical query point, θ and β denote local poses and body shape of the human which capture pose-dependent cloth deformations, and $\mathcal{Z}$ denotes a per-frame optimizable latent code which compensates for time-dependent dynamic cloth deformations. For brevity, we write this neural SDF as $f_{\sigma_f}(\hat{\mathbf{x}})$ in the remainder of the paper.

Similar to the canonical SDF network, we define a canonical color network with parameters σ_c as $f_{\sigma_c}(\hat{\mathbf{x}}, \mathbf{n}, \mathbf{v}, \mathbf{z}, \mathcal{Z}) : \mathbb{R}^{9+|\mathbf{z}|+|\mathcal{Z}|} \mapsto \mathbb{R}^3$. Here, $\mathbf{n}$ denotes a normal vector in the observation space. $\mathbf{n}$ is computed by transforming the canonical normal vectors using the rotational part of forward transformations $\sum_{b=1}^{24} f_{\sigma_\omega}(\hat{\mathbf{x}}^{(i)})_b \mathbf{B}_b$ (Eq. (1)). $\mathbf{v}$ denotes viewing direction. Similar to [69,79,80], $\mathbf{z}$ denotes an SDF feature which is extracted from the output of the second-last layer of the neural SDF. $\mathcal{Z}$ denotes a per-frame latent code which is shared with the SDF network. It compensates for time-dependent dynamic lighting effects. The outputs of f_{σ_c} are RGB color values in the range $[0,1]$.

3.3 Joint Root-Finding

While surface rendering [47,80] could be used to learn the network parameters introduced in Sects. 3.1 and 3.2, it cannot handle abrupt changes in depth, as demonstrated in [69]. We also observe severe geometric artifacts when applying surface rendering to our setup, we refer readers to the Supp. Mat. for such an

[1] With slight abuse of notation, we also use $\hat{\mathbf{x}}$ to represent points in homogeneous coordinates when necessary.

ablation. On the other hand, volume rendering can better handle abrupt depth changes in articulated human rendering. However, volume rendering requires multi-step dense sampling on camera rays [69,79], which, when combined naively with the iterative root-finding algorithm [11], requires significantly more memory and becomes prohibitively slow to train and test. We thus employ a hybrid method similar to [49]. We first search the ray-surface intersection and then sample near/far surface points on the ray. In practice, we initialize our SDF network with [71]. Thus, we fix the sampling depth interval around the surface to $[-5\,\text{cm}, +5\,\text{cm}]$.

A naive way of finding the ray-surface intersection is to use sphere tracing [19] and map each point to canonical space via root-finding [11]. In this case, we need to solve the costly root-finding problem during each step of the sphere tracing. This becomes prohibitively expensive when the number of rays is large. Thus, we propose an alternative solution. We leverage the skinning weights of the nearest neighbor on the registered SMPL mesh to the query point $\bar{\mathbf{x}}$ and use the inverse of the linearly combined forward bone transforms to map $\bar{\mathbf{x}}$ to its rough canonical correspondence. Combining this approximate backward mapping with sphere tracing, we obtain rough estimations of intersection points. Then, starting from these rough estimations, we apply a novel joint root-finding algorithm to search the precise intersection points and their correspondences in canonical space. In practice, we found that using a single initialization for our joint root-finding works well already. Adding more initializations incurs drastic memory and runtime overhead while not achieving any noticeable improvements. We hypothesize that this is due to the fact that our initialization is obtained using inverse transformations with SMPL skinning weights rather than rigid bone transformations (as was done in [11]).

Formally, we define a camera ray as $\mathbf{r} = (\mathbf{c}, \mathbf{v})$ where $\mathbf{c}$ is the camera center and $\mathbf{v}$ is a unit vector that defines the direction of this camera ray. Any point on the camera ray can be expressed as $\mathbf{c} + \mathbf{v} \cdot d$ with $d >= 0$. The joint root-finding aims to find canonical point $\hat{\mathbf{x}}$ and depth d on the ray in observation space, such that:

$$\begin{aligned} f_{\sigma_f}(\hat{\mathbf{x}}) &= 0 \\ LBS_{\sigma_\omega}(\hat{\mathbf{x}}) - (\mathbf{c} + \mathbf{v} \cdot d) &= \mathbf{0} \end{aligned} \tag{2}$$

in which $\mathbf{c}, \mathbf{v}$ are constants per ray. Denoting the joint vector-valued function as $g_{\sigma_f,\sigma_\omega}(\hat{\mathbf{x}}, d)$ and the joint root-finding problem as:

$$g_{\sigma_f,\sigma_\omega}(\hat{\mathbf{x}}, d) = \begin{bmatrix} f_{\sigma_f}(\hat{\mathbf{x}}) \\ LBS_{\sigma_\omega}(\hat{\mathbf{x}}) - (\mathbf{c} + \mathbf{v} \cdot d) \end{bmatrix} = \mathbf{0} \tag{3}$$

we can then solve it via Newton's method

$$\begin{bmatrix} \hat{\mathbf{x}}_{k+1} \\ d_{k+1} \end{bmatrix} = \begin{bmatrix} \hat{\mathbf{x}}_k \\ d_k \end{bmatrix} - \mathbf{J}_k^{-1} \cdot g_{\sigma_f,\sigma_\omega}(\hat{\mathbf{x}}_k, d_k) \tag{4}$$

where:

$$\mathbf{J}_k = \begin{bmatrix} \frac{\partial f_{\sigma_f}}{\partial \hat{\mathbf{x}}}(\hat{\mathbf{x}}_k) & 0 \\ \frac{\partial LBS_{\sigma_\omega}}{\partial \hat{\mathbf{x}}}(\hat{\mathbf{x}}_k) & -\mathbf{v} \end{bmatrix} \tag{5}$$

Following [11], we use Broyden's method to avoid computing $\mathbf{J}_k$ at each iteration.

Amortized Complexity: Given the number of sphere-tracing steps as N and the number of root-finding steps as M, the amortized complexity for joint root-finding is $O(M)$ while naive alternation between sphere-tracing and root-finding is $O(MN)$. In practice, this results in about 5× speed up of joint root-finding compared to the naive alternation between sphere-tracing and root-finding. We also note that from a theoretical perspective, our proposed joint root-finding converges quadratically while the secant-method-based root-finding in the concurrent work [85] converges only superlinearly.

We describe how to compute implicit gradients wrt. the canonical SDF and the forward LBS in the Supp. Mat. In the main paper, we use volume rendering which does not need to compute implicit gradients wrt. the canonical SDF.

3.4 Differentiable Volume Rendering

We employ a recently proposed SDF-based volume rendering formulation [79]. Specifically, we convert SDF values into density values σ using the scaled CDF of the Laplace distribution with the negated SDF values as input

$$\sigma(\hat{\mathbf{x}}) = \frac{1}{b}\left(\frac{1}{2} + \frac{1}{2}\mathrm{sign}(-f_{\sigma_f}(\hat{\mathbf{x}}))\big(1 - \exp(-\frac{|-f_{\sigma_f}(\hat{\mathbf{x}})|}{b})\right) \tag{6}$$

where b is a learnable parameter. Given the surface point found via solving Eq. (3), we sample 16 points around the surface points and another 16 points between the near scene bound and the surface point, and map them to canonical space along with the surface point. For rays that do not intersect with any surface, we uniformly sample 64 points for volume rendering. With N sampled points on a ray $\mathbf{r} = (\mathbf{c}, \mathbf{v})$, we use standard volume rendering [46] to render the pixel color

$$\hat{C}(\mathbf{r}) = \sum_{i=1}^{N} T^{(i)} \left(1 - \exp(-\sigma(\hat{\mathbf{x}}^{(i)})\delta^{(i)})\right) f_{c_\sigma}(\hat{\mathbf{x}}^{(i)}, \mathbf{n}^{(i)}, \mathbf{v}, \mathbf{z}, \mathcal{Z}) \tag{7}$$

$$T^{(i)} = \exp\left(-\sum_{j<i} \sigma(\hat{\mathbf{x}}^{(j)})\delta^{(j)}\right) \tag{8}$$

where $\delta^{(i)} = |d^{(i+1)} - d^{(i)}|$.

3.5 Loss Function

Our loss consists of a photometric loss in observation space and multiple regularizers in canonical space

$$\mathcal{L} = \lambda_C \cdot \mathcal{L}_C + \lambda_E \cdot \mathcal{L}_E + \lambda_O \cdot \mathcal{L}_O + \lambda_I \cdot \mathcal{L}_I + \lambda_S \cdot \mathcal{L}_S \tag{9}$$

$\mathcal{L}_C$ is the L1 loss for color predictions. $\mathcal{L}_E$ is the Eikonal regularization [16]. $\mathcal{L}_O$ is an off-surface point loss, encouraging points far away from the SMPL mesh to have positive SDF values. Similarly, $\mathcal{L}_I$ regularizes points inside the canonical SMPL mesh to have negative SDF values. $\mathcal{L}_S$ encourages the forward LBS network to predict similar skinning weights to the canonical SMPL mesh. Different from [26,74,80], we do not use an explicit silhouette loss. Instead, we utilize foreground masks and set all background pixel values to zero. In practice, this encourages the SDF network to predict positive SDF values for points on rays that do not intersect with foreground masks. For detailed definitions of loss terms and model architectures, please refer to the Supp. Mat.

4 Experiments

We validate the generalization ability and reconstruction quality of our proposed method against several recent baselines [54,56,65]. As was done in [56], we consider a setup with 4 cameras positioned equally spaced around the human subject. For an ablation study on different design choices of our model, including ray sampling strategy, LBS networks, and number of initializations for root-finding, we refer readers to the Supp. Mat.

Datasets: We use the ZJU-MoCap [56] dataset as our primary testbed because its setup includes 23 cameras which allows us to extract pseudo-ground-truth geometry to evaluate our model. More specifically, the dataset consists of 9 sequences captured with 23 calibrated cameras. We use the training/testing splits from Neural Body [56] for both the cameras and the poses. As one of our goals is learn to detailed geometry, we collect pseudo-ground-truth geometry for the training poses. We use all 23 cameras and apply NeuS with a background NeRF model [69], a state-of-the-art method for multi-view reconstruction. Note that we refrain from using the masks provided by Neural Body [56] as these masks are noisy and insufficient for accurate static scene reconstruction. We observe that geometry reconstruction with NeuS [69] fails when subjects wear black clothes or the environmental light is not bright enough. Therefore, we manually exclude bad reconstructions and discard sequences with less than 3 valid reconstructions. For completeness, we also tested our approach on the H36M dataset [25] and report a quantitative comparison to [48,54] in the Supp. Mat.

Baselines: We compare against three major baselines: Neural Body [56](NB), Ani-NeRF [54](AniN), and A-NeRF [65](AN). Neural Body diffuses per-SMPL-vertex latent codes into observation space as additional conditioning for NeRF models to achieve state-of-the-art novel view synthesis results on training poses.

Ani-NeRF learns a canonical NeRF model and a backward LBS network which predicts residuals to the deterministic SMPL-based backward LBS. Consequently, the LBS network needs to be re-trained for each test sequence. A-NeRF employs a deterministic backward mapping with bone-relative embeddings for query points and only uses keypoints and joint rotations instead of surface models (*i.e.* SMPL surface). For the detailed setups of these baselines, please refer to the Supp. Mat.

Benchmark Tasks: We benchmark our approach on three tasks: generalization to unseen poses, geometry reconstruction, and novel-view synthesis. To analyze generalization ability, we evaluate the trained models on unseen testing poses. Due to the stochastic nature of cloth deformations, we quantify performance via perceptual similarity to the ground-truth images with the LPIPS [84] metric. We report PSNR and SSIM in the Supp. Mat. We also encourage readers to check out qualitative comparison videos at https://neuralbodies.github.io/arah/.

For geometry reconstruction, we evaluate our method and baselines on the training poses. We report point-based L2 Chamfer distance (CD) and normal consistency (NC) wrt. the pseudo-ground-truth geometry. During the evaluation, we only keep the largest connected component of the reconstructed meshes. Note that is in favor of the baselines as they are more prone to producing floating blob artifacts. We also remove any ground-truth or predicted mesh points that are below an estimated ground plane to exclude outliers from the ground plane from the evaluation. For completeness, we also evaluate novel-view synthesis with PSNR, SSIM, and LPIPS using the poses from the training split.

Table 1. Generalization to Unseen Poses. We report LPIPS [84] on synthesized images under unseen poses from the testset of the ZJU-MoCap dataset [56] (*i.e.* all views except 0, 6, 12, and 18). Our approach consistently outperforms the baselines by a large margin. We report PSNR and SSIM in the Supp. Mat.

Sequence	Metric	NB	AniN	AN	Ours
313	LPIPS ↓	0.126	0.115	0.209	**0.092**
315	LPIPS ↓	0.152	0.167	0.232	**0.105**
377	LPIPS ↓	0.119	0.153	0.165	**0.093**
386	LPIPS ↓	0.171	0.187	0.241	**0.127**
387	LPIPS ↓	0.135	0.145	0.162	**0.099**
390	LPIPS ↓	0.163	0.173	0.226	**0.126**
392	LPIPS ↓	0.135	0.169	0.183	**0.106**
393	LPIPS ↓	0.132	0.155	0.175	**0.104**
394	LPIPS ↓	0.150	0.171	0.199	**0.111**

Table 2. Geometry Reconstruction. We report L2 Chamfer Distance (CD) and Normal Consistency (NC) on the training poses of the ZJU-MoCap dataset [56]. Note that AniN and AN occasionally produce large background blobs that are connected to the body resulting in large deviations from the ground truth.

Sequence	Metric	NB	AniN	AN	Ours
313	CD ↓	1.258	1.242	9.174	**0.707**
	NC ↑	0.700	0.599	0.691	**0.809**
315	CD ↓	2.167	2.860	1.524	**0.779**
	NC ↑	0.636	0.450	0.610	**0.753**
377	CD ↓	1.062	1.649	1.008	**0.840**
	NC ↑	0.672	0.541	0.682	**0.786**
386	CD ↓	2.938	23.53	3.632	**2.880**
	NC ↑	0.607	0.325	0.596	**0.741**
393	CD ↓	1.753	3.252	1.696	**1.342**
	NC ↑	0.600	0.481	0.605	**0.739**
394	CD ↓	1.510	2.813	558.8	**1.177**
	NC ↑	0.628	0.540	0.639	**0.762**

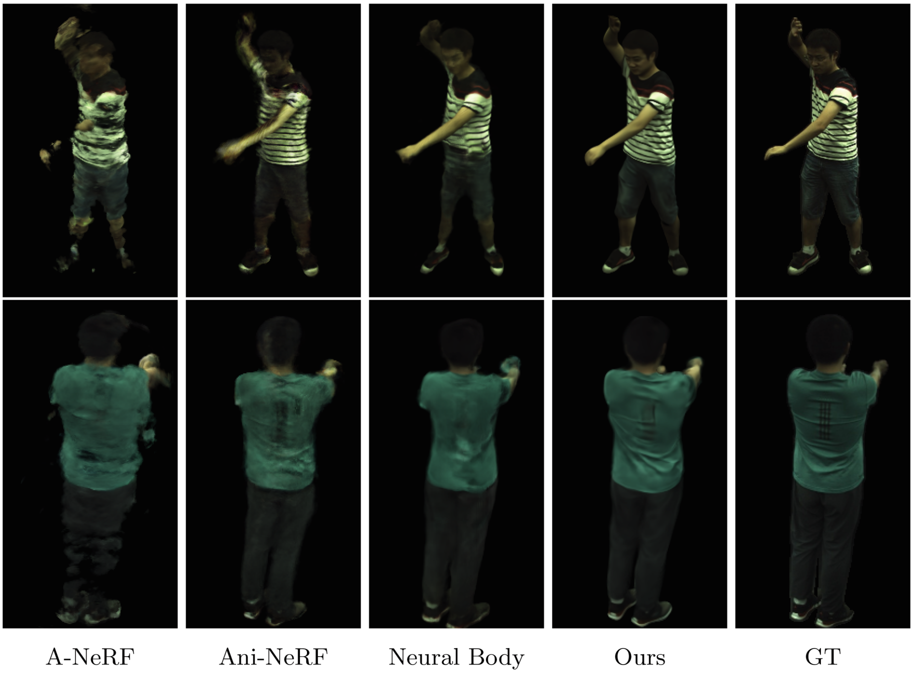


Fig. 3. Generalization to Unseen Poses on the testing poses of ZJU-MoCap. A-NeRF struggles with unseen poses due to the limited training poses and the lack of a SMPL surface prior. Ani-NeRF produces noisy images as it uses an inaccurate backward mapping function. Neural Body loses details, e.g. wrinkles, because its conditional NeRF is learned in observation space. Our approach generalizes well to unseen poses and can model fine details like wrinkles.

4.1 Generalization to Unseen Poses

We first analyze the generalization ability of our approach in comparison to the baselines. Given a trained model and a pose from the test set, we render images of the human subject in the given pose. We show qualitative results in Fig. 3 and quantitative results in Table 1. We significantly outperform the baselines both qualitatively and quantitatively. The training poses of the ZJU-MoCap dataset are extremely limited, usually comprising just 60–300 frames of repetitive motion. This limited training data results in severe overfitting for the baselines. In contrast, our method generalizes well to unseen poses, even when training data is limited.

We additionally animate our models trained on the ZJU-MoCap dataset using extreme out-of-distribution poses from the AMASS [40] and AIST++ [32] datasets. As shown in Fig. 5, even under extreme pose variation our approach produces plausible geometry and rendering results while all baselines show severe artifacts. We attribute the large improvement on unseen poses to our root-finding-based backward skinning, as the learned forward skinning weights are

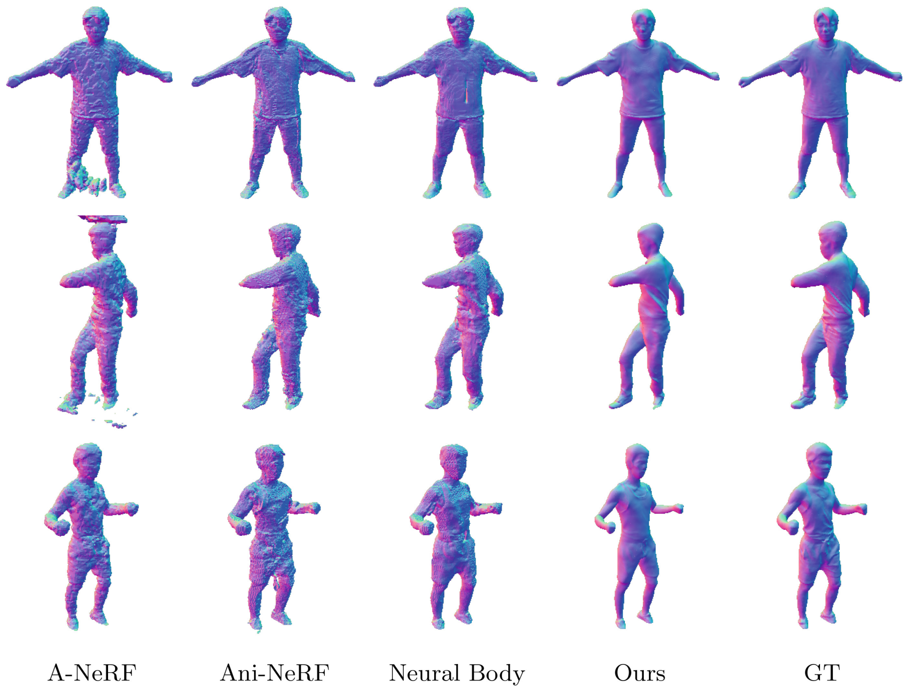


Fig. 4. Geometry Reconstruction. Our approach reconstructs more fine-grained geometry than the baselines while preserving high-frequency details such as wrinkles. Note that we remove an estimated ground plane from all meshes.

constants per subject, while root-finding is a deterministic optimization process that does not rely on learned neural networks that condition on inputs from the observation space. More comparisons can be found in the Supp. Mat.

4.2 Geometry Reconstruction on Training Poses

Next, we analyze the geometry reconstructed with our approach against reconstructions from the baselines. We compare to the pseudo-ground-truth obtained from NeuS [69]. We show qualitative results in Fig. 4 and quantitative results in Table 2. Our approach consistently outperforms existing NeRF-based human models on geometry reconstruction. As evidenced in Fig. 4, the geometry obtained with our approach is much cleaner compared to NeRF-based baselines, while preserving high-frequency details such as wrinkles.

Table 3. Novel View Synthesis. We report PSNR, SSIM, and LPIPS [84] for novel views of training poses of the ZJU-MoCap dataset [56]. Due to better geometry, our approach produces more consistent rendering results across novel views than the baselines. We include qualitative comparisons in the Supp. Mat. Note that we crop slightly larger bounding boxes than Neural Body [56] to better capture loose clothes, *e.g.* sequence 387 and 390. Therefore, the reported numbers vary slightly from their evaluation.

	313			315			377		
Method	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
NB	30.5	0.967	0.068	26.4	0.958	0.079	**28.1**	**0.956**	0.080
Ani-N	29.8	0.963	0.075	23.1	0.917	0.138	24.2	0.925	0.124
A-NeRF	29.2	0.954	0.075	25.1	0.948	0.087	27.2	0.951	0.080
Ours	**31.6**	**0.973**	**0.050**	**27.0**	**0.965**	**0.058**	27.8	**0.956**	**0.071**
	386			387			390		
Method	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
NB	29.0	**0.935**	0.112	26.7	0.942	0.101	**27.9**	0.928	0.112
Ani-N	25.6	0.878	0.199	25.4	0.926	0.131	26.0	0.912	0.148
A-NeRF	28.5	0.928	0.127	26.3	0.937	0.100	27.0	0.914	0.126
Ours	**29.2**	0.934	**0.105**	**27.0**	**0.945**	**0.079**	**27.9**	**0.929**	**0.102**
	392			393			394		
Method	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
NB	**29.7**	**0.949**	0.101	**27.7**	0.939	0.105	28.7	0.942	0.098
Ani-N	28.0	0.931	0.151	26.1	0.916	0.151	27.5	0.924	0.142
A-NeRF	28.7	0.942	0.106	26.8	0.931	0.113	28.1	0.936	0.103
Ours	29.5	0.948	**0.090**	**27.7**	**0.940**	**0.093**	**28.9**	**0.945**	**0.084**

4.3 Novel View Synthesis on Training Poses

Lastly, we analyze our approach for novel view synthesis on training poses. Table 3 provides a quantitative comparison to the baselines. While not the main focus of this work, our approach also outperforms existing methods on novel view synthesis. This suggests that more faithful modeling of geometry is also beneficial for the visual fidelity of novel views. Particularly when few training views are available, NeRF-based methods produce blob/cloud artifacts. By removing such artifacts, our approach achieves high image fidelity and better consistency across novel views. Due to space limitations, we include further qualitative results on novel view synthesis in the Supp. Mat.

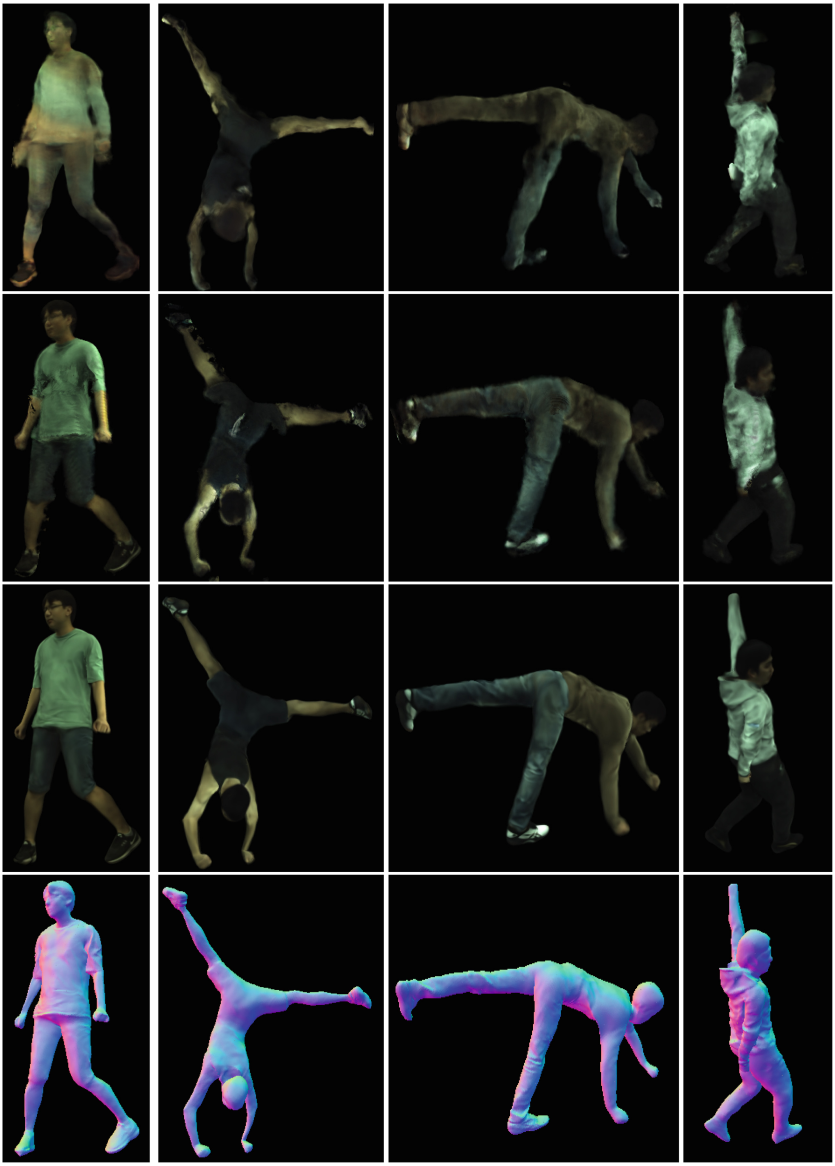


Fig. 5. Qualitative Results on Out-of-distribution Poses from the AMASS [40] and AIST++ [32] datasets. From top to bottom row: Neural Body, Ani-NeRF, our rendering, and our geometry. Note that Ani-NeRF requires re-training their backward LBS network on novel pose sequence. We did not show A-NeRF results as it already produces severe overfitting effects on ZJU-MoCap test poses. For more qualitative comparisons, please refer to the Supp. Mat.

5 Conclusion

We propose a new approach to create animatable avatars from sparse multi-view videos. We largely improve geometry reconstruction over existing approaches by modeling the geometry as articulated SDFs. Further, our novel joint root-finding algorithm enables generalization to extreme out-of-distribution poses. We discuss limitations of our approach in the Supp. Mat.

Acknowledgement. Shaofei Wang and Siyu Tang acknowledge the SNF grant 200021 204840. Katja Schwarz was supported by the BMWi in the project KI Delta Learning (project number 19A19013O). Andreas Geiger was supported by the ERC Starting Grant LEGO-3D (850533) and the DFG EXC number 2064/1 - project number 390727645.

References

1. Alldieck, T., Magnor, M., Bhatnagar, B.L., Theobalt, C., Pons-Moll, G.: Learning to reconstruct people in clothing from a single RGB camera. In: Proceedings of CVPR (2019)
2. Alldieck, T., Magnor, M., Xu, W., Theobalt, C., Pons-Moll, G.: Video based reconstruction of 3d people models. In: Proceedings of CVPR (2018)
3. Alldieck, T., Xu, H., Sminchisescu, C.: imGHUM: implicit generative models of 3d human shape and articulated pose. In: Proceedings of CVPR (2021)
4. Alldieck, T., Zanfir, M., Sminchisescu, C.: Photorealistic monocular 3d reconstruction of humans wearing clothing. In Proceedings of CVPR (2022)
5. Anguelov, D., Srinivasan, P., Koller, D., Thrun, S., Rodgers, J., Davis, J.: Scape: shape completion and animation of people. In: ACM Transasctions Graphics, vol. 24 (2005)
6. Bhatnagar, B.L., Sminchisescu, C., Theobalt, C., Pons-Moll, G.: Combining implicit function learning and parametric models for 3d human reconstruction. In: proceedings of ECCV (2020)
7. Bhatnagar, B.L., Sminchisescu, C., Theobalt, C., Pons-Moll, G.: LoopReg: self-supervised learning of implicit surface correspondences, pose and shape for 3d human mesh registration. In: Proceedings of NeurIPS (2020)
8. Burov, A., Nießner, M., Thies, J.: Dynamic surface function networks for clothed human bodies. In: proceedings of ICCV (2021)
9. Chen, J., et al.: Animatable neural radiance fields from monocular RGB videos. arXiv preprint arXiv:2106.13629 (2021)
10. Chen, X., et al.: gDNA: towards generative detailed neural avatars. In: Proceedings of CVPR (2022)
11. Chen, X., Zheng, Y., Black, M., Hilliges, O., Geiger, A.: Snarf: differentiable forward skinning for animating non-rigid neural implicit shapes. In: Proceedings of ICCV (2021)
12. Chen, Z., Zhang, H.: Learning implicit fields for generative shape modeling. In: Proceedings of CVPR (2019)
13. Chibane, J., Alldieck, T., Pons-Moll, G.: Implicit functions in feature space for 3d shape reconstruction and completion. In: Proceedings of CVPR (2020)

14. Corona, E., Pumarola, A., Alenyà, G., Pons-Moll, G., Moreno-Noguer, F.: SMPLicit: topology-aware generative model for clothed people. In: Proceedings of CVPR (2021)
15. Dong, Z., Guo, C., Song, J., Chen, X., Geiger, A., Hilliges, O.: Pina: Learning a personalized implicit neural avatar from a single RGB-D video sequence. In: In Proceedings of of CVPR (2022)
16. Gropp, A., Yariv, L., Haim, N., Atzmon, M., Lipman, Y.: Implicit geometric regularization for learning shapes. In: Proceedings of of ICML (2020)
17. Guan, P., Reiss, L., Hirshberg, D.A., Weiss, E., Black, M.J.: Drape: dressing any person. ACM Trans. Graph. **31**(4), 1–10 (2012)
18. Gundogdu, E., Constantin, V., Seifoddini, A., Dang, M., Salzmann, M., Fua, P.: Garnet: a two-stream network for fast and accurate 3d cloth draping. In: Proceedings of of ICCV (2019)
19. Hart, J.C.: Sphere tracing: a geometric method for the antialiased ray tracing of implicit surfaces. Vis. Comput. **12**(10), 527–545 (1995)
20. Hasler, N., Stoll, C., Sunkel, M., Rosenhahn, B., Seidel, H.P.: A statistical model of human pose and body shape. Comput. Graph. Forum **28**, 337–346 (2009)
21. He, T., Collomosse, J., Jin, H., Soatto, S.: Geo-PIFu: geometry and pixel aligned implicit functions for single-view human reconstruction. In: Proceedings of of NeurIPS (2020)
22. He, T., Xu, Y., Saito, S., Soatto, S., Tung, T.: Arch++: animation-ready clothed human reconstruction revisited. In: Proceedings of of ICCV (2021)
23. Hu, T., Yu, T., Zheng, Z., Zhang, H., Liu, Y., Zwicker, M.: HVTR: hybrid volumetric-textural rendering for human avatars. arXiv preprint arXiv:2112.10203 (2021)
24. Huang, Z., Xu, Y., Lassner, C., Li, H., Tung, T.: ARCH: animatable reconstruction of Clothed Humans. In: Proceedings of CVPR (2020)
25. Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: Human3.6 m: large scale datasets and predictive methods for 3d human sensing in natural environments. IEEE Trans. Pattern Anal. Mach. Intell. **36**(7), 1325–1339 (2014)
26. Jiang, B., Hong, Y., Bao, H., Zhang, J.: SelfRecon: self reconstruction your digital avatar from monocular video. In: Proceedings of CVPR (2022)
27. Joo, H., Simon, T., Sheikh, Y.: Total capture: A 3d deformation model for tracking faces, hands, and bodies. In: Proceedings of CVPR (2018)
28. Kwon, Y., Kim, D., Ceylan, D., Fuchs, H.: Neural human performer: Learning generalizable radiance fields for human performance rendering. In: Proceedings of NeurIPS (2021)
29. Lähner, Z., Cremers, D., Tung, T.: DeepWrinkles: accurate and realistic clothing modeling. In: Proceedings of ECCV (2018)
30. Li, R., et al.: TAVA: template-free animatable volumetric actors. In: Proceedings of ECCV (2022)
31. Li, R., Xiu, Y., Saito, S., Huang, Z., Olszewski, K., Li, H.: Monocular real-time volumetric performance capture. In: Proceedings of ECCV (2020)
32. Li, R., Yang, S., Ross, D.A., Kanazawa, A.: Ai choreographer: music conditioned 3d dance generation with aist++. In: Proceedings of ICCV (2021)
33. Li, Y., Habermann, M., Thomaszewski, B., Coros, S., Beeler, T., Theobalt, C.: Deep physics-aware inference of cloth deformation for monocular human performance capture. In: Proceedings of 3DV (2021)
34. Li, Z., Yu, T., Pan, C., Zheng, Z., Liu, Y.: Robust 3d self-portraits in seconds. In: Proceedings of CVPR (2020)

35. Li, Z., Yu, T., Zheng, Z., Guo, K., Liu, Y.: POSEFusion: pose-guided selective fusion for single-view human volumetric capture. In: Proceedings of CVPR (2021)
36. Liu, L., Habermann, M., Rudnev, V., Sarkar, K., Gu, J., Theobalt, C.: Neural actor: neural free-view synthesis of human actors with pose control. ACM Trans. Graph. (ACM SIGGRAPH Asia) **40**(6), 1–16 (2021)
37. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: SMPL: a skinned multi-person linear model. ACM Trans. Graph. **34**(6), 1–16 (2015)
38. Ma, Q., Saito, S., Yang, J., Tang, S., Black, M.J.: SCALE: modeling clothed humans with a surface codec of articulated local elements. In: Proceedings of CVPR (2021)
39. Ma, Q., Yang, J., Tang, S., Black, M.J.: The power of points for modeling humans in clothing. In: Proceedings of ICCV (2021)
40. Mahmood, N., Ghorbani, N., Troje, N.F., Pons-Moll, G., Black, M.J.: AMASS: archive of motion capture as surface shapes. In: Proceedings of ICCV (2019)
41. Mescheder, L., Oechsle, M., Niemeyer, M., Nowozin, S., Geiger, A.: Occupancy networks: Learning 3d reconstruction in function space. In: Proceedings of CVPR (2019)
42. Michalkiewicz, M., Pontes, J.K., Jack, D., Baktashmotlagh, M., Eriksson, A.: Implicit surface representations as layers in neural networks. In: Proceedings of ICCV (2019)
43. Mihajlovic, M., Bansal, A., Zollhoefer, M., Tang, S., Saito, S.: KeypointNeRF: generalizing image-based volumetric avatars using relative spatial encoding of keypoints. In: Proceedings of ECCV (2022)
44. Mihajlovic, M., Saito, S., Bansal, A., Zollhoefer, M., Tang, S.: COAP: compositional articulated occupancy of people. In: Proceedings of CVPR (2022)
45. Mihajlovic, M., Zhang, Y., Black, M.J., Tang, S.: LEAP: learning articulated occupancy of people. In: Proceedings of CVPR (2021)
46. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: representing scenes as neural radiance fields for view synthesis. In: Proceedings of ECCV (2020)
47. Niemeyer, M., Mescheder, L., Oechsle, M., Geiger, A.: Differentiable volumetric rendering: learning implicit 3d representations without 3d supervision. In: Proceedings of CVPR (2020)
48. Noguchi, A., Sun, X., Lin, S., Harada, T.: Neural articulated radiance field. In: Proceedings of ICCV (2021)
49. Oechsle, M., Peng, S., Geiger, A.: UNISURF: unifying neural implicit surfaces and radiance fields for multi-view reconstruction. In: Proceedings of ICCV (2021)
50. Osman, A.A.A., Bolkart, T., Black, M.J.: Star: Sparse trained articulated human body regressor. In: Proceedings of ECCV (2020)
51. Park, J.J., Florence, P., Straub, J., Newcombe, R., Lovegrove, S.: DeepSDF: learning continuous signed distance functions for shape representation. In: Proceedings of CVPR (2019)
52. Patel, C., Liao, Z., Pons-Moll, G.: TailorNet: predicting clothing in 3d as a function of human pose, shape and garment style. In: Proceedings of CVPR (2020)
53. Pavlakos, G., et al.: Expressive body capture: 3d hands, face, and body from a single image. In: Proceedings of CVPR (2019)
54. Peng, S., et al.: Animatable neural radiance fields for modeling dynamic human bodies. In: Proceedings of ICCV (2021)
55. Peng, S., et al.: Animatable neural implict surfaces for creating avatars from videos. arXiv preprint arXiv:2203.08133 (2022)

56. Peng, S., et al.: Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In: Proceedings of CVPR (2021)
57. Peng, S., et al.: Shape as points: a differentiable poisson solver. In: Proceedings of NeurIPS (2021)
58. Prokudin, S., Black, M.J., Romero, J.: SMPLpix: neural avatars from 3D human models. In: Proceedings WACV (2021)
59. Raj, A., Tanke, J., Hays, J., Vo, M., Stoll, C., Lassner, C.: Anr-articulated neural rendering for virtual avatars. In: Proceedings of CVPR (2021)
60. Saito, S., Huang, Z., Natsume, R., Morishima, S., Kanazawa, A., Li, H.: PIFu: pixel-aligned implicit function for high-resolution clothed human digitization. In: Proceedings of ICCV (2019)
61. Saito, S., Simon, T., Saragih, J., Joo, H.: PIFuHD: multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In: Proceedings of CVPR (2020)
62. Saito, S., Yang, J., Ma, Q., Black, M.J.: SCANimate: Weakly supervised learning of skinned clothed avatar networks. In: Proceedings of CVPR (2021)
63. Santesteban, I., Thuerey, N., Otaduy, M.A., Casas, D.: Self-Supervised Collision Handling via Generative 3D Garment Models for Virtual Try-On. In: Proceedings of CVPR (2021)
64. Su, S.Y., Bagautdinov, T., Rhodin, H.: DANBO: disentangled articulated neural body representations via graph neural networks. In: Proceedings of ECCV (2022)
65. Su, S.Y., Yu, F., Zollhoefer, M., Rhodin, H.: A-neRF: articulated neural radiance fields for learning human shape, appearance, and pose. In: Proceedings of NeurIPS (2021)
66. Su, Z., Xu, L., Zheng, Z., Yu, T., Liu, Y., Fang, L.: RobustFusion: human volumetric capture with data-driven visual cues using a RGBD camera. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12349, pp. 246–264. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58548-8_15
67. Tiwari, G., Sarafianos, N., Tung, T., Pons-Moll, G.: Neural-GIF: neural generalized implicit functions for animating people in clothing. In: Proceedings of ICCV (2021)
68. Tiwari, L., Bhowmick, B.: DeepDraper: fast and accurate 3d garment draping over a 3d human body. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops (2021)
69. Wang, P., Liu, L., Liu, Y., Theobalt, C., Komura, T., Wang, W.: NeuS: learning neural implicit surfaces by volume rendering for multi-view reconstruction. In: Proceedings NeurIPS (2021)
70. Wang, S., Geiger, A., Tang, S.: Locally aware piecewise transformation fields for 3d human mesh registration. In: In Proceedings of CVPR (2021)
71. Wang, S., Mihajlovic, M., Ma, Q., Geiger, A., Tang, S.: MetaAvatar: learning animatable clothed human models from few depth images. In: Proceedings of NeurIPS (2021)
72. Weng, C.Y., Curless, B., Srinivasan, P.P., Barron, J.T., Kemelmacher-Shlizerman, I.: HumanNeRF: free-viewpoint rendering of moving people from monocular video. In: Proceedings CVPR (2022)
73. Xiu, Y., Yang, J., Tzionas, D., Black, M.J.: ICON: implicit clothed humans obtained from Normals. In: Proceedings of CVPR (2022)
74. Xu, H., Alldieck, T., Sminchisescu, C.: H-neRF: neural radiance fields for rendering and temporal reconstruction of humans in motion. In: Proceedings of NeurIPS (2021)
75. Xu, H., Bazavan, E.G., Zanfir, A., Freeman, W.T., Sukthankar, R., Sminchisescu, C.: GHUM & GHUML: generative 3d human shape and articulated pose models. In: Proceedings of CVPR (2020)

76. Xu, L., Su, Z., Han, L., Yu, T., Liu, Y., Fang, L.: UnstructuredFusion: real-time 4d geometry and texture reconstruction using commercial RGBD cameras. IEEE Trans. Pattern Anal. Mach. Intell. **42**(10), 2508–2522 (2020)
77. Xu, T., Fujita, Y., Matsumoto, E.: Surface-aligned neural radiance fields for controllable 3d human synthesis. In: CVPR (2022)
78. Yang, J., Franco, J.S., Hétroy-Wheeler, F., Wuhrer, S.: Analyzing clothing layer deformation statistics of 3d human motions. In: Proceedings of ECCV (2018)
79. Yariv, L., Gu, J., Kasten, Y., Lipman, Y.: Volume rendering of neural implicit surfaces. In: Proceedings of NeurIPS (2021)
80. Yariv, L., et al.: Multiview neural surface reconstruction by disentangling geometry and appearance. In: Proceedings of NeurIPS (2020)
81. Yu, T., et al.: DoubleFusion: real-time capture of human performances with inner body shapes from a single depth sensor. In: Proceedings of CVPR (2018)
82. Zakharkin, I., Mazur, K., Grigorev, A., Lempitsky, V.: Point-based modeling of human clothing. In: Proceedings of ICCV (2021)
83. Zhang, C., Pujades, S., Black, M.J., Pons-Moll, G.: Detailed, accurate, human shape estimation from clothed 3d scan sequences. In: Proceedings of CVPR (2017)
84. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of CVPR (2018)
85. Zheng, Y., Abrevaya, V.F., Bühler, M.C., Chen, X., Black, M.J., Hilliges, O.: I M Avatar: implicit morphable head avatars from videos. In: Proceedings of CVPR (2022)
86. Zheng, Z., Yu, T., Liu, Y., Dai, Q.: Pamir: parametric model-conditioned implicit representation for image-based human reconstruction. IEEE Trans. Pattern Anal. Mach. Intell. 1 (2021). https://doi.org/10.1109/TPAMI.2021.3050505
87. Zuo, X., Wang, S., Sun, Q., Gong, M., Cheng, L.: Self-supervised 3d human mesh recovery from noisy point clouds. arXiv preprint arXiv:2107.07539 (2021)