SelfRecon: Self Reconstruction Your Digital Avatar from Monocular Video

Boyi Jiang[1,2] Yang Hong[1] Hujun Bao[3] Juyong Zhang[1*]

[1]University of Science and Technology of China [2]Image Derivative Inc [3]Zhejiang University

Abstract

We propose SelfRecon, a clothed human body reconstruction method that combines implicit and explicit representations to recover space-time coherent geometries from a monocular self-rotating human video. Explicit methods require a predefined template mesh for a given sequence, while the template is hard to acquire for a specific subject. Meanwhile, the fixed topology limits the reconstruction accuracy and clothing types. Implicit representation supports arbitrary topology and can represent high-fidelity geometry shapes due to its continuous nature. However, it is difficult to integrate multi-frame information to produce a consistent registration sequence for downstream applications. We propose to combine the advantages of both representations. We utilize differential mask loss of the explicit mesh to obtain the coherent overall shape, while the details on the implicit surface are refined with the differentiable neural rendering. Meanwhile, the explicit mesh is updated periodically to adjust its topology changes, and a consistency loss is designed to match both representations. Compared with existing methods, SelfRecon can produce high-fidelity surfaces for arbitrary clothed humans with self-supervised optimization. Extensive experimental results demonstrate its effectiveness on real captured monocular videos. The source code is available at https://github.com/jby1993/SelfReconCode.

1. Introduction

Clothed body reconstruction has been an important and challenging research topic in the community for years. In the film and gaming industry, high-fidelity human reconstruction usually requires pre-captured templates, multi-camera systems, controlled studios, and long-term works of talented artists. However, these requirements exceed the application scenarios of general customers, such as personalized avatars for telepresence, AR/VR, anthropometry, and virtual try-on, etc. Therefore, directly reconstruction high-fidelity digital avatar from monocular video will have significant practical application value.

The state-of-the-art marker-less monocular human performance capture approaches [17, 18, 52] are mainly designed based on explicit mesh representation. They require actor-specific rigged templates and utilize detected 2D/3D joints and silhouettes to estimate per frame's posture and non-rigid deformation. DeepCap [18] additionally uses multi-view information during training to resolve deep ambiguity and improve tracking accuracy for monocular inference. The explicit representation has some advantages, including space-time coherence and compatibility with existing graphics control pipelines, like texture editing and reposing. Moreover, skinning deformation is suitable under this paradigm to model the body's large-scale articulated deformations. However, actor-specific templates limit the extension of these methods to unseen human sequences. For videos of self-rotation humans under rough A-pose, VideoAvatar [2] can estimate general clothed humans with the SMPL+D parametric representation [1–3,6,31,49], while it can not recover folds and loose clothing, like skirts.

Recently, some neural implicit representation based monocular human reconstruction approaches have demonstrated compelling results [10, 20, 21, 23, 41, 42, 46, 50, 55, 56]. These methods can handle various topologies, and thus can represent various clothing and hairstyles. However, they require high-quality 3D data for supervision, and they only reconstruct for a specific frame and can not keep the space-time coherence of surface vertices for the whole sequence. A simple solution to guarantee the coherence and correct body structure is to maintain an implicit template surface in the canonical space, and then utilize backward deformation fields to map current points to canonical space to assist their implicit function queries. The backward deformation strategy has been widely applied recently and works well for small-scale deformations [29,36,40,45]. However, it is not very suitable for articulated skinning deformation due to its irreversibility in some parts of current space [11, 24]. To this end, technologies such as pose-related skinning weights prediction [24,38,47] and specific inverse articulated deformation design [13] are proposed at the cost of high complexity and poor generalization.

In this work, we propose SelfRecon, which combines the explicit and implicit representations together to reconstruct

*Corresponding author

high-fidelity digital avatar from a monocular video. Specifically, SelfRecon utilizes a learnable signed distance field (SDF) rather than a template with fixed topology to represent the canonical shape. To improve the generalization of the deformation and reduce the optimization difficulty, we adopt the forward deformation to map canonical points to the current frame space [11, 54]. During optimization, we periodically extract the explicit canonical mesh and warp it to each frame with the deformation fields. For these meshes, we utilize mask loss and smooth constraints to recover the overall shape. For the implicit part, a differential formulation is designed to intersect the deformed surface and follow IDR's neural rendering [53] to refine the geometry. A consistency loss is designed to match both geometric representations as close as possible.

SelfRecon alleviates the dependence on actor-specific templates and extracts a space-time coherent mesh sequence from a monocular video. Extensive evaluations on self-rotating human videos demonstrate that it outperforms existing methods. We believe that SelfRecon will inspire more studies on combining implicit and explicit representations for 3D reconstruction for articulated object.

2. Related Work

Implicit Human Reconstruction. PIFu [41] adopts a deep network to extract image features and concatenates pixel's feature and its corresponding 3D point depth information as the input of a Multi-Layer Perceptron (MLP) to obtain high-fidelity 3D clothed human occupancy field. However, it may generate incorrect body structures for humans under challenging poses. StereoPIFu [22] aims at binocular images, utilizes volume alignment feature and predicted high-precision depth to guide implicit function prediction, can effectively alleviate the depth ambiguity and restore absolute scale information. PIFuHD [42] utilizes higher resolution features and predicted normal information to refine the geometric details of PIFu. PaMIR [55] utilizes parameterized human body to decrease the influence of deep ambiguity in implicit function training, reduces the occurrence of abnormal human body structure, and improves the reconstruction accuracy. These methods train an MLP to represent the human's implicit geometry from single or several images and achieve impressive results. However, they require the corresponding high-quality 3D data of color images to train the model, which is hard to obtain and thus limits their generalization to in-the-wild images.

Besides, overfitting the implicit neural representation of a person's movement sequence to acquire actor-specific reconstructions becomes popular. NASA [13] coarsely models the naked body as the union of articulated parts, and each part is an implicit occupancy field. SCANimate [43] proposes an end-to-end trainable framework that turns raw 3D scans of a clothed human into an animatable avatar. SNARF [11] learns a forward deformation field to improve its generalization for unseen human poses. All these methods need 4D scan data to train their clothed body representation, and thus are difficult to be widely used for general image data.

Recently, some implicit representation methods, which can extract geometry and synthesize novel views based on multi-view images, attract researchers' attention. NeuralBody [39] reconstructs per frame's NeRF [34] field conditioned at body structured latent codes and utilizes the NeRF field to synthesize new images. However, the extracted geometry from NeRF suffers from noise. H-NeRF [51] utilizes an implicit parametric model [4] to reconstruct the temporal motion of humans. Neural Actor [30] integrates texture map features to refine volume rendering. IDR [53] combines implicit signed distance field and differential neural rendering to generate high-quality rigid reconstruction from multi-view images. Concurrent IMAvatar [54] expand IDR to learn implicit head avatars from monocular videos.

Explicit Human Reconstruction. With the help of human statistical model [5, 25, 31], some works utilize image cues to automatically obtain model parameters [9, 16, 28, 35]. To represent human clothing, some methods add displacements on SMPL [31] vertices to model tight clothing [2, 7, 33, 37]. However, this SMPL+D representation can only support tight clothing types and recover coarse level geometry shape. To improve the representation ability, some works adopt separate clothing representation and combine with SMPL body to do reconstruction [26, 32], but they need clothing type and high-quality 3D supervision.

Besides, to capture the performance of a specific person, many prior works use an actor-specific template to assist tracking. Monoperfcap [52] optimizes the deformation of template mesh to match 2D cues. LiveCap [17] refines the optimization pipeline and achieves real-time tracking for a specific person with monocular RGB input. DeepCap [18] adopts a network to predict per frame's template deformation for a specific person. However, the requirement for pre-defined templates limits their broader applications.

3. Method

SelfRecon aims to reconstruct a high-fidelity and space-time coherent clothed body shape from a monocular video depicting a self-rotating person, and the whole algorithm pipeline is given in Fig. 1. Both explicit and implicit geometric representations are utilized to achieve the above target. Specifically, we utilize the forward deformation field to generate space-time coherent explicit meshes. The deformation fields are decomposed into two parts, where the first one represents per frame's non-rigid deformation with a learnable MLP, and the second is the skinning deformation field. Differentiable masks, regular and smooth losses are adopted to control the shape of explicit meshes. To

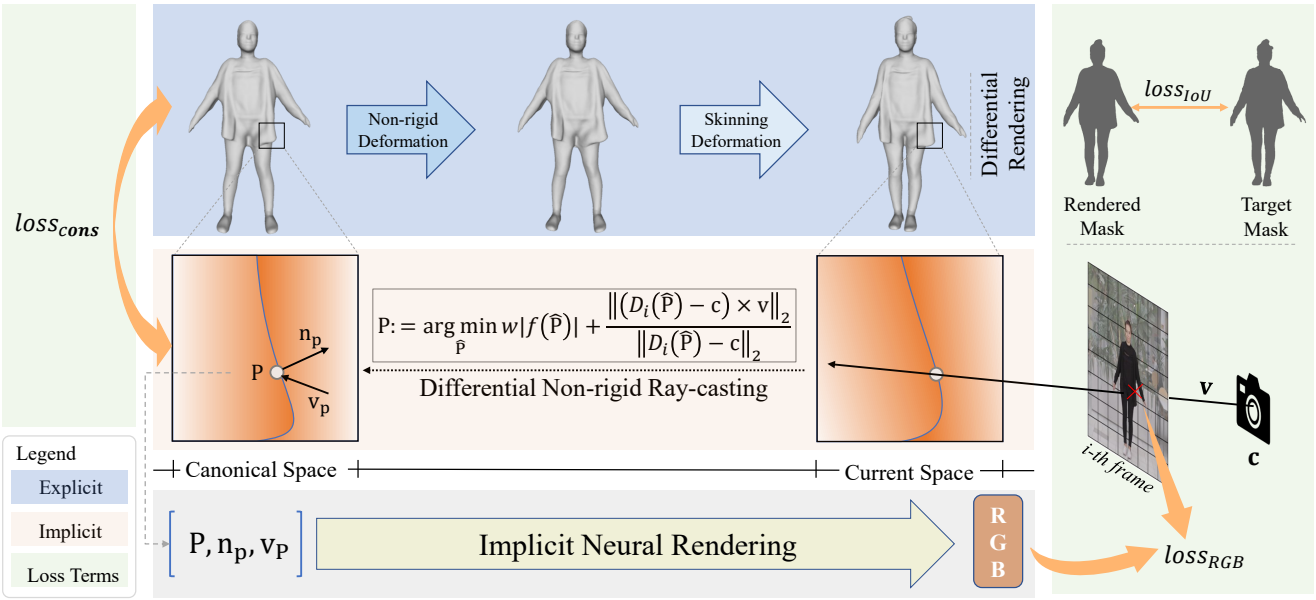

Figure 1. The pipeline of SelfRecon. We simultaneously maintain the explicit and implicit geometry representations and use forward deformation fields to transform canonical geometry to the current frame space. For explicit representation, we mainly use differentiable mask loss to recover the overall shape. As for implicit representation, sampled neural rendering loss and predicted normals are used to refine geometry details. Finally, a consistency loss is used to keep both geometric representations matched.

update the shape of implicit neural representation, we use non-rigid ray-casting (sec 3.3) to find the differentiable intersection points of rays and the deformed implicit surface. Then, the implicit rendering network (sec 3.4) will utilize the rays' color information to improve the geometry. Unless otherwise indicated, we also utilize the predicted normal map [42] to refine the details. Finally, a consistency loss is designed to match both representations.

For a self-rotating video with N frames, we adopt the method described in VideoAvatar [2] to generate the initial shape parameter $\boldsymbol{\beta}$ and per-frame's pose parameters $\{\boldsymbol{\theta}_i | i \in \{1, ..., N\}\}$ of SMPL model. We pre-defined a template pose and generate initial canonical SMPL body mesh $\mathcal{B}$ with $\boldsymbol{\beta}$ and this pose parameter. Our implicit and explicit representations are both initialized with $\mathcal{B}$. In the following, we present the algorithm details of each component.

3.1. Canonical Implicit SDF

In the similar work of VideoAvatar [2], they adopt the SMPL+D representation for clothed human body. However, SMPL+D has limited resolution and representation ability, and thus it can not represent high-fidelity geometry shape and various clothing types. In this work, we represent the canonical template shape $\mathcal{S}_\eta$ as the zero isosurface of a SDF, which is expressed by an MLP f with learnable weights η:

$$\mathcal{S}_\eta = \{\mathbf{p} \in \mathbb{R}^3 | f(\mathbf{p}; \eta) = 0\}. \quad (1)$$

To avoid unexpected solution, we use IGR [15] to initialize $\mathcal{S}_\eta$ as the initial canonical body $\mathcal{B}$.

3.2. Deformation Fields

Following prior works [21, 23], we utilize skeleton skinning transformation to control human body's large-scale movements due to the articulated structure. However, garments' non-rigid deformation cannot be fully represented by skinning transformation. Therefore, we extend to model non-rigid deformation with another MLP.

Non-rigid Deformation Field. We use an MLP d with learnable weights ϕ to represent the non-rigid deformation field. For i-th frame, d takes its optimizable conditional variable $\mathbf{h}_i$ as input and deform points in the canonical space with i-th frame's specific non-rigid deformation.

Skinning Transformation Field. Given i-th frame's pose parameter $\boldsymbol{\theta}_i$, we have to define a canonical-to-current space skinning transformation field $\mathcal{W}$. As the initial template body $\mathcal{B}$ has well-defined skinning weights relate to its SMPL skeleton, an intuitive idea is to expand the skinning weights of $\mathcal{B}$'s vertices to the whole canonical space to define the skinning transformation field. Specifically, we first pre-defined a sparse grid containing $\mathcal{B}$ in the canonical space. For each grid point, we find its nearest 30 vertices on $\mathcal{B}$ and average their skinning weights with IDW (inverse distance weight) as its initial weight. Then, we smooth all grid points' weights with Laplace smoothing. Finally, given

a point in the canonical space, we compute its skinning weights by trilinear interpolation in the grid. During our optimization, the grid is pre-computed and fixed. This forward deformation design avoids the trouble of inverse skinning transformation [11, 13, 24, 47] and provides a regular constrain for human articulated movement.

Finally, by compositing d and $\mathcal{W}$, we get the final deformation field $\mathcal{D} = \mathcal{W}(d(\cdot))$. It takes i-th frame's conditional variable $\mathbf{h}_i$ and SMPL pose parameter $\boldsymbol{\theta}_i$ as input, and transform canonical points to the i-th frame space. For brevity of description, we use $\mathcal{D}_i$ to denote i-th frame's deformation field, $\mathcal{S}_i$ for i-th frame's zero isosurface $\mathcal{D}_i(\mathcal{S}_\eta)$ and ψ_i for $\mathcal{D}_i$'s optimizable parameters $\{\phi, \mathbf{h}_i, \boldsymbol{\theta}_i\}$.

3.3. Differentiable Non-rigid Ray-casting

For rigid scenes, the sphere tracing algorithm [19, 27, 53] is widely used to find the intersection point of a ray and the SDF. However, it is not feasible here due to the deformation fields. Inspired by the method in [44], which proposes a strategy to render a deformed SDF, we utilize the explicit mesh to help find the intersection point of a ray and $\mathcal{S}_i$.

As shown in Fig 2, we extract an explicit template mesh $\mathbf{T}$ from the canonical surface $\mathcal{S}_\eta$. With deformation $\mathcal{D}_i$, we can get i-th frame's mesh $\mathbf{T}_i$. Theoretically, $\mathbf{T}_i$ is a piecewise linear approximation of $\mathcal{S}_i$. Therefore, consider a ray emitted from the camera position $\mathbf{c}$ along the direction $\mathbf{v}$, its first intersection $\hat{\mathbf{x}}$ with $\mathbf{T}_i$ is a good approximation of its intersection with $\mathcal{S}_i$. Moreover, with the intersected triangle on $\mathbf{T}_i$, we can find $\hat{\mathbf{x}}$'s corresponding point $\hat{\mathbf{p}}$ on the template $\mathbf{T}$ by consistent barycentric weights. Obviously, $\hat{\mathbf{p}}$ is close to $\mathcal{S}_\eta$ and is a good approximation of $\mathcal{D}_i^{-1}(\hat{\mathbf{x}})$. With $\hat{\mathbf{p}}$ as good initialization, we can find a point $\mathbf{p}$ on $\mathcal{S}_\eta$, whose deformed point $\mathbf{x} = \mathcal{D}_i(\mathbf{p})$ is exactly the intersection point of the ray $\mathbf{r}$ and $\mathcal{S}_i$. Specifically, we solve $\mathbf{p}$ by:

$$\mathbf{p} = \arg\min_{\hat{\mathbf{p}}} \ \omega |f(\hat{\mathbf{p}})| + \frac{\|(\mathcal{D}_i(\hat{\mathbf{p}}) - \mathbf{c}) \times \mathbf{v}\|_2}{\|\mathcal{D}_i(\hat{\mathbf{p}}) - \mathbf{c}\|_2}, \quad (2)$$

where the first item constrains $\hat{\mathbf{p}}$ to be close with $\mathcal{S}_\eta$ and the second item restricts $\mathcal{D}_i(\hat{\mathbf{p}})$ on the ray. In our implementation, we set $\omega = 3.05$ and execute 10 gradient descent iterations to solve $\mathbf{p}$. To guarantee accuracy, we reject those samples with large losses.

Differentiable Formula. The above-mentioned solving process of $\mathbf{p}$ is an iterative optimization process, which is not differentiable. For the ray in i-th frame, camera position $\mathbf{c}$, view direction $\mathbf{v}$, $\mathcal{D}_i$'s parameters ψ_i and f's parameter η uniquely determine the $\mathbf{p}$. Therefore, $\mathbf{p}$ can be seen as a function of these parameters, and we need to compute partial derivatives of $\mathbf{p}$ to all these parameters. For brevity, we only clarify our calculation for η here, and other partial derivatives are computed similarly.

Through the above analysis, $\mathbf{p}$ satisfies the surface and ray constraints: $f(\mathbf{p}) \equiv 0$ and $(\mathcal{D}_i(\mathbf{p}) - \mathbf{c}) \times \mathbf{v} \equiv 0$. We

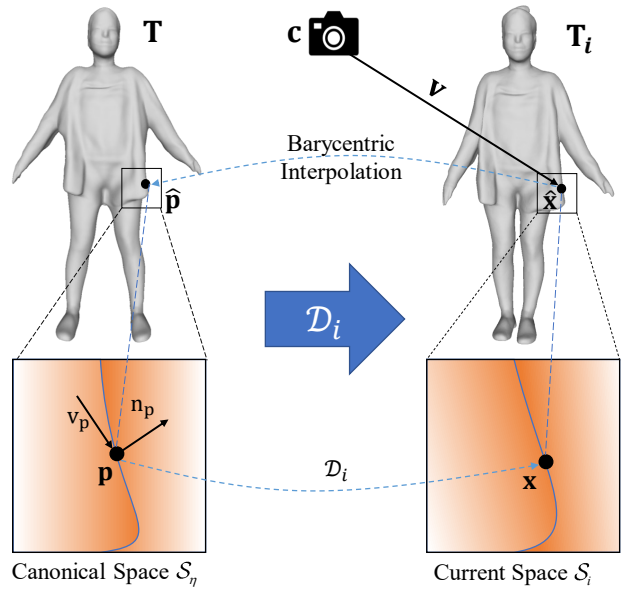

Figure 2. Some related symbols and illustration about differentiable non-rigid ray-casting and implicit neural rendering.

differentiate these two equations w.r.t η to get:

$$\frac{\partial f}{\partial \mathbf{p}}^T \frac{\partial \mathbf{p}}{\partial \eta} = -\frac{\partial f}{\partial \eta} \quad [\mathbf{v}]_\times \frac{\partial \mathbf{x}}{\partial \mathbf{p}} \frac{\partial \mathbf{p}}{\partial \eta} = 0, \quad (3)$$

where $\mathbf{x} = \mathcal{D}_i(\mathbf{p})$ and $[\mathbf{v}]_\times$ is $\mathbf{v}$'s cross product matrix. We concatenate these two equations to get a 4×3 linear system, then $\frac{\partial \mathbf{p}}{\partial \eta}$ is computed by solving its normal equation.

3.4. Implicit Rendering Network

IDR [53] proposes an MLP M to approximate the rendering equation and demonstrates certain disentangle ability of lighting and material. In their rigid configuration, M takes the zero isosurface's point, its normal, the view direction and its global geometry feature vector as input to estimate the point's color along the view direction. We subtly transfer their design to non-rigid scenarios by converting related current frame's attributes to canonical space. As shown in Fig. 2, considering a ray emitted from camera center $\mathbf{c}$ along a sampled pixel, whose direction $\mathbf{v}$ is determined by camera's intrinsic parameters τ, we compute its intersection point $\mathbf{p}$ on $\mathcal{S}_\eta$ with the algorithm described in sec 3.3. In the meantime, we compute its normal $\mathbf{n}_\mathbf{p} = \nabla f(\mathbf{p}; \eta)$ by gradient calculation. Then, the view direction $\mathbf{v}_\mathbf{p}$ in canonical space can be computed by transferring $\mathbf{v}$ with the Jacobian matrix $J_\mathbf{x}(\mathbf{p})$ of the deformed point $\mathbf{x} = \mathcal{D}_i(\mathbf{p})$ w.r.t $\mathbf{p}$. As for the global geometry feature, we similarly use a larger MLP $F(\mathbf{p}; \eta) = (f(\mathbf{p}; \eta), \mathbf{z}(\mathbf{p}; \eta))$ to additionally compute it, which implies the geometry information around $\mathbf{p}$ and can be used to help the prediction of global shadow [53].

Finally, we use an MLP M with learnable weights γ to compute $\mathbf{p}$'s color $L_{\mathbf{p}}(\eta, \psi_i, \gamma, \tau)$, formulated as:

$$\begin{aligned} L_{\mathbf{p}}(\eta, \psi_i, \gamma, \tau) &= M(\mathbf{p}, \mathbf{n}_{\mathbf{p}}, \mathbf{v}_{\mathbf{p}}, \mathbf{z}(\mathbf{p}; \eta); \gamma) \\ \mathbf{n}_{\mathbf{p}} &= \nabla f(\mathbf{p}; \eta) \\ \mathbf{x} &= \mathcal{D}_i(\mathbf{p}) \\ \mathbf{v}_{\mathbf{p}} &= J_{\mathbf{x}}(\mathbf{p})^{-1}\mathbf{v}, \end{aligned} \tag{4}$$

where related symbols have been described above. It can be seen that the color along direction $\mathbf{v}$ of the deformed point $\mathbf{x}$ in i-th frame is determined by the MLP weights η and γ, camera parameters τ and deformation field parameters ψ_i.

3.5. Loss Function

According to the description above, for a N-frame self-rotating video, the set $\mathcal{X}$ of all optimizable parameters is:

$$\mathcal{X} = \{\eta, \gamma, \phi, \tau\} \cup \{\mathbf{h}_i, \boldsymbol{\theta}_i | i \in 1, ..., N\},$$

which includes camera parameters, the learnable weights of MLPs shared by the whole sequence and per frame's specific pose parameters and non-rigid deformation field's conditional variable. Our target is to design a loss function and optimize $\mathcal{X}$ to match the mask and RGB images $\{O_i, I_i | i \in 1, ..., N\}$ of the input video. Besides, a predicted normal map $\{N_i | i \in 1, ..., N\}$ is added to the optimization. As SelfRecon maintains both explicit and implicit geometry, the loss terms can be divided into two parts.

3.5.1 Explicit Loss

During the computation of explicit losses, we temporarily regard the canonical mesh vertices $\mathbf{T}$ as an optimizable variable and compute its gradient together with $\mathcal{X}$. Then in the consistency loss, we associate its variations with our implicit representation. At present, explicit losses mainly include mask loss, deformation regularization loss, and the smoothness loss of the skeleton.

Mask Loss. We utilize a differentiable renderer [48] based on point cloud to render the mask $O(\mathbf{T}_i)$ of i-th frame's mesh $\mathbf{T}_i = \mathcal{D}_i(\mathbf{T})$ with camera parameters, and compute the IoU loss with target mask O_i:

$$loss_{IoU} = 1 - \frac{\|O(\mathbf{T}_i) \otimes O_i\|_1}{\|O(\mathbf{T}_i) \oplus O_i - O(\mathbf{T}_i) \otimes O_i\|_1}, \tag{5}$$

where $\otimes$ and $\oplus$ are the operators that perform element-wise product and sum respectively.

Deformation Regularization Loss. As stated in Sec. 3.2, the i-th frame's deformation field $\mathcal{D}_i$ contains variable d and fixed $\mathcal{W}$. d represents the deformation that can not be represented by skinning transformation $\mathcal{W}$, and this deformation should be relatively small. To associate the skeleton pose, we design the following regularization loss:

$$loss_{regu} = \frac{1}{|\mathbf{T}|} \sum_{\mathbf{t} \in \mathbf{T}} \rho(\|\mathcal{W}(\mathbf{t}; \boldsymbol{\theta}_i) - \mathcal{D}_i(\mathbf{t})\|_2), \tag{6}$$

where $\mathbf{t}$ is a vertex coordinate of $\mathbf{T}$, $|\mathbf{T}|$ is the vertices number of $\mathbf{T}$, and ρ is the Geman-McClure robust loss [14].

Skeleton Smoothness Loss. The motion trajectory of the joints should be low frequency. Similar with MonoPerf [52], we smooth the skeleton coordinates of 30 consecutive frames by minimizing the distance to a 10 dimensional linear subspace $\mathbf{B} \in \mathbb{R}^{30\times10}$ spanned by the 10 lowest frequency basis vectors of the discrete cosine transform:

$$loss_{ske} = \frac{1}{30}\|\mathbf{J}\text{Null}(\mathbf{B})\|_F^2. \tag{7}$$

Here, $\text{Null}(\mathbf{B})$ denotes the nullspace of the $\mathbf{B}$ matrix, the matrix $\mathbf{J} \in \mathbb{R}^{72\times30}$ stacks all skeleton coordinates of consecutive 30 frames, and $\|\cdot\|_F$ denotes the Frobenius norm.

Finally, the loss for the explicit representation is:

$$Loss_{exp} = loss_{IoU} + \lambda_{e1} loss_{regu} + \lambda_{e2} loss_{ske}. \tag{8}$$

λ_{e1} and λ_{e2} adjust the weights of related losses. After each iteration, we reserve $\mathcal{X}$'s gradients and wait for the implicit loss iteration to update together. For canonical mesh vertices, we use SGD to update $\mathbf{T}$ to $\hat{\mathbf{T}}$, which will be used in consistency loss to match two representations.

3.5.2 Implicit Loss

We sample pixels within the ground truth mask and utilize Sec. 3.3 to get the ray's intersection $\mathbf{p}$ on $\mathcal{S}_\eta$ and its corresponding ground truth color $I_{\mathbf{p}}$ and predicted normal $N_{\mathbf{p}}$ if available. Then, based on this sampled points set P, we construct two losses.

Color Loss. By referring to Eq. (4), we formulate the color loss as:

$$loss_{RGB} = \frac{1}{|\mathrm{P}|} \sum_{\mathbf{p} \in \mathrm{P}} |L_{\mathbf{p}}(\mathcal{X}) - I_{\mathbf{p}}|. \tag{9}$$

Here, we use $\mathcal{X}$ to substitute related parameters in Eq. (4). Intuitively, this loss requires that the rendered images should match the input images.

Normal Loss. We utilized the predicted normal map by PIFuHD [42] to further refine the geometry shape. Referring to Eq. (4), we can easily compute $\mathbf{p}$'s normal $\mathbf{n}_{\mathbf{p}}$. Besides, we need to transform the corresponding predicted normal $N_{\mathbf{p}}$ from the space of current frame to canonical space, which can be computed with $J_{\mathbf{x}}(\mathbf{p})^T N_{\mathbf{p}}$, where $J_{\mathbf{x}}(\mathbf{p})$ is the Jacobian matrix of the forward deformation field at $\mathbf{p}$ [44]. Thus, the normal loss is:

$$loss_{norm} = \frac{1}{|\mathrm{P}|} \sum_{\mathbf{p} \in \mathrm{P}} \omega_{\mathbf{p}} \|\mathbf{n}_{\mathbf{p}} - unit(J_{\mathbf{x}}(\mathbf{p})^T N_{\mathbf{p}})\|_2. \tag{10}$$

Here, $unit(\cdot)$ means to normalize the vector. $\omega_{\mathbf{p}}$ is the weight defined by the cosine of angle between $\mathbf{n_p}$ and corresponding view direction. Since the predicted normals are noisy and inconsistent between frames, we use this weights to alleviate the impact of normal that deviates from its view direction and avoid geometry artifacts.

We also design regular losses for the implicit representation, and these losses are defined on the set of sampled points S near the implicit surface [15].

Rigidity Loss. We require the first deformation field d to be as rigid as possible to avoid distortion. Following Park *et al.* [36], we design our loss as:

$$loss_{rigid} = \frac{1}{|\mathrm{S}|}\sum_{\mathbf{p}\in\mathrm{S}}\rho(\|\log\Sigma_{\mathbf{p}}\|_F), \tag{11}$$

where $\Sigma_{\mathbf{p}}$ is the singular value diagonal matrix of the Jacobian of d on $\mathbf{p}$ and ρ is the robust function [14].

Eikonal Loss. We adopt the regular loss of IGR [15] to make f to be sign distance function:

$$loss_{sdf} = \frac{1}{|\mathrm{S}|}\sum_{\mathbf{p}\in\mathrm{S}}(\|\mathbf{n_p}\|_2 - 1)^2, \tag{12}$$

where $\mathbf{n_p}$ is obtained by differentiating f at $\mathbf{p}$.

Finally, the implicit loss can be represented as:

$$Loss_{imp} = loss_{RGB} + \lambda_{i1} loss_{norm} + \lambda_{i2} loss_{rigid} + \lambda_{i3} loss_{sdf}, \tag{13}$$

where λ_{i1}, λ_{i2} and λ_{i3} are balancing weights.

3.5.3 Explicit/Implicit Consistency

After explicit iteration, the canonical mesh has been updated to $\hat{\mathbf{T}}$, to make the implicit SDF consistent with the updated explicit mesh during implicit iteration, we design a consistency loss:

$$Loss_{cons} = \frac{1}{|\hat{\mathbf{T}}|}\sum_{\hat{\mathbf{t}}\in\hat{\mathbf{T}}}|f(\hat{\mathbf{t}};\eta)|, \tag{14}$$

where $\hat{\mathbf{t}}$ is a vertex coordinate of $\hat{\mathbf{T}}$. Intuitively, the loss requires $\hat{\mathbf{T}}$ to match the implicit surface $\mathcal{S}_\eta$.

In each optimization step, we first perform the explicit iteration to obtain $\hat{\mathbf{T}}$ and reserve $\mathcal{X}$'s gradients. Then, we compute implicit and consistency losses to accumulate new $\mathcal{X}$'s gradients. Finally, Adam is utilized to update $\mathcal{X}$ with computed gradients.

4. Experiments

We conduct quantitative and qualitative experiments to demonstrate the effectiveness of SelfRecon. For quantitative evaluation, we synthesize several sequences with commercial software due to the lack of high-quality geometry data of human body with general clothing. For qualitative evaluation, we mainly utilize the PeopleSnapshot [2] dataset and several real sequences collected by ourselves. We also present ablation study for the loss term design and present an avatar generation application.

Table 1. The errors (cm) on the synthetic five sequences. We report three error metrics: the average distance from reconstructed to GT meshes (Recon), the average distance from GT to reconstructed meshes (GT), and the Chamfer distance. For each error metric, we report the values of VideoAvatar and ours in two consecutive rows.

Subject	f1	f2	f3	m1	m2	mean
Recon	**1.59**	1.71	1.93	1.81	1.27	1.66
	1.67	**1.32**	**1.63**	**1.53**	**1.17**	**1.46**
GT	2.08	1.50	2.40	1.92	1.42	1.86
	1.62	**1.16**	**1.92**	**1.53**	**1.17**	**1.48**
Chamfer	1.84	1.60	2.17	1.86	1.34	1.76
	1.64	**1.24**	**1.77**	**1.53**	**1.17**	**1.47**

4.1. Quantitative Evaluation

We synthesize data to quantitatively evaluate our reconstruction algorithm. Specifically, we use Blender [8] to design the self-rotating motions for male and female avatars. Then, we utilize CLO3D [12] to design several clothes and animate the clothed body with motions. Finally, we synthesized two sets of male and three sets of female dressing sequences. We reconstruct these sequences with VideoAvatar [2] and our method, and report the registration error for the canonical posture results in Tab. 1. Compared with VideoAvatar, our method significantly reduces the values of various error metrics. In Fig. 3, we also present four group results and their error maps. Intuitively, our results capture the overall shape and have some reasonable details. As VideoAvatar is based on the SMPL+D representation, it has plausible results for tighter clothing, like the male examples, but lacks detailed reconstruction ability. Moreover, it can not correctly reconstruct loose clothing, especially for the females dressed in skirts.

4.2. Qualitative Evaluation

We also qualitatively compare SelfRecon with multi-frame prediction algorithm PaMIR [55], optimization method VideoAvatar [2] and NeRF [34] based neural rendering method NeuralBody [39] on several sequences of PeopleSnapshot dataset. In Fig. 4, we present the first frame of input video, our rendered image, and reconstruction results of all methods from two perspectives. And we compare with PaMIR in the first two rows and the others with NeuralBody. As we can see, VideoAvatar based on SMPL+D can only approximately capture the overall shape, but details such as hairstyle and clothing folds are lost. PaMIR uses multi-frame input to improve its results,

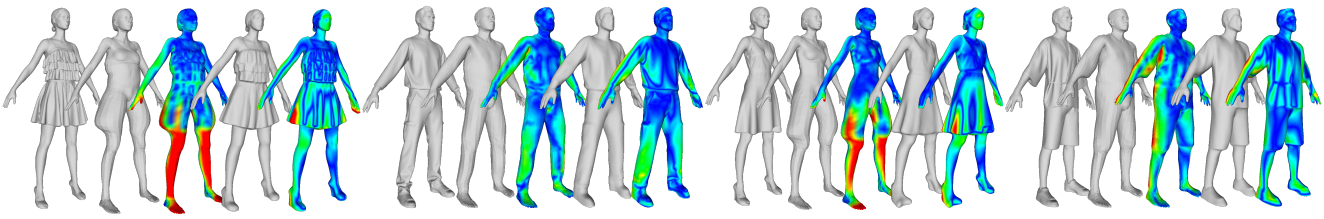

Figure 3. Reconstructions in canonical pose and their error maps for four synthetic self-rotating sequences. In each group, we show the GT mesh, results of VideoAvatar and SelfRecon in turn (red means $\geq$ 6cm).

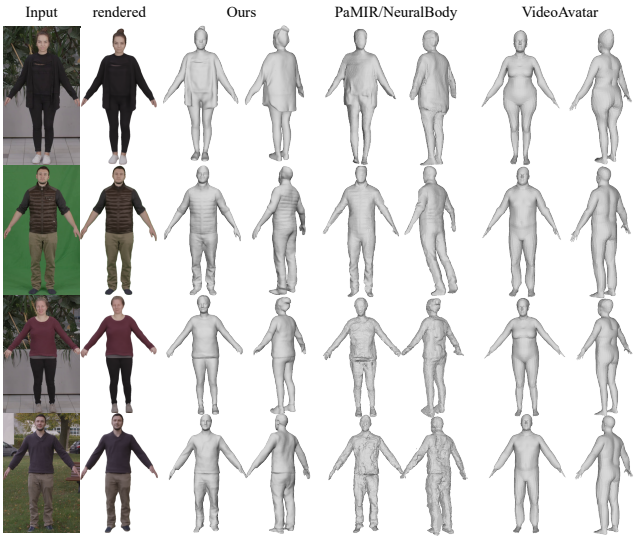

Figure 4. Comparison results with methods that use video or multi-frame images, including PaMIR [55], NeuralBody [39] and VideoAvatar [2]. For the second comparison group, the method in the first two rows is PaMIR, and the rest is NeuralBody. We also present our rendered image as reference. SelfRecon can reconstruct high-fidelity geometry shape of standing posture, including facial features and clothing folds.

but still suffers from the deep ambiguity. As in the second example, its reconstructed human is not upright from the side view. Besides, its results have some details but miss facial features, while our method has better details and can recover certain facial features. Similar with SelfRecon, NeuralBody also inputs a video to do self-supervised optimization. It mainly focuses on novel view synthesis, but still can extract geometry from underlying NeRF representation. We can see that its reconstructions alleviate deep ambiguity and conform to the human's overall structure while suffering from large noise on the surface, which may be caused by excessive freedom of volume rendering. Different from their method, implicit surface representation based SelfRecon can recover high-fidelity geometry shape without noises.

Figure 5. Reconstruction results from videos taken by smartphones. Each group shows the first image of the video, corresponding neural rendered image and reconstructed shape.

Fig. 5 shows reconstruction results on our collected videos with smartphones. For each group, we present the first frame of the video, our rendered image and reconstruction. Our results have high-fidelity geometry shapes for kinds of clothing and body, and our neural rendered images are also quite close to the input images.

4.3. Ablation Study

Our complete algorithm requires color images, masks and normal maps as inputs. Fig. 6 shows ablation experiments of two examples on three inputs. As the results show, if only using the mask loss, the recovered geometry shape is inside the convex hull formed by the silhouettes but lacks details and has noticeable concavities. After adding the color loss, it significantly improves the details and reduces the unnatural concavities. For the second example, the result has been very close to the result obtained by adding the normal loss. However, for the first example, it can not completely eliminate the depressed geometry with-

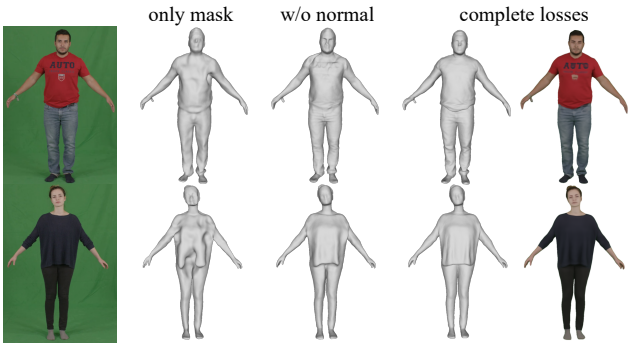

Figure 6. Ablation study for color, mask and normal losses. With mask loss only, the results lack details and have much concave geometry. After adding color loss, geometry is significantly improved but can not completely eliminate the concavities. With the predicted normals, the results are further improved. In the last column, we also show the neural rendering images as reference.

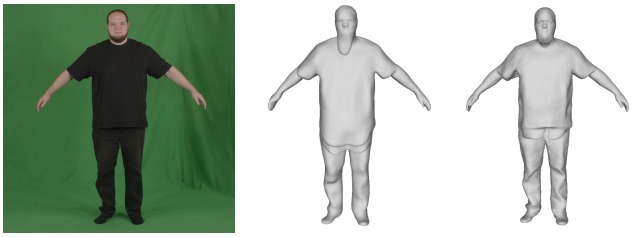

Figure 7. Without the adaptive weights in Eq. (10) and small λ_{i1} in Eq. (13), normal loss will result artifacts (middle). After adjusting the weights, the corresponding result (right) is more plausible.

out normal loss. This may be caused by lacking rich texture and multi-view observations in these areas. With the normal loss, our results are further improved, and the unnatural pits are eliminated while the details are preserved.

Since the normal prediction network [42] is trained with synthetic images, its prediction may be not accurate for real tests and may be not consistent for different frames. As shown in Fig. 7, without the adaptive weights in Eq. (10), the normal loss might result unexpted results.

4.4. Avatar Generation

Thanks to our forward deformation field design, we can extract a mesh sequence with consistent topology. Based on the tracking results, we can extract a texture template mesh from the images with bound skinning weights from the skinning transformation field. Then, an animatable avatar is generated and can be driven with SMPL pose parameters. For texture extraction, we follow the method of VideoAvatar [2]. Fig. 8 shows two examples of texture generation and driving from the PeopleSnapshot dataset. Our method recovers better geometric details like facial, shoes

Figure 8. The reconstructed texture mesh and driving results. The left shows the reference image, texture meshes generated by VideoAvatar and SelfRecon. On the right, we use three pose parameters to drive our texture mesh and generate plausible results.

and clothing folds thanks to more accurate tracking results. Besides, our driving results look plausible and may be of sufficient quality for some applications.

5. Conclusion and Discussion

We proposed SelfRecon, a self-supervised reconstruction method based on neural implicit representation and neural rendering. With forward deformation, our method can be easily applied to body movement and recover space-time coherent surfaces, which is convenient for downstream applications. Moreover, combining the explicit representation, we proposed a non-rigid ray casting algorithm, which makes it possible for differentiable intersecting with the deformed implicit surface. SelfRecon can reconstruct high-fidelity clothed body shape from a self-rotating video without pre-computed templates. We also show high-fidelity avatar generation with our tracking results, demonstrating potential applications of SelfRecon.

SelfRecon still has several limitations. First, it requires relatively long time to optimize, which limits its convenient applications. However, this problem can be alleviated with the help of body priors and the fast growing field of neural rendering. Second, current method relies on the predicted normal maps to improve the geometric details. How to recover the geometric details directly from the self-supervised rendering loss is worthy of future study. Third, the proposed method mainly works well for self-rotating motions, and it is worthy of study for more general motion sequences.

Acknowledgement This research was supported by National Natural Science Foundation of China (No. 62122071), the Youth Innovation Promotion Association CAS (No. 2018495), “the Fundamental Research Funds for the Central Universities”(No. WK3470000021).

References

[1] Thiemo Alldieck, Marcus Magnor, Bharat Lal Bhatnagar, Christian Theobalt, and Gerard Pons-Moll. Learning to reconstruct people in clothing from a single RGB camera. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, jun 2019. 1

[2] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Video based reconstruction of 3d people models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 1, 2, 3, 6, 7, 8

[3] Thiemo Alldieck, Gerard Pons-Moll, Christian Theobalt, and Marcus Magnor. Tex2shape: Detailed full human body geometry from a single image. In *IEEE International Conference on Computer Vision (ICCV)*. IEEE, oct 2019. 1

[4] Thiemo Alldieck, Hongyi Xu, and Cristian Sminchisescu. imghum: Implicit generative models of 3d human shape and articulated pose. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5461–5470, 2021. 2

[5] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. Scape: shape completion and animation of people. In *ACM SIGGRAPH 2005 Papers*, pages 408–416. 2005. 2

[6] Bharat Lal Bhatnagar, Garvita Tiwari, Christian Theobalt, and Gerard Pons-Moll. Multi-garment net: Learning to dress 3d people from images. In *IEEE International Conference on Computer Vision (ICCV)*. IEEE, oct 2019. 1

[7] Bharat Lal Bhatnagar, Garvita Tiwari, Christian Theobalt, and Gerard Pons-Moll. Multi-garment net: Learning to dress 3d people from images. In *IEEE International Conference on Computer Vision (ICCV)*, 2019. 2

[8] Blender. 2021. https://www.blender.org/. 6

[9] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *European conference on computer vision*, pages 561–578. Springer, 2016. 2

[10] Andrei Burov, Matthias Nießner, and Justus Thies. Dynamic neural radiance fields for monocular 4d facial avatar reconstruction, 2021. 1

[11] Xu Chen, Yufeng Zheng, Michael J Black, Otmar Hilliges, and Andreas Geiger. Snarf: Differentiable forward skinning for animating non-rigid neural implicit shapes. In *International Conference on Computer Vision (ICCV)*, 2021. 1, 2, 4

[12] CLO3D. 2021. https://www.clo3d.com/. 6

[13] Boyang Deng, John P Lewis, Timothy Jeruzalski, Gerard Pons-Moll, Geoffrey Hinton, Mohammad Norouzi, and Andrea Tagliasacchi. Nasa neural articulated shape approximation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16*, pages 612–628. Springer, 2020. 1, 2, 4

[14] Stuart Ganan and D McClure. Bayesian image analysis: An application to single photon emission tomography. *Amer. Statist. Assoc*, pages 12–18, 1985. 5, 6

[15] Amos Gropp, Lior Yariv, Niv Haim, Matan Atzmon, and Yaron Lipman. Implicit geometric regularization for learning shapes. In *Proceedings of Machine Learning and Systems 2020*, pages 3569–3579. 2020. 3, 6

[16] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7297–7306, 2018. 2

[17] Marc Habermann, Weipeng Xu, Michael Zollhoefer, Gerard Pons-Moll, and Christian Theobalt. Livecap: Real-time human performance capture from monocular video. *ACM Transactions On Graphics (TOG)*, 38(2):1–17, 2019. 1, 2

[18] 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. 1, 2

[19] John C Hart. Sphere tracing: A geometric method for the antialiased ray tracing of implicit surfaces. *The Visual Computer*, 12(10):527–545, 1996. 4

[20] Tong He, John Collomosse, Hailin Jin, and Stefano Soatto. Geo-pifu: Geometry and pixel aligned implicit functions for single-view human reconstruction. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2020. 1

[21] Tong He, Yuanlu Xu, Shunsuke Saito, Stefano Soatto, and Tony Tung. Arch++: Animation-ready clothed human reconstruction revisited. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11046–11056, 2021. 1, 3

[22] Yang Hong, Juyong Zhang, Boyi Jiang, Yudong Guo, Ligang Liu, and Hujun Bao. Stereopifu: Depth aware clothed human digitization via stereo vision. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[23] Zeng Huang, Yuanlu Xu, Christoph Lassner, Hao Li, and Tony Tung. Arch: Animatable reconstruction of clothed humans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3093–3102, 2020. 1, 3

[24] Timothy Jeruzalski, David I. W. Levin, Alec Jacobson, Paul Lalonde, Mohammad Norouzi, and Andrea Tagliasacchi. Nilbs: Neural inverse linear blend skinning. *CoRR*, abs/2004.05980, 2020. 1, 4

[25] Boyi Jiang, Juyong Zhang, Jianfei Cai, and Jianmin Zheng. Disentangled human body embedding based on deep hierarchical neural network. *IEEE Transactions on Visualization and Computer Graphics*, 2020. 2

[26] Boyi Jiang, Juyong Zhang, Yang Hong, Jinhao Luo, Ligang Liu, and Hujun Bao. Bcnet: Learning body and cloth shape from a single image. In *European Conference on Computer Vision*. Springer, 2020. 2

[27] Yue Jiang, Dantong Ji, Zhizhong Han, and Matthias Zwicker. Sdfdiff: Differentiable rendering of signed distance fields for 3d shape optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1251–1261, 2020. 4

[28] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and

pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7122–7131, 2018. 2

[29] Zhengqi Li, Simon Niklaus, Noah Snavely, and Oliver Wang. Neural scene flow fields for space-time view synthesis of dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6498–6508, 2021. 1

[30] 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 SIGGRAPH Asia*, 2021. 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)*, 34(6):1–16, 2015. 1, 2

[32] Qianli Ma, Shunsuke Saito, Jinlong Yang, Siyu Tang, and Michael J Black. Scale: Modeling clothed humans with a surface codec of articulated local elements. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16082–16093, 2021. 2

[33] Qianli Ma, Jinlong Yang, Anurag Ranjan, Sergi Pujades, Gerard Pons-Moll, Siyu Tang, and Michael J Black. Learning to dress 3d people in generative clothing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6469–6478, 2020. 2

[34] 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. In *European Conference on Computer Vision*, pages 405–421. Springer, 2020. 2, 6

[35] Mohamed Omran, Christoph Lassner, Gerard Pons-Moll, Peter Gehler, and Bernt Schiele. Neural body fitting: Unifying deep learning and model based human pose and shape estimation. In *2018 international conference on 3D vision (3DV)*, pages 484–494. IEEE, 2018. 2

[36] 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. 1, 6

[37] Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. Learning to estimate 3d human pose and shape from a single color image. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 459–468, 2018. 2

[38] Sida Peng, Junting Dong, Qianqian Wang, Shangzhan Zhang, Qing Shuai, Xiaowei Zhou, and Hujun Bao. Animatable neural radiance fields for modeling dynamic human bodies. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14314–14323, 2021. 1

[39] 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. In *CVPR*, 2021. 2, 6, 7

[40] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10318–10327, 2021. 1

[41] 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. 1, 2

[42] 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. 1, 2, 3, 5, 8

[43] 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

[44] Dario Seyb, Alec Jacobson, Derek Nowrouzezahrai, and Wojciech Jarosz. Non-linear sphere tracing for rendering deformed signed distance fields. *ACM Transactions on Graphics*, 38(6), 2019. 4, 5

[45] 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. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12959–12970, 2021. 1

[46] Shaofei Wang, Marko Mihajlovic, Qianli Ma, Andreas Geiger, and Siyu Tang. Metaavatar: Learning animatable clothed human models from few depth images. In *Advances in Neural Information Processing Systems*, 2021. 1

[47] Chung-Yi Weng, Brian Curless, and Ira Kemelmacher-Shlizerman. Vid2actor: Free-viewpoint animatable person synthesis from video in the wild. *arXiv preprint arXiv:2012.12884*, 2020. 1, 4

[48] Olivia Wiles, Georgia Gkioxari, Richard Szeliski, and Justin Johnson. Synsin: End-to-end view synthesis from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7467–7477, 2020. 5

[49] Donglai Xiang, Fabian Prada, Chenglei Wu, and Jessica Hodgins. Monoclothcap: Towards temporally coherent clothing capture from monocular rgb video. In *2020 International Conference on 3D Vision (3DV)*, pages 322–332. IEEE, 2020. 1

[50] Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J Black. Icon: Implicit clothed humans obtained from normals. *arXiv preprint arXiv:2112.09127*, 2021. 1

[51] 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

[52] 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. 1, 2, 5

[53] Lior Yariv, Yoni Kasten, Dror Moran, Meirav Galun, Matan Atzmon, Basri Ronen, and Yaron Lipman. Multiview neural surface reconstruction by disentangling geometry and appearance. *Advances in Neural Information Processing Systems*, 33, 2020. 2, 4

[54] Yufeng Zheng, Victoria Fernández Abrevaya, Xu Chen, Marcel C Bühler, Michael J Black, and Otmar Hilliges. Im avatar: Implicit morphable head avatars from videos. *arXiv preprint arXiv:2112.07471*, 2021. 2

[55] Zerong Zheng, Tao Yu, Yebin Liu, and Qionghai Dai. Pamir: Parametric model-conditioned implicit representation for image-based human reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 1, 2, 6, 7

[56] Zerong Zheng, Tao Yu, Yixuan Wei, Qionghai Dai, and Yebin Liu. Deephuman: 3d human reconstruction from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7739–7749, 2019. 1