AvatarOne: Monocular 3D Human Animation

Akash Karthikeyan[1] Robert Ren[1] Yash Kant[1] Igor Gilitschenski[1,2]
[1]University of Toronto [2]Vector Institute for AI

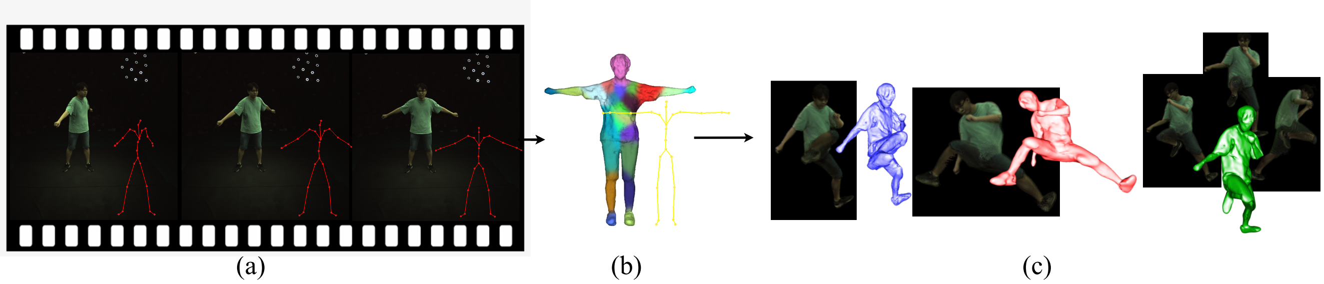

Figure 1. **Overview of AvatarOne.** We present AvatarOne, which models the 3D deformable models only from monocular videos and tracked skeleton. **(a)**. Our method can model a temporally consistent human avatar that is dynamically updated **(b)** and supports novel pose and view synthesis **(c)**.

Abstract

Reconstructing realistic human avatars from monocular videos is a challenge that demands intricate modeling of 3D surface and articulation. In this paper, we introduce a comprehensive approach that synergizes three pivotal components: (1) a Signed Distance Field (SDF) representation with volume rendering and grid-based ray sampling to prune empty raysets, enabling efficient 3D reconstruction; (2) faster 3D surface reconstruction through a warmup stage for human surfaces, which ensures detailed modeling of body limbs; and (3) temporally consistent subject-specific forward canonical skinning, which helps in retaining correspondences across frames, all of which can be trained in an end-to-end fashion under 15 minutes.

Leveraging warmup and grid-based ray marching, along with a faster voxel-based correspondence search, our model streamlines the computational demands of the problem. We further experiment with different sampling representations to improve ray radiance approximations and obtain a floater free surface. Through rigorous evaluation, we demonstrate that our method is on par with current techniques while offering novel insights and avenues for future research in 3D avatar modeling. This work showcases a fast and robust solution for both surface modeling and novel-view animation. Project website: https://aku02.github.io/projects/avatarone

1. Introduction

Building 3D models of humans is essential for a variety applications like telepresence and digital entertainment. While traditional solutions have multi camera rigged environments or controlled studios with calibrated depth sensors, latest advancements in neural rendering have led to more scalable and cost-effective solutions. Recent works utilize a deformation module parametrized by the neural fields to capture the dynamic human motions. These approaches choose to define the skinning weights (continuous skinning weights field) in deformed space (backward skinning leads to limited ability to generalize skinning weights to unfamiliar poses, as it relies on memorizing these weights in previously encountered configurations). To utilize the avatars and animatable models in general and personalized scenarios, *it's essential to build 3D animatable model directly from monocular videos, which can be readily rendered at novel poses.*

While some other works [21, 44, 57] utilizes the blend weights from template models, such as SMPL [31]. These model often encounters limitations in accurately representing clothing regions, as SMPL lacks specific modeling for these regions. Consequently, learning a deformation field that is both robust and capable of generalization remains an ongoing challenge in the field.

To recover photo-realistic avatars from monocular videos that are capable of generalizing to unseen poses, several components are essential. **1)** The correspondence search and skinning weight field should be defined in the canonical space, thereby ensuring their inherent pose-agnostic properties. This independence ensures that the model does not suffer from generalization issues [9, 63, 64]; and **2)** A deformation field that retains correspondences and alleviates ambiguous ones.

To summarize, our main contributions are:

- We present a new approach that can obtain both novel view and pose rendering of a human actor through explicit pose control, only requiring a monocular video as supervision, in less than **15 minutes**.
- We propose the use of SDF-based SMPL canonical human surface initialization and warm-up stages to make the model aware of a temporally consistent human surface. This helps in faster convergence of root-finding algorithm.
- We adopt a grid based ray-sampler to enable faster rendering and importance based Sampling via Transmittance. It also helps us estimate the weights of the ray samples without the memory intensive integration part.
- We adopt voxelized skinning weights to create an end-to-end learnable pipeline for reposing. This enables the rendering of human actors in a diverse range of poses while also maintaining correspondences across those poses.

2. Related Work

2.1. Neural Scene Representations and Animation

In recent years, there has been notable advancement in neural scene representations, especially in coordinate-based methods, showing impressive success in shape encoding [34, 40] and appearance [30, 35, 54]. Yet, enabling these representations to support deformability and animation presents a problem [28,41,42,60]. Notably, the current state-of-the-art methods are not built to control the scene beyond interpolations and do not preserve correspondences across different poses, impeding content creation or editing.

Recent strides towards animatable Neural Radiance Fields (NeRFs) offer promising solutions to these limitations. Several studies have proposed controllable animatable NeRFs [29, 38, 44, 45, 56, 64], introducing an array of techniques such as pose-dependent radiance fields, latent codes anchored on deformable meshes, and transformation optimization between view and canonical space. However, most of these methods still fall short in handling creatures beyond humans, relying heavily on SMPL [31] body templates and often not generalizing well to unseen poses.

2.2. Non-Rigid Shape Reconstruction and Animatable Shapes

Traditional non-rigid shape reconstruction methods typically establish a fixed canonical space across frames and employ a deformation model to map canonical to deformed space [2, 5–7, 11, 16, 24, 25, 31, 55, 61, 66]. Contemporary approaches have shifted towards modeling inverse deformation fields [12, 37, 41, 47, 50], yet they face difficulties in generalizing to unseen poses. A notable exception is SNARF [9], which leverages a forward deformation field. However, unlike our method, these methods generally require 3D geometry supervision and often do not optimize for appearance.

2.3. Human Performance Capture and Rendering from Monocular Video

Traditional human performance capture and rendering techniques often rely on multi-view videos [53, 66] or depth cameras [36, 52, 69, 74] for human body geometry reconstruction and albedo map generation. With the advent of NeRFs, new methods have proposed modeling human geometry as radiance fields [10, 21, 23, 41, 42, 44, 45, 64, 78] or distance functions [59, 67], offering more flexibility and improved rendering quality.

The limitations of multi-view constraints have prompted researchers to explore human reconstruction from a single image or monocular video. Pioneering studies in this domain have achieved static clothed 3D human recovery [48,49,65], full body reconstruction [1, 17, 19, 70], and dynamic human modeling [21, 22, 41, 42, 46, 63, 71]. Specifically [20] explores explicit grid based methods for dynamic reconstruction, and struggle capturing pose dependent deformations resulting in non-uniform geometry. Despite their accomplishments, these methods tend to overfit to training data, leading to unwanted artifacts in novel views. Notably, recent efforts have started to address these issues by introducing motion priors [21, 63, 64, 75] to regularize the deformation. However, these methods primarily aim at rendering free-viewpoint human images, whereas our focus is on creating realistic reposed avatars and handling out-of-distribution poses.

3. Approach

Problem Statement. Our goal is to construct and control a detailed implicit neural avatar in a free view-point and arbitrary novel pose, only using monocular video and known 3D skeleton data.

Overview. We aim to achieve the above goal with the help of three components (1) a canonical representation of the actor (2) a deformation module based on forward skinning (3) a grid-based volumetric rendering with importance sampling via transmittance. The pipeline is illustrated in Fig. 2. To obtain the animatable model, our method first samples rays from world frame and deforms the points along those rays back to canonical frame via root finding, then query the color and signed distance functions (SDF) values in canonical space. Specifically, we formulate the problem as a pose conditioned implicit signed distance field and texture field in canonical space Sec. 3.4. The dynamically updated canonical human shape helps optimize the skinning fields. Unlike other methods we initialize the skinning weights in canonical space to allow pose generalization. *Canonical*

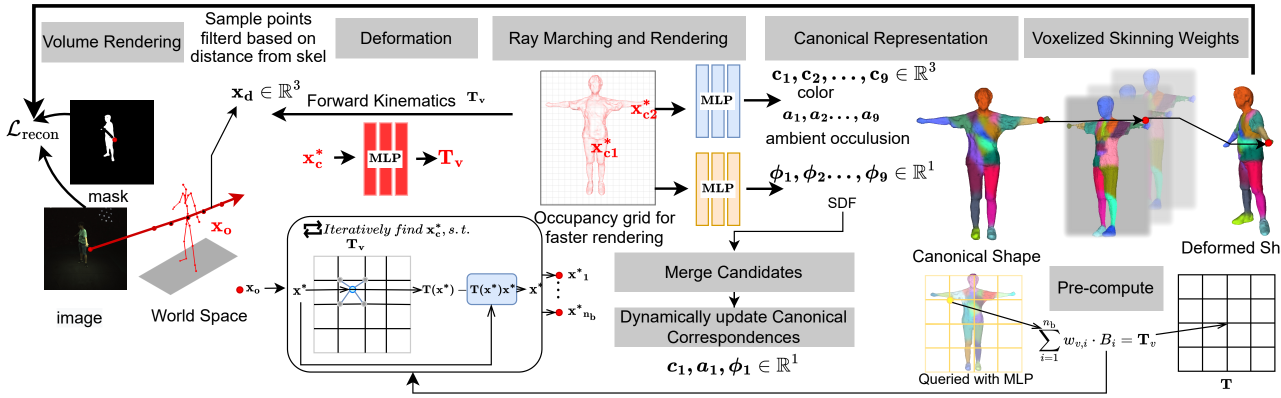
Volume Rendering
Sample points filterd based on distance from skel
Deformation
Ray Marching and Rendering
Canonical Representation
Voxelized Skinning Weights
$\mathcal{L}_{\text{recon}}$
mask
image
World Space
$\mathbf{x_d} \in \mathbb{R}^3$
Forward Kinematics
$\mathbf{T_v}$
$\mathbf{x_c^*}$
MLP
$\mathbf{T_v}$
$\mathbf{x_o}$
$\mathbf{x_o}$
Iteratively find $\mathbf{x_c^*}, s.t.$
$\mathbf{T_v}$
$\mathbf{x^*}$
$\mathbf{T(x^*)}$
$\mathbf{T(x^*)x^*}$
$\mathbf{x^*}$
$\mathbf{x^*}_1$
$\mathbf{x^*}_{n_b}$
$\mathbf{x^*_{c2}}$
$\mathbf{x^*_{c1}}$
Occupancy grid for faster rendering
MLP
MLP
$\mathbf{c_1, c_2, \ldots, c_9} \in \mathbb{R}^3$
color
$\boldsymbol{a_1, a_2 \ldots, a_9}$
ambient occulusion
$\boldsymbol{\phi_1, \phi_2 \ldots, \phi_9} \in \mathbb{R}^1$
SDF
Merge Candidates
Dynamically update Canonical Correspondences
$\boldsymbol{c_1, a_1, \phi_1} \in \mathbb{R}^1$
Canonical Shape
Deformed Shape
Pre-compute
Queried with MLP
$\sum_{i=1}^{n_b} w_{v,i} \cdot B_i = \mathbf{T}_v$
$\mathbf{T}$

Figure 2. **Method Overview.** In the Deformation module, the points along a ray in world space $\mathbf{x_o}$ are mapped to canonical space $\mathbf{x_c}$ by solving for the root of LBS blend weights. In canonical space, we use a color, SDF, and ambient occlusion networks to parameterize the appearance. Finally $F_{\Theta_{lbs}}$ allows learning the forward skinning weights, defined in this canaonical space.

Correspondence Search module maps points between observation space and canonical space and maintains 1-1 correspondence as detailed in Sec. 3.3. Details regarding the training objectives and the volume rendering process can be found in Section 3.6.

3.1. Preliminaries

Deformed and Canonical Spaces: We denote a sample along a ray in the observation space as $\mathbf{x_o} \in \mathbb{R}^3$, and a point in the canonical space as $\mathbf{x_c} \in \mathbb{R}^3$. It is important to note that the canonical space is independent of pose and maintains temporal consistency, as seen in [27, 75].

Networks and Parameters: Given a sequence of RGB frames of a human subject, along with their tracked skeleton, segmentation masks, and the camera parameters, following the TAVA methodology [27], we aim to learn an implicit representation of our human avatar through surface-based volume rendering [72]. The implicit neural avatar representation is dynamically updated in the canonical space. This representation is dynamically updated in the canonical space, where the color and SDF at point $\mathbf{x}_c$ are specifically modeled as:

$$F_{\Theta_{surf}} : (\mathbf{x}_c) \rightarrow (sdf, \mathbf{n}, \mathbf{feat}) \tag{1}$$

$$F_{\Theta_{rgb}} : (\mathbf{feat}, \mathbf{n}) \rightarrow (c, h) \tag{2}$$

$$F_{\Theta_a}(h, B) \rightarrow ao \tag{3}$$

$$F_{\Theta_{lbs}} : \mathbf{x_c} \rightarrow \mathbf{w}_v, \tag{4}$$

where $F_{\Theta_{surf}}$ is a coordinate-based Multilayer Perceptron (MLP) network. It takes a point $\mathbf{x_c} \in \mathbb{R}^3$ in canonical space as input and outputs its SDF $\in \mathbb{R}^1$, normal $\mathbf{n} \in \mathbb{R}^3$ which are spatial gradient of the SDF w.r.t. the points $\mathbf{x_c}$, and features $\mathbf{feat} : \mathbb{R}^3$. The $F_{\Theta_{rgb}}$ is a network which takes it $\mathbf{feat} \in \mathbb{R}^3$ and normal to return an intermediate activation $h \in \mathbb{R}^{256}$ and color $c \in \mathbb{R}^3$ of values in range $[0, 1]$. The F_{Θ_a} is a shading network, which takes in the intermediate activation h and body pose B to return a scalar that is used to compensate for ambient occlusion. $F_{\Theta_{lbs}}$ is also a MLP which learns the neural blend weights $\mathbf{w}$ in canonical space to enable forward skinning of the avatar $\mathbf{x_c} \rightarrow \mathbf{x_o}$, which describes how to animate it given a pose $B \in \mathbb{R}^{78}$, following Fast-SNARF [8] we "re-parameterize the skinning weight field w as a low-resolution voxel grid $\mathbf{w}_v$" for faster processing. For simplicity, we use the same notation w to represent voxel skinning weights. Consequently, we develop an end-to-end trainable model capable of rendering and reposing a human actor across a diverse range of poses and viewpoints.

3.2. Surface based volume rendering

To learn the dynamic avatar surface, we estimate the integral of samples within the volume of discrete samples similar to [33, 35] which are then subject to reconstruction losses Eq. (15). Following Volume Rendering of Neural Implicit Surfaces [73], we represent volume density σ as a transformed version of the SDF to the scene's surface obtained from Cumulative Distribution Function (Φ_β) of Laplace distribution's with zero mean and a scale of β:

$$\sigma(\mathbf{x_c}) = \alpha \Phi_\beta(-F_{\Theta_{surf}}(\mathbf{x_c})) \tag{5}$$

where α, $\beta > 0$ are learnable parameters. This hybrid representation, when integrated with an occupancy grid-based [26] ray sampling algorithm, minimizes floating artifacts and effectively decouples shape and texture in volume rendering. Additionally, this approach enables us to delineate the actor's geometry using the zero level set of $F_{\Theta_{surf}}$. The canonical shape $\mathcal{S}$ is represented as:

$$\mathcal{S} = \{ \mathbf{x}_c \mid F_{\Theta_{surf}}(\mathbf{x}_c) = 0 \} \tag{6}$$

In practice, we adopt the methodology of Li et al. [26], where each sample is represented as an interval along the ray. This approach allows us to decouple the sampling process from the differentiable computational graph, thereby facilitating the exclusion of empty rays and background pixels. Our sampling strategy employs a transmittance estimator, enabling us to directly compute the Cumulative Distribution Function (CDF) as $1-T(t)$. During the optimization process, the radiance field undergoes modifications between iterations. This necessitates dynamically updating the transmittance estimator at each step k:

$$F_{\Theta_{surf}} : T_{k-1} \rightarrow T_k. \tag{7}$$

This dynamic updating often poses challenges, as the constantly changing radiance field makes transmittance estimation more difficult. To mitigate this, our warm-up stage, described in Section 3.7, helps in regularizing and estimating the transmittance. Notably, our approach eliminates the need for costly numerical integration typically required to accumulate weights, as seen in TAVA [27] and NeRF [13,35]. The volume rendering for $\mathbf{N}$ points are done as follows:

$$C(\mathbf{r}) = \sum_{i=1}^{N} T_i \alpha_i \mathbf{c}_i, \tag{8}$$

$$\text{where, } T_i = \prod_{j=1}^{i-1} (1-\alpha_j)\,,\ \alpha_i = (1-\exp(\sigma_i \delta_i)), \tag{9}$$

Here, $\delta^{(i)}$, the distance between samples on ray $\mathbf{r}$, a ray sample's influence on pixel color $C(\mathbf{r})$ diminishes with low transmittance T_i. This T_i indicates the likelihood of preceding points being empty. The render weight of a sample is $w_i = T_i \alpha_i$, with α_i representing the non-emptiness probability of that voxel. The overall ray opacity is $\alpha(\mathbf{r}) = \sum_{i=1}^{N} w_i$.

3.3. Canonical Correspondence search

Traditional parametric human body models [3,18,32,39,43,68] use linear blend skinning (LBS) to deform a canonical surface according to rigid bone transformations and skinning weights. Following SNARF [9], the LBS weights are parameterized via $F_{\Theta_{lbs}}$ helps warping a deformed point $\mathbf{x_o}$ to the corresponding canonical point $\mathbf{x_c}$ as follows:

$$F_{\Theta_{lbs}}(\mathbf{x_c}, \boldsymbol{B}) = \mathbf{T}_v(\mathbf{x_c})\ ;\ w_v \cdot \mathbf{x_c} = \mathbf{x_o} \tag{10}$$

The process of warping points from $\mathbf{x_c} \rightarrow \mathbf{x_o}$ is defined as forward skinning. While, learning this voxelized weights field w_v which is parametrized by $F_{\Theta_{lbs}}$ we use the backward skinning function i.e. $\mathbf{x_o} \rightarrow \mathbf{x_c}$, which is implicitly defined as the solution to the equation:

$$F_{\Theta_{lbs}}(\mathbf{x_c}, \boldsymbol{B}) - \mathbf{x_o} = \mathbf{0}. \tag{11}$$

The Eq. (11) cannot be solved analytically. Therefore, we map the sampled points $\mathbf{x_o}$ to canonical space $\mathbf{x_c}^*$ by inverting the forward skinning function. This inversion is performed using iterative numerical approximation to estimate the true canonical correspondences. Following Fast-SNARF [8], the skinning weight field is parameterized as a voxel grid. For each pose, we first pre-compute the Linear Blend Skinning (LBS) for each grid point, generating a transformation field T_v. For each queried deformed point $\mathbf{x_o}$ in deformed space, we initialize its roots $\mathbf{x_c}^*$ and iteratively solve them such that it satisfies $\mathbf{T_v}(\mathbf{x_c}^*) \cdot \mathbf{x_c}^* = \mathbf{x_o}$. The vertex points of the skinning fields is updated by the $F_{\Theta_{lbs}}$ network. For other points, we employ tri-linear interpolation. Additionally the canonical correspondence module has Identity transform $I \in \mathbb{R}^{4\times4} : w_{bg} \cdot I$. This inclusion aids in the removal of background points, a feature also observed in HumanNeRF [64].

3.4. Building in Canonical Space

To construct canonical space, we sample pixels from the Monocular video and warp the samples from observation space ($\mathbf{x_o}$) along each pixel ray to canonical space following pose data $\mathbf{B}$, where we search for canonical correspondences using the root finding operation based on the inversion of Eq. (11). This yields multiple possible correspondences $\mathbf{x}_c^*$ on the canonical iso-surface for each $\mathbf{x_o} \xrightarrow{\text{r.f.}} \{\mathbf{x}^*_{c,1}, \mathbf{x}^*_{c,2}, ..., \mathbf{x}^*_{c,K}\}$ where $K \in joints$. Unlike static avatars, where we can directly query the color and SDF of $\mathbf{x_o}$ in the observation space. We dynamically update the parameters of surface representation and deformation networks in canonical space:

$$F_{\Theta_{surf}} : (\mathbf{x}^*_{c,i}) \rightarrow sdf^*_i. \tag{12}$$

We query the SDF values of these points and identify the true correspondences based on $\text{argmin}\ |F_{\Theta_{surf}}(\mathbf{x}^*_{c,i})| \rightarrow \mathbf{x}_c$. Subsequently, we utilize these surface points $\mathbf{x}_c$ to update the parameters of both $F_{\Theta_{surf}}$ and $F_{\Theta_{lbs}}$ networks, as detailed in the losses described in Sec. 3.6. The voxelized skinning weights w, parameterized by $F_{\Theta_{lbs}}$, are also jointly optimized. Since these weights are defined in the canonical space, they offer a temporally consistent representation that remains valid across a broad range of poses. This allows the model to animate the actor in novel poses.

3.5. Texture Fields

Once the canonical surface points $\mathbf{x}_c$ are obtained for each observation point $\mathbf{x}_o$, we proceed to update the color in the canonical space. Using $F_{\Theta_{rgb}}$, we query the color of these points. To obtain the final color, we also incorporate a scaling factor from ao, resulting in $c : c \cdot ao$. Subsequently, all the ray samples are aggregated as in Sec. 3.2 to compute the final rendered pixel value. This value serves as the basis for the reconstruction loss, as described in Equation 15.

3.6. Losses and Objectives

We use a two stage approach for optimization and train the model in a end-to-end fashion. The primary changes between the stages is the various losses that is used to bootstrap certain networks that help improve convergence speed.

Normal Consistency Loss. Since our model is focused on human actor, we exploit the canonical SMPL mesh. This loss primarily helps bootstrapping the canonical surface as seen in Fig. 4.

$$\mathcal{L}_{\text{n}}^i = ||\widehat{\mathbf{n}_c^{\text{smpl}}} - \mathbf{n}_{\mathbf{c}}^{\text{smpl}}||_2^2 \tag{13}$$

$$\widehat{\mathbf{n}_c^{\text{smpl}}} = \frac{\partial F_{\Theta_{surf}}(\mathbf{x}_c^{\text{smpl}}, B)}{\partial \mathbf{x}_c^{\text{smpl}}}$$

Bone Weight Loss. Considering that each point along a bone undergoes an identical transformation, we force the skinning weights $\mathbf{w}$ for samples $\bar{\mathbf{x}}_c$ situated on the bones should resemble one-hot vectors $\mathbf{w_{gt}}$. We sample $\bar{\mathbf{x}}_c$ points between joints for this purpose. The loss is defined as $\mathcal{L}_w = ||\mathbf{w}(\bar{\mathbf{x}}_c) - \mathbf{w_{gt}})||_2^2$. We observed that initializing this along with $\mathcal{L}_{\text{n}}^i$ helps in faster convergence of root-finding step as seen in [9]

Opacity Sparseness Regularization. following [15] we use L_{sparse} to bootstrap the ray opacity at the start of training to improve the transmittance estimation as well, which as seen in Sec. 3.2 influences the samples, $\mathcal{R}_{\text{empty}}$ is the empty rays.

$$\mathcal{L}_{\text{sparse}}^i = \frac{1}{|\mathcal{R}_{\text{empty}}^i|} \sum_{\mathbf{r} \in \mathcal{R}_{\text{empty}}^i} |\alpha(\mathbf{r})|. \tag{14}$$

Reconstruction Loss. The $\mathcal{L}_{\text{rgb}}^i$ for frame i is determined by computing the $L2$ distance between the rendered color $\hat{C}(\mathbf{r})$ and the actual pixel's RGB value $C(\mathbf{r})$:

$$\mathcal{L}_{\text{rgb}}^i = \frac{1}{|\mathcal{R}^i|} \sum_{\mathbf{r} \in \mathcal{R}^i} \left|\left|C(\mathbf{r}) - \hat{C}(\mathbf{r})\right|\right|_2^2. \tag{15}$$

Similarly we also obtain $\mathcal{L}_{\text{mask}}^i$ where, $\alpha(\mathbf{r})$ is the ground truth mask and $\widehat{\alpha}(\mathbf{r})$ rendered mask

$$\mathcal{L}_{\text{mask}}^i = \sum_{\mathbf{r} \in \mathcal{R}^i} \left|\left|\alpha(\mathbf{r}) - \widehat{\alpha}(\mathbf{r})\right|\right|_2^2. \tag{16}$$

Eikonal Loss. We follow the same implementation as in IGR [14] and [15], to ensure the gradient norms of the geometry network $F_{\Theta_{surf}}$ are regularized and to improve surface details.

Our final loss is: $\mathcal{L} = \mathcal{L}_{rgb}^i + \mathcal{L}_{mask}^i + \mathcal{L}_{\text{sparse}}^i + \mathcal{L}_{\text{eik}}^i + \mathcal{L}_{\text{n}}^i + \lambda\mathcal{L}_w$ where λ is set to 1.0 in warm-up and stage I while the bone weights are frozen in stage II in all our experiments.

3.7. Optimization and Sampling Strategy

- **Warm-up stage:** Initialize canonical SMPL based SDF and normal. This helps stabilize the transmittance estimation and increase converge speed. Loss at this stage is $\mathcal{L} = \mathcal{L}_{\text{eik}}^i + \lambda\mathcal{L}_w + \mathcal{L}_{\text{n}}^i$.
- **Stage I:** Resume training with randomized ray sampling. For all our experiments, we initialize an occupancy grid of $\mathbb{R} \in 112^3$ with bounds $[-10, 10]$ along all axes. We randomly sample 1024 rays from each training image. This stage primarily aids in updating the canonical surface and in the learning of the deformation weights. The loss at this stage is $\mathcal{L} = \mathcal{L}_{rgb}^i + \mathcal{L}_{mask}^i + \mathcal{L}_{\text{sparse}}^i + \mathcal{L}_{\text{eik}}^i + \lambda\mathcal{L}_w$. Note that we are no longer conditioning on the SMPL mesh.
- **Stage II:** We observed that the skinning weights are learned rapidly due to voxel re-parameterization. At this stage, we freeze the $F_{\Theta_{lbs}}$ network and adopt a patch-based sampling approach, akin to that used in [51, 64]. This method effectively assists the $F_{\Theta_{rgb}}$ network in learning textures, thereby enhancing both texture and geometry capture. The loss function at this stage is $\mathcal{L} = \mathcal{L}_{rgb}^i + \mathcal{L}_{mask}^i + \mathcal{L}_{\text{sparse}}^i + \mathcal{L}_{\text{eik}}^i + \mathcal{L}_{\text{LPIPS}}$. The implementation of $\mathcal{L}_{\text{LPIPS}}$ aligns with the method described in [76]. For patch-based sampling, we feed in 3 patches of 32×32 pixels each and concurrently increase the learning rate from 5×10^{-6} to 5×10^{-5}.

Occupancy Grids. In terms of implementation we follow NerfAcc [26] which suggests that inverse sampling of CDF is equivalent to inverse sampling of the transmittance $T(t)$. We can compute the CDF directly using $1 - T(t)$ as seen in Eq. (8) and Eq. (9), instead of the computationally expensive integral $\int_{t_n}^{t} T(v)\sigma(v)\,dv$, which is the standard implementation adopted by many popular codebases [4, 35].

Conditioning on Canonical SMPL Mesh. We initialize the SDF network $F_{\Theta_{surf}}$ using the standard SMPL model, following the normal loss Eq. (13). This initialization is performed during the warm-up stage to estimate the human T-pose. Our empirical observations indicate that this approach facilitates more efficient network convergence and effectively eliminates incorrect correspondences, as discussed Sec. 3.3. Consequently, this initialization strategy also aids in optimizing $F_{\Theta_{lbs}}$ network. It is to be noted that there is always an inherent possibility that the root-finding process may not converge.. In practice, we initially filter out canonical points based on SDF query and the function $argmin|F_{\Theta_{surf}}(x_{c,i}^*)| \rightarrow \mathbf{x_c}$. This approach may occasionally lead to a minor fraction of root-finding solutions failing. For these points, the rendering step interpolates the nearest available value.

4. Experiments

4.1. Dataset and Preprocessing

ZJU-MoCap dataset [45]: We use subjects 313, 387 and 393 from the ZJU-MoCap dataset, which contain calibrated and accurately annotated segmentation masks, and camera transformations. Therefore, we train our model using the provided ground truths, while restricting ourselves to a single camera view.

Human Motion Diffusion Model (MDM) [58]: We employ the MDM approach to acquire text-guided action sequences, which allows us to assess the adaptability of our model to extreme variations in poses.

4.2. Baseline Experiments

We compare our method with (1) Neural Body [45], which adopts an implicit neural representations with structured latent codes; (2) TAVA [27] (Template-free Animatable Volumetric Actors), which also utilizes a forward skinning module and has a coarse to fine sampling strategy as in NeRF [35]. We use the same splits for all experiments. For more details on pre-processing and implementation refer supplementary material.

Evaluation metrics. To evaluate the quality of rendered images, we use PSNR and SSIM [62]. Since PSNR measures the magnitude of pixel-level differences between the reference and generated images, smoother images tend to have a higher PSNR value [64, 77]. To complement this, we calculate the SSIM as well to ensure that the resulting images are of high fidelity to human visual perception. These two metrics combined provides a more holistic understanding of the performance of the models.

Comparison settings. We compare our method against Neural Body and TAVA in terms of both novel view and novel pose synthesis. We use the validation splits to evaluate all the models as briefed in Sec. 4. The quantitative evaluation is reported in Tab. 1 and the qualitative results are presented in Fig. 3.

4.3. Results and Discussion

Baseline Comparison We find that our method consistently outperforms both Neural Body [45] and TAVA [27] across a variety of settings Tab. 1. While TAVA's PSNR metrics are occasionally comparable to ours, a qualitative analysis reveals significant shortcomings. As illustrated in Fig. 3, the textures and colors rendered by TAVA appear overly smooth, compromising the realism of the reconstructed subject. In contrast, Neural Body performs poorly in novel-view settings. This is attributed to the entanglement of texture and geometry within the latent codes, making it less robust to changes in viewpoint. It is worth noting that the PSNR metric tends to favor smoother images, which explains why TAVA may score comparably in some cases yet still produce less realistic results.

4.3.1 Ablation experiments

Forward-Skinning Module (w/o fs): We substitute the forward deformation module with a MLP designed to model skinning weights. However, this approach encounters difficulties in decoupling deformation from geometry, resulting in mesh collapse. Consequently, the performance metrics dramatically deteriorate, as indicated in row 1 of Tab. 2. The SDF surface get's unstable when the importance base sampling is poor.

Occupancy Grid (w/o grid): In contrast to the importance-based sampling discussed in Sec. 3.6, we also experimented with hierarchical sampling. This alternative was not only memory-intensive but also problematic in terms of mesh quality. Specifically, it produced numerous floaters and introduced extraneous surface components, as illustrated in the bottom left corner of Fig. 6. Additionally, these floaters often obscured missing parts in the mesh, further complicating the analysis. This causes the model's score to falter as seen in row 2 of Tab. 2. Additionally the occupancy grid based ray-sampler enables faster rendering and importance based Sampling via Transmittance.

Ambient Occlusion (w/o ao): We conducted experiments in which we disabled the ambient occlusion or shading MLP, which is parameterized based on the body pose B. In the absence of the ao scaling factor, the resulting images exhibit darker tones than usual, attributable to self-body occlusion effects. Importantly, since this layer exclusively influences shading and not the geometric properties, the surface structure remains unaffected. Consequently, the results display the least deviation among all the variants tested, as evidenced in row 3 of Tab. 2 and Fig. 7.

Training Progression: We opted to train the $F_{\Theta_{surf}}$ network without conditioning it on the canonical SMPL mesh and skipped the warm-up stages. This approach resulted in missing parts in the model as in the images on row 1 in Fig. 4, particularly in areas with large deformations (i.e., arms). Solving for the numerical root from scratch becomes a formidable challenge in this setup, as each point in the observation space x_o can have up to nine different canonical initialization. This makes it a difficult problem to solve without prior knowledge of the surface, often causing the Broyden iterations to terminate prematurely and fail to converge with a root. When a warm-up stage is employed, the SDF surface serves as a guiding mechanism for the $F_{\Theta_{lbs}}$ network. Conversely, the skinning process also updates

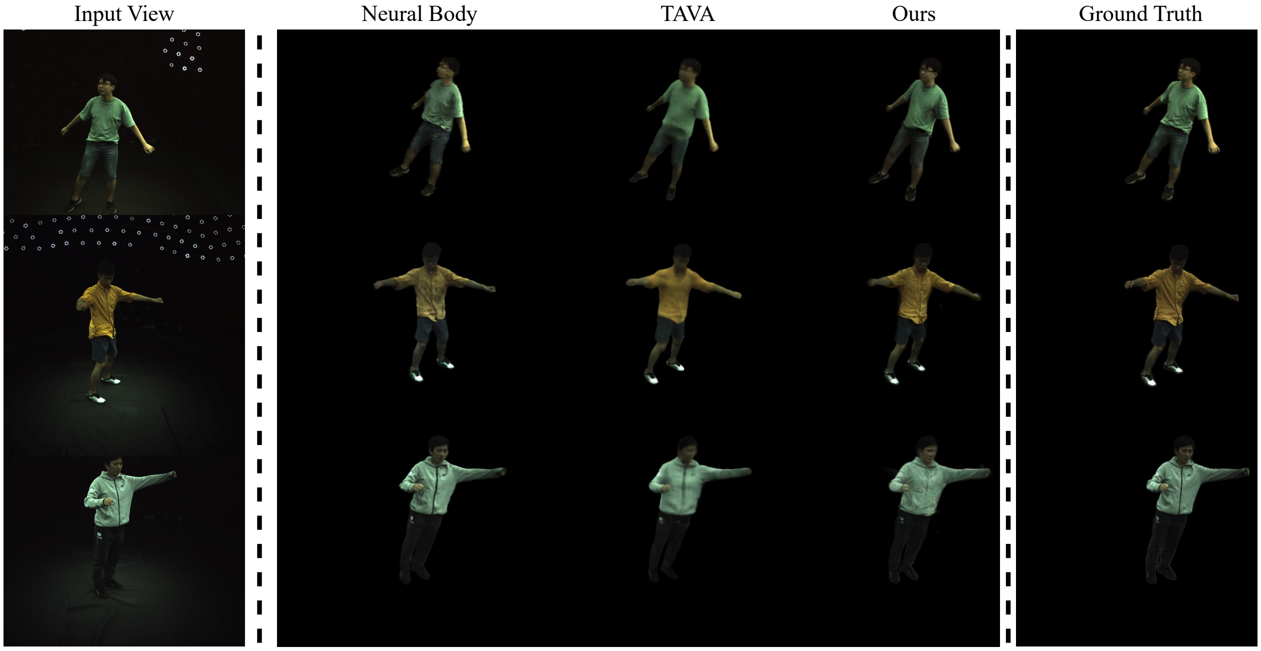

Figure 3. **Qualitative results under novel view setting for ZJU-MoCap Dataset.** Comparisons of novel view synthesis with other baseline methods for ZJU-Mocap. Results show that our method produces realistic images. The quantitative data of the same is present in Tab. 1

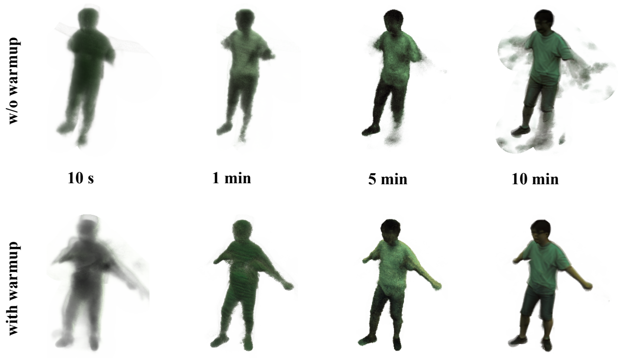

Figure 4. **Training Progression.** We present rendered images at various stages of the training process. The use of Canonical SDF initialization not only accelerates the convergence but also effectively preserves regions of the model that undergo significant deformations.

	Novel-view		Novel-pose (ind)		Novel-pose (ood)	
	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑
Subject 313						
NeuralBody [45]	29.03	0.96	29.02	0.96	29.03	0.96
TAVA [27]	33.48	0.98	33.07	0.98	30.13	0.97
Ours	**34.97**	**0.98**	**33.08**	**0.98**	**30.94**	**0.97**
Subject 387						
NeuralBody [45]	26.81	0.95	26.77	0.95	26.79	0.95
TAVA [27]	30.41	0.97	31.31	0.97	29.48	0.96
Ours	**32.79**	**0.98**	**31.56**	**0.97**	**29.62**	**0.96**
Subject 393						
NeuralBody [45]	27.45	0.96	27.44	0.95	27.45	0.95
TAVA [27]	31.64	0.97	**32.75**	0.97	29.96	0.97
Ours	**33.13**	**0.98**	32.32	**0.98**	**30.42**	**0.97**

Table 1. *Per-Subject Comparisons on the ZJU-MoCap Dataset Monocular camera setup against NeuralBody and TAVA*

$F_{\Theta_{surf}}$ network as seen from the progression images in row 2 Fig. 4. The joint optimization of these networks not only accelerates convergence but also improves surface estimation. These refined surfaces can be further extracted from the zero-level set of the SDF, as described in Eq. (6).

Normal Consistency Losses in Warm-up: We conducted experiments where we removed the normal consistency loss, as defined in Eq. (13), during the warm-up stage. Although the SDF was capable of capturing major components, it fell short in preserving finer texture details and facial features such as the eyes and mouth. The inadequacy in capturing these details is evident in Fig. 5.

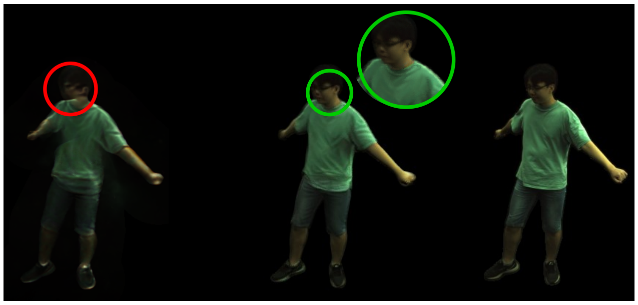

Figure 5. **Qualitative results of using normal consistency loss to bootstrap SDF initialization.** The model without normal consistency loss fails to capture the texture of the person's face.

	Novel-view		Novel-pose (ind)		Novel-pose (ood)	
	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑
Subject 313						
w/o fs	26.69	0.94	24.58	0.91	25.20	0.90
w/o grid	28.96	0.97	29.04	0.97	27.94	0.95
w/o ao	29.66	0.97	29.41	0.97	28.76	0.96
Full	**34.97**	**0.98**	**33.08**	**0.98**	**30.94**	**0.97**
Subject 387						
w/o fs	24.26	0.87	24.64	0.87	26.23	0.88
w/o grid	28.24	0.94	28.62	0.95	27.84	0.95
w/o ao	30.89	0.97	31.05	0.97	29.48	0.96
Full	**32.79**	**0.98**	**31.56**	**0.97**	**29.62**	**0.96**
Subject 393						
w/o fs	26.30	0.94	25.92	0.95	25.70	0.92
w/o grid	29.05	0.96	27.89	0.95	27.43	0.93
w/o ao	31.62	0.97	31.18	0.97	28.06	0.97
Full	**33.13**	**0.98**	**32.32**	**0.98**	**30.42**	**0.97**

Table 2. *Ablation experiments on the ZJU-MoCap subjects*

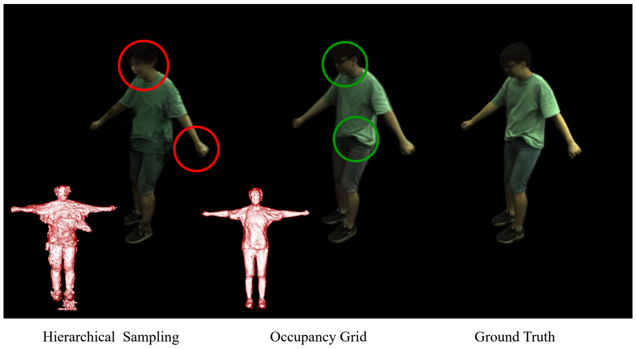

Figure 6. **Qualitative results from using both hierarchical sampling and occupancy grid-based sampling.** The occupancy grid-based sampling approach primarily helps eliminate floaters and preserve the surface structure.

Training Speed: By employing voxelized skinning weights and leveraging the memory-efficient rendering strategy proposed in [26], we achieve faster convergence rates for our model. Additionally, initializing the model with a SDF further accelerates the training process. As evidenced by Fig. 8, the PSNR metric for our method reaches saturation after approximately 10,000 steps, indicating rapid and efficient training.

5. Conclusion

We propose AvatarOne, which addresses the problem of jointly optimizing for iso-surface and canonical correspondences, resulting in fast, view-consistent and re-posable human avatars. We propose a novel initialization and bootstrap losses for surface and transmittance estimation which helps us to exploit the voxelized skinning weights. Furthermore, we employ a holistic sampling strategy that injects a useful geometrical inductive bias into the neural volume rendering , which helps us obtain more accurate ray radiance approximations. To evaluate our approach, we validate our methodology by animating our avatar with challenging poses generated through text-prompts, based on the Human Motion Diffusion Model [58]. The results show that the model can handle unseen poses, generating realistic images and avatars.

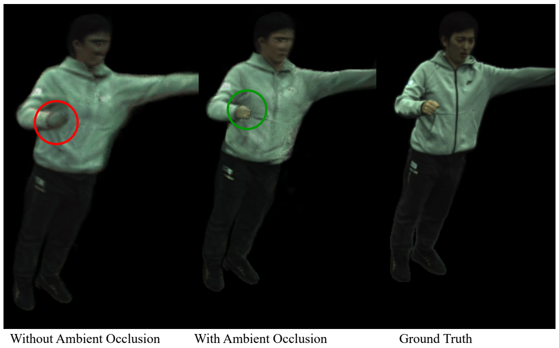

Figure 7. **Qualitative results from without and with ambient occlusions.** The model without using ambient occlusions fails at capturing the color of the hand as shown in the red circle.

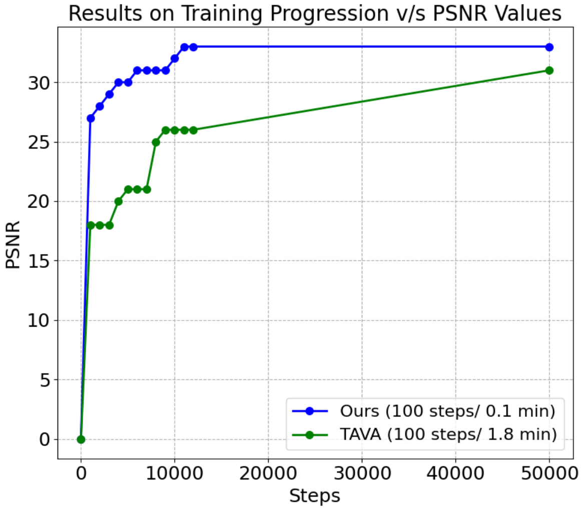

Figure 8. **Comparative Analysis of Training Progression.** Our method outperforms TAVA while being almost **18x** faster on ZJU-Subject 313.

Limitations and Future work. 1. The quality of the rendered results is dependent on the accuracy of pose and mask annotations. Future work could investigate techniques for real-time pose and mask correction during the training process. **2.** The proposed method is trained in a subject-specific manner. Investigating approaches for transitioning from one-shot rendering to a more generalizable model, particularly by leveraging a wide range of pre-trained modules presents a valuable direction for future research.

References

[1] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Video based reconstruction of 3d people models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8387–8397, 2018. 2

[2] D. Anguelov, P. Srinivasan, D. Koller, S. Thrun, J. Rodgers, and J. Davis. Scape: shape completion and animation of people. In *ACM SIGGRAPH 2005 Papers*, pages 408–416, 2005. 2

[3] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. Scape: shape completion and animation of people. *ACM Transasctions Graphics*, 24, 2005. 4

[4] Jonathan T. Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P. Srinivasan. Mip-NeRF: A multiscale representation for anti-aliasing neural radiance fields. *ICCV*, 2021. 5

[5] G. Borshukov, D. Piponi, O. Larsen, J.P. Lewis, and C. Tempelaar-Lietz. Universal capture-image-based facial animation for "the matrix reloaded". In *SIGGRAPH 2005 Courses*, 2005. 2

[6] Joel Carranza, Christian Theobalt, Marcus A Magnor, and Hans-Peter Seidel. Free-viewpoint video of human actors. *ACM transactions on graphics (TOG)*, 2003. 2

[7] Dan Casas, Marco Volino, John Collomosse, and Adrian Hilton. 4D video textures for interactive character appearance. *Computer Graphics Forum*, 2014. 2

[8] Xu Chen, Tianjian Jiang, Jie Song, Max Rietmann, Andreas Geiger, Michael J. Black, and Otmar Hilliges. Fast-snarf: A fast deformer for articulated neural fields. *Pattern Analysis and Machine Intelligence (PAMI)*, 2023. 3, 4

[9] Xu Chen, Yufeng Zheng, Michael J Black, Otmar Hilliges, and Andreas Geiger. Snarf: Differentiable forward skinning for animating non-rigid neural implicit shapes. In *ICCV*, 2021. 1, 2, 4, 5

[10] Wei Cheng, Su Xu, Jingtan Piao, Chen Qian, Wayne Wu, Kwan-Yee Lin, and Hongsheng Li. Generalizable neural performer: Learning robust radiance fields for human novel view synthesis. *arXiv preprint arXiv:2204.11798*, 2022. 2

[11] Alvaro Collet, Ming Chuang, Pat Sweeney, Don Gillett, Dennis Evseev, David Calabrese, Hugues Hoppe, Adam Kirk, and Steve Sullivan. High-quality streamable free-viewpoint video. *ACM Transactions on Graphics (ToG)*, 2015. 2

[12] B. Deng, J.P. Lewis, T. Jeruzalski, G. Pons-Moll, G. Hinton, M. Norouzi, and A. Tagliasacchi. Nasa neural articulated shape approximation. In *European Conference on Computer Vision*, pages 612–628. Springer, 2020. 2

[13] Thibaud Ehret, Roger Marí, and Gabriele Facciolo. Regularization of nerfs using differential geometry, 2022. 4

[14] Amos Gropp, Lior Yariv, Niv Haim, Matan Atzmon, and Yaron Lipman. Implicit geometric regularization for learning shapes. *arXiv preprint arXiv:2002.10099*, 2020. 5

[15] Chen Guo, Tianjian Jiang, Xu Chen, Jie Song, and Otmar Hilliges. Vid2avatar: 3d avatar reconstruction from videos in the wild via self-supervised scene decomposition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023. 5

[16] Kaiwen Guo, Peter Lincoln, Philip Davidson, Jay Busch, Xueming Yu, Matt Whalen, Geoff Harvey, Sergio Orts-Escolano, Rohit Pandey, Jason Dourgarian, et al. The relightables: Volumetric performance capture of humans with realistic relighting. *ACM Transactions on Graphics (TOG)*, 2019. 2

[17] Marc Habermann, Weipeng Xu, Michael Zollhofer, Gerard Pons-Moll, and Christian Theobalt. Deepcap: Monocular human performance capture using weak supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5052–5063, 2020. 2

[18] N. Hasler, C. Stoll, M. Sunkel, B. Rosenhahn, and H.-P. Seidel. A Statistical Model of Human Pose and Body Shape. *Computer Graphics Forum*, 28:337–346, 2009. 4

[19] Boyi Jiang, Yang Hong, Hujun Bao, and Juyong Zhang. Selfrecon: Self reconstruction your digital avatar from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5605–5615, 2022. 2

[20] Tianjian Jiang, Xu Chen, Jie Song, and Otmar Hilliges. Instantavatar: Learning avatars from monocular video in 60 seconds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16922–16932, June 2023. 2

[21] Wei Jiang, Kwang Moo Yi, Golnoosh Samei, Oncel Tuzel, and Anurag Ranjan. Neuman: Neural human radiance field from a single video. *arXiv preprint arXiv:2203.12575*, 2022. 1, 2

[22] Tianshu Kuai, Akash Karthikeyan, Yash Kant, Ashkan Mirzaei, and Igor Gilitschenski. Camm: Building category-agnostic and animatable 3d models from monocular videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 6586–6596, June 2023. 2

[23] Youngjoong Kwon, Dahun Kim, Duygu Ceylan, and Henry Fuchs. Neural human performer: Learning generalizable radiance fields for human performance rendering. *Advances in Neural Information Processing Systems*, 34:24741–24752, 2021. 2

[24] H. Li, L. Luo, D. Vlasic, P. Peers, J. Popović, M. Pauly, and S. Rusinkiewicz. Temporally coherent completion of dynamic shapes. *ACM Transactions on Graphics (TOG)*, 31(1):1–11, 2012. 2

[25] K. Li, J. Yang, L. Liu, R. Boulic, Y.K. Lai, Y. Liu, Y. Li, and E. Molla. Spa: Sparse photorealistic animation using a single rgb-d camera. *Transactions on Circuits and Systems for Video Technology*, 2016. 2

[26] Ruilong Li, Hang Gao, Matthew Tancik, and Angjoo Kanazawa. Nerfacc: Efficient sampling accelerates nerfs. *arXiv preprint arXiv:2305.04966*, 2023. 3, 5, 8

[27] Ruilong Li, Julian Tanke, Minh Vo, Michael Zollhofer, Jurgen Gall, Angjoo Kanazawa, and Christoph Lassner. Tava: Template-free animatable volumetric actors. *arXiv preprint arXiv:2206.08929*, 2022. 3, 4, 6, 7

[28] Zhengqi Li, Simon Niklaus, Noah Snavely, and Oliver Wang. Neural scene flow fields for space-time view synthesis of dynamic scenes. *CVPR*, 2021. 2

[29] Lingjie Liu, Marc Habermann, Viktor Rudnev, Kripasindhu Sarkar, Jiatao Gu, and Christian Theobalt. Neural actor: Neural free-view synthesis of human actors with pose control. *ACM Trans. Graph.(ACM SIGGRAPH Asia)*, 2021. 2

[30] Stephen Lombardi, Tomas Simon, Jason Saragih, Gabriel Schwartz, Andreas Lehrmann, and Yaser Sheikh. Neural volumes: Learning dynamic renderable volumes from images. *ACM Transactions on Graphics (TOG)*, 2019. 2

[31] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. SMPL: A skinned multi-person linear model. *ACM transactions on graphics (TOG)*, 2015. 1, 2

[32] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Transasctions Graphics*, 34(6), 2015. 4

[33] Nelson Max. Optical models for direct volume rendering. *IEEE Transactions on Visualization and Computer Graphics*, 1995. 3

[34] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *CVPR*, pages 4455–4465, 2019. 2

[35] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. *ECCV*, 2020. 2, 3, 4, 5, 6

[36] Richard A Newcombe, Dieter Fox, and Steven M Seitz. Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 343–352, 2015. 2

[37] M. Niemeyer, L. Mescheder, M. Oechsle, and A. Geiger. Occupancy flow: 4d reconstruction by learning particle dynamics. In *International Conference on Computer Vision*, 2019. 2

[38] A. Noguchi, X. Sun, S. Lin, and T. Harada. Neural articulated radiance field. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5762–5772, 2021. 2

[39] Ahmed A. A. Osman, Timo Bolkart, and Michael J. Black. Star: Sparse trained articulated human body regressor. In *Proc. of ECCV*, 2020. 4

[40] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proc. of CVPR*, 2019. 2

[41] Keunhong Park, Utkarsh Sinha, Jonathan T. Barron, Sofien Bouaziz, Dan B Goldman, Steven M. Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. *ICCV*, 2021. 2

[42] Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T. Barron, Sofien Bouaziz, Dan B Goldman, Ricardo Martin-Brualla, and Steven M. Seitz. HyperNeRF: A higher-dimensional representation for topologically varying neural radiance fields. *SIGGRAPH Asia*, 2021. 2

[43] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proc. of CVPR*, 2019. 4

[44] Sida Peng, Junting Dong, Qianqian Wang, Shangzhan Zhang, Qing Shuai, Xiaowei Zhou, and Hujun Bao. Animatable neural radiance fields for modeling dynamic human bodies. *ICCV*, 2021. 1, 2

[45] Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. *CVPR*, 2021. 2, 6, 7

[46] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-NeRF: Neural radiance fields for dynamic scenes. *CVPR*, 2020. 2

[47] A. Pumarola, E. Corona, G. Pons-Moll, and F. Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. In *Conference on Computer Vision and Pattern Recognition*, 2021. 2

[48] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2304–2314, 2019. 2

[49] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 84–93, 2020. 2

[50] Shunsuke Saito, Jinlong Yang, Qianli Ma, and Michael J Black. Scanimate: Weakly supervised learning of skinned clothed avatar networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2886–2897, 2021. 2

[51] Katja Schwarz, Yiyi Liao, Michael Niemeyer, and Andreas Geiger. GRAF: Generative radiance fields for 3d-aware image synthesis. *NeurIPS*, 2020. 5

[52] Ari Shapiro, Andrew Feng, Ruizhe Wang, Hao Li, Mark Bolas, Gerard Medioni, and Evan Suma. Rapid avatar capture and simulation using commodity depth sensors. *Computer Animation and Virtual Worlds*, 25(3-4):201–211, 2014. 2

[53] Leonid Sigal, Alexandru O Balan, and Michael J Black. Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *International journal of computer vision*, 87(1):4–27, 2010. 2

[54] Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. Scene representation networks: continuous 3D-structure-aware neural scene representations. *NeurIPS*, 2019. 2

[55] J. Starck and A. Hilton. Surface capture for performance-based animation. *IEEE Computer Graphics and Applications*, 27(3):21–31, 2007. 2

[56] S.Y. Su, F. Yu, M. Zollhöfer, and H. Rhodin. A-nerf: Articulated neural radiance fields for learning human shape,

appearance, and pose. In *Advances in Neural Information Processing Systems*, volume 34, 2021. 2

[57] Gusi Te, Xiu Li, Xiao Li, Jinglu Wang, Wei Hu, and Yan Lu. Neural capture of animatable 3d human from monocular video. *arXiv preprint arXiv:2208.08728*, 2022. 1

[58] Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *The Eleventh International Conference on Learning Representations*, 2023. 6, 8

[59] Garvita Tiwari, Dimitrije Antić, Jan Eric Lenssen, Nikolaos Sarafianos, Tony Tung, and Gerard Pons-Moll. Pose-ndf: Modeling human pose manifolds with neural distance fields. In *European Conference on Computer Vision*, pages 572–589. Springer, 2022. 2

[60] Edgar Tretschk, Ayush Tewari, Vladislav Golyanik, Michael Zollhöfer, Christoph Lassner, and Christian Theobalt. Non-rigid neural radiance fields: Reconstruction and novel view synthesis of a dynamic scene from monocular video. *ICCV*, 2021. 2

[61] M. Volino, D. Casas, J.P. Collomosse, and A. Hilton. Optimal representation of multi-view video. In *British Machine Vision Conference*, 2014. 2

[62] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 6

[63] Chung-Yi Weng, Brian Curless, and Ira Kemelmacher-Shlizerman. Vid2Actor: Free-viewpoint animatable person synthesis from video in the wild. *arXiv:2012.12884*, 2020. 1, 2

[64] Chung-Yi Weng, Brian Curless, Pratul P Srinivasan, Jonathan T Barron, and Ira Kemelmacher-Shlizerman. Humannerf: Free-viewpoint rendering of moving people from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16210–16220, 2022. 1, 2, 4, 5, 6

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

[66] Feng Xu, Yebin Liu, Carsten Stoll, James Tompkin, Gaurav Bharaj, Qionghai Dai, Hans-Peter Seidel, Jan Kautz, and Christian Theobalt. Video-based characters: creating new human performances from a multi-view video database. *SIGGRAPH*, 2011. 2

[67] Hongyi Xu, Thiemo Alldieck, and Cristian Sminchisescu. H-nerf: Neural radiance fields for rendering and temporal reconstruction of humans in motion. *Advances in Neural Information Processing Systems*, 34, 2021. 2

[68] Hongyi Xu, Eduard Gabriel Bazavan, Andrei Zanfir, William T. Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Ghum & ghuml: Generative 3d human shape and articulated pose models. In *Proc. of CVPR*, 2020. 4

[69] Lan Xu, Wei Cheng, Kaiwen Guo, Lei Han, Yebin Liu, and Lu Fang. Flyfusion: Realtime dynamic scene reconstruction using a flying depth camera. *IEEE transactions on visualization and computer graphics*, 27(1):68–82, 2019. 2

[70] Weipeng Xu, Avishek Chatterjee, Michael Zollhöfer, Helge Rhodin, Dushyant Mehta, Hans-Peter Seidel, and Christian Theobalt. Monoperfcap: Human performance capture from monocular video. *ACM Transactions on Graphics (ToG)*, 37(2):1–15, 2018. 2

[71] Gengshan Yang, Minh Vo, Natalia Neverova, Deva Ramanan, Andrea Vedaldi, and Hanbyul Joo. Banmo: Building animatable 3d neural models from many casual videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2863–2873, June 2022. 2

[72] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021. 3

[73] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. In *Proc. of NeurIPS*, 2021. 3

[74] Tao Yu, Zerong Zheng, Kaiwen Guo, Jianhui Zhao, Qionghai Dai, Hao Li, Gerard Pons-Moll, and Yebin Liu. Doublefusion: Real-time capture of human performances with inner body shapes from a single depth sensor. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7287–7296, 2018. 2

[75] Zhengming Yu, Wei Cheng, Xian Liu, Wayne Wu, and Kwan-Yee Lin. Monohuman: Animatable human neural field from monocular video. *CVPR*, 2023. 2, 3

[76] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018. 5

[77] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. *CVPR*, 2018. 6

[78] Fuqiang Zhao, Wei Yang, Jiakai Zhang, Pei Lin, Yingliang Zhang, Jingyi Yu, and Lan Xu. Humannerf: Generalizable neural human radiance field from sparse inputs. *arXiv preprint arXiv:2112.02789*, 2021. 2