Capturing and Animation of Body and Clothing from Monocular Video

Yao Feng  
MPI for Intelligent Systems  
Tübingen, Germany  
E-mail: yfeng@tuebingen.mpg.de

Jinlong Yang  
MPI for Intelligent Systems  
Tübingen, Germany  
E-mail: jyang@tuebingen.mpg.de

Marc Pollefeys  
ETH Zürich  
Zürich, Switzerland  
E-mail: marc.pollefeys@inf.ethz.ch

Michael J. Black  
MPI for Intelligent Systems  
Tübingen, Germany  
E-mail: black@tuebingen.mpg.de

Timo Bolkart  
MPI for Intelligent Systems  
Tübingen, Germany  
E-mail: tbolkart@tuebingen.mpg.de

ABSTRACT

While recent work has shown progress on extracting clothed 3D human avatars from a single image, video, or a set of 3D scans, several limitations remain. Most methods use a holistic representation to jointly model the body and clothing, which means that the clothing and body cannot be separated for applications like virtual try-on. Other methods separately model the body and clothing, but they require training from a large set of 3D clothed human meshes obtained from 3D/4D scanners or physics simulations. Our insight is that the body and clothing have different modeling requirements. While the body is well represented by a mesh-based parametric 3D model, implicit representations and neural radiance fields are better suited to capturing the large variety in shape and appearance present in clothing. Building on this insight, we propose SCARF (Segmented Clothed Avatar Radiance Field), a hybrid model combining a mesh-based body with a neural radiance field.

Integrating the mesh into the volumetric rendering in combination with a differentiable rasterizer enables us to optimize SCARF directly from monocular videos, without any 3D supervision. The hybrid modeling enables SCARF to (i) animate the clothed body avatar by changing body poses (including hand articulation and facial expressions), (ii) synthesize novel views of the avatar, and (iii) transfer clothing between avatars in virtual try-on applications. We demonstrate that SCARF reconstructs clothing with higher visual quality than existing methods, that the clothing deforms with changing body pose and body shape, and that clothing can be successfully transferred between avatars of different subjects. The code and models are available at https://github.com/YadiraF/SCARF.

CCS CONCEPTS

• Computing methodologies → Shape modeling.

KEYWORDS

Avatar creation, clothing capture

ACM Reference Format:

Yao Feng, Jinlong Yang, Marc Pollefeys, Michael J. Black, and Timo Bolkart. 2022. Capturing and Animation of Body and Clothing from Monocular Video. In SIGGRAPH Asia 2022 Conference Papers (SA ’22 Conference Papers), December 6–9, 2022, Daegu, Republic of Korea. ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3550469.3555423

Figure 1: Given a monocular video (a), our method (SCARF) builds an avatar where the body and clothing are disentangled (b). The body is represented by a traditional mesh, while the clothing is captured by an implicit neural representation. SCARF enables animation with detailed control over the face and hands (c) as well as clothing transfer between subjects (d).
1 INTRODUCTION

Realistic avatar creation is a key enabler of the metaverse, and it supports many applications in virtual presence, fitness, digital fashion, and entertainment. Traditional ways to build avatars require either complex capture systems or manual design by artists, both of which are time-consuming and inefficient for large-scale avatar creation. To address this, previous work explores more practical ways to create avatars directly from single RGB images or monocular videos, which are more accessible to consumers.

The majority of work (e.g., [Choutas et al. 2020; Feng et al. 2021a; Kanazawa et al. 2018; Kolotouros et al. 2019; Pavlakos et al. 2019; Rong et al. 2021; Zanfir et al. 2021]) creates 3D human body avatars from images by estimating parameters of statistical 3D mesh models such as SCAPE [Anguelov et al. 2005], Adam [Joo et al. 2018], SMPL/SMPL-X [Loper et al. 2015; Pavlakos et al. 2019], GHUM [Xu et al. 2020], or STAR [Osman et al. 2020], or implicit surface models like imGHUM [Allieck et al. 2021] and LEAP [Mihajlovic et al. 2021]. As these models are trained from minimally clothed body scans, they are unable to capture clothing shape and appearance variations. Methods that recover clothed bodies from images are instead trained with a large set of 3D clothed human scans [Saito et al. 2019, 2020; Xiu et al. 2022], or optimize the clothed avatar directly from multi-view images or videos [Chen et al. 2021b; Jiang et al. 2022; Liu et al. 2021b; Feng et al. 2021a, 2022, 2021b; Xu et al. 2021]. To handle the complex topology of different clothing types, these methods model the body and clothing with a holistic implicit representation. Hence, hands and faces are typically poorly reconstructed and are not articulated. Additionally, holistic models of the body and clothing do not permit virtual try-on applications, which require the body and clothing to be represented separately. While neural radiance fields (NeRF) can model the head well (e.g., which require the body and clothing to be represented separately. The NeRF is modeled in canonical space, and we use the skinning transformation from the SMPL-X body model to deform points in observation space to the canonical space. This requires accurate estimates of body shape and pose for every video frame. We estimate body pose and shape parameters with PIXIE [Feng et al. 2021a]. However, these estimates are not accurate enough, resulting in blurry reconstructions. Thus, we refine the body pose and shape during optimization. Second, the cloth deformations are not fully explained by the SMPL-X skinning, particularly in the presence of loose clothing. To overcome this, we learn a non-rigid deformation field to correct clothing deviations from the body. Third, SCARF’s hybrid representation, combining a NeRF and a mesh, requires customized volumetric rendering. Specifically, rendering the clothed body must account for the occlusions between the body mesh and the clothing layer. To integrate a mesh into volume rendering, we sample a ray from camera’s optical center until it intersects the body mesh, and accumulate the colors along the ray up to the intersection point with the colored mesh surface. Fourth, to disentangle the body and clothing, we must prevent the NeRF from capturing all image information including the body. To that end, we use clothing segmentation masks to penalize the NeRF outside of clothed regions.

In summary, SCARF automatically creates a 3D clothed human avatar from monocular video (Fig. 1) with disentangled clothing on top of the human body. SCARF offers the best of two worlds by combining different representations – a 3D parametric model for the body and a NeRF for the clothing. Based on SMPL-X, the reconstructed avatar offers animator control over body shape, pose, hand articulation, and facial expression. Since SCARF factors clothing from the body, the clothing can be extracted and transferred between avatars, enabling applications such as virtual try-on.

2 RELATED WORK

3D Bodies from images. The 3D surface of a human body is typically represented by a learned statistical 3D model [Allieck et al. 2021; Anguelov et al. 2005; Joo et al. 2018; Loper et al. 2015; Osman et al. 2020; Pavlakos et al. 2019; Xu et al. 2020]. Numerous optimization and regression methods have been proposed to compute 3D shape and pose parameters from images, videos, and scans. See [Liu et al. 2021a; Tian et al. 2022] for recent surveys. We focus on methods that capture full-body pose and shape, including the hands and facial expressions [Allieck et al. 2021; Choutas et al. 2020; Feng et al. 2021a; Pavlakos et al. 2019; Xiang et al. 2019]. Such methods, however, do not capture hair, clothing, or anything that deviates the body. Also, they rarely recover the appearance, due to large geometric discrepancy between the clothed human in images and captured body meshes. Unlike prior work, we consider clothing as an important component for human appearance and capture both the parametric body and non-parametric clothing from videos.

Capturing clothed humans from images. Clothing is more complex than body in terms of geometry, non-rigid deformation, and appearance, making the capture of clothing from images challenging. Mesh-based methods to capture clothing often use additional
Capturing both clothing and body. Several methods model clothing as a separate layer on top of the human body. They use training data produced by physics-based simulations [Bertiche et al. 2020; Patel et al. 2020; Santesteban et al. 2019; Vidalaurre et al. 2020] or require template meshes fit to 3D scans [Chen et al. 2021a; Pons-Moll et al. 2017; Tiwari et al. 2020; Xiang et al. 2021]. It is a much harder problem to recover the body and clothing from images alone, where 3D data is not provided. Jiang et al. [2020] and Zhu et al. [2020] train a multi-clothing model on 3D datasets with various clothing styles. Then during inference, a trained network produces the 3D clothing as a separate layer by recognizing and predicting the clothing style from an image. Zhu et al. [2022] fit template meshes to non-parametric 3D reconstructions. While these methods recover the clothing and body from images, they are limited in visual fidelity, as they do not capture clothing appearance. Additionally, methods with such predefined clothing style templates can not easily handle the real clothing variations, limiting their applications. In contrast, Corona et al. [2021] represent clothing layers with deep unsigned distance functions [Chibane et al. 2020], and learn the clothing style and clothing cut space with an auto-decoder. Once trained, the clothing latent code can be optimized to match image observations, but it produces over-smooth results without detailed wrinkles. Instead, SCARF models the clothing layer with a neural radiance field, and optimizes the body and clothing layer from scratch instead of the latent space of a learned model. Therefore, SCARF produces avatars with higher visual fidelity (see Section 4).

3 METHOD

SCARF extracts a clothed 3D avatar from a monocular video. SCARF enables us to synthesize novel views of the reconstructed avatar, and to animate the avatar with SMPL-X identity shape and pose control. The disentanglement of body and clothing further enables us to transfer clothing between subjects for virtual try-on applications.

Key idea. SCARF is grounded in the observation that statistical mesh models can represent human bodies well, but are ill-suited for clothing due to the large variation in clothing shape and topology (e.g., open & closed jackets, shirt, trousers, and skirts cannot be modeled with meshes of the same topology). Instead, NeRF [Mildenhall et al. 2020] offers more flexibility for modeling clothing, but is less appropriate for bodies where good models already exist. In particular, body NeRFs often lack facial details, poorly reconstruct hands, and lack fine-grained control of hand articulation and facial expression [Chen et al. 2021b; Peng et al. 2022, 2021b; Su et al. 2021]. Motivated by the strengths and weaknesses of the different representations, we use a hybrid representation that combines the strengths of body mesh models (specifically SMPL-X) with the flexibility of NeRFs; see Figure 2 for an overview.
We define the clothed body model in a canonical space, where body and clothing are represented separately.

**Body representation.** We represent the body with the expressive body model, SMPL-X [Pavlakos et al. 2019], which captures whole-body shape and pose variations, including finger articulation, and facial expressions. Given parameters for identity body shape $\mathbf{b} \in \mathbb{R}^{\beta_0}$, pose $\mathbf{\theta} \in \mathbb{R}^{\beta_0 \times 3}$, and facial expression $\mathbf{\psi} \in \mathbb{R}^{\beta_1}$, SMPL-X is defined as a differentiable function $M(\beta, \theta, \psi) \rightarrow (V, F)$ that outputs a 3D human body mesh with $n_v$ vertices $V \in \mathbb{R}^{n_v \times 3}$, and $n_f$ faces $F \in \mathbb{R}^{n_f \times 3}$. To increase the flexibility of the model, we add an additional set of vertex offsets $O \in \mathbb{R}^{n_v \times 3}$ to capture localized geometric details, and define the model as

$$M(\beta, \theta, \psi, O) = \text{LBS}(T_p(\beta, \theta, \psi, O), J(\beta), \theta, W),$$

where $T_p(\beta, \theta, \psi, O) = T + O + B(\beta, \theta, \psi)$,

$$T_p(\beta, \theta, \psi, O) = T + O + B(\beta, \theta, \psi),$$

and $B(\beta, \theta, \psi) = B_3(\beta, S) + B_p(\theta, \mathcal{P}) + B_1(\psi; E)$.

Here, $B_3(\beta, S) : \mathbb{R}^{\beta_0} \rightarrow \mathbb{R}^{n_v \times 3}$ are the identity body shapes, $B_p(\theta, \mathcal{P}) : \mathbb{R}^{3n_v \times 3} \rightarrow \mathbb{R}^{n_v \times 3}$ are the pose body shapes, and $B_1(\psi; E) : \mathbb{R}^{\beta_1} \rightarrow \mathbb{R}^{n_v \times 3}$ are the expression blend shapes with the learned identity $S$, pose $\mathcal{P}$, and expression $E$ subspaces.

Specifically, given a template vertex $t_i$, the vertex $v_i$ ($t_i$ and $v_i$ are column vectors in homogeneous coordinates) is computed as $v_i = M_i(\beta, \theta, \psi, O)t_i$, with $M_i(\cdot) \in \mathbb{R}^{4 \times 4}$ as

$$M_i(\beta, \theta, \psi, O) = \sum_{k=1}^{n_k} w_k \mathbf{G}_k(\theta, J) \begin{bmatrix} I_{4 \times 4} \end{bmatrix} 1,$$

where $w_k$ is a blend weight element of $W$, $\mathbf{G}_k(\theta, J) \in \mathbb{R}^{4 \times 4}$ is the world transformation of joint $k$, $I \in \mathbb{R}^{3 \times 3}$ is the identity matrix, and $o_i$ and $b_i(\cdot)$ are the elements of the i-th vertex of $O$ and $B_i(\cdot)$, respectively. For more details regarding the SMPL-X formulation, we refer to Pavlakos et al. [2019]. To capture more geometric details, we use an upsampling version of SMPL-X with $n_v = 38,703$ vertices and $n_f = 77,336$ faces. We obtain this by subdividing a quad mesh of the model’s template, and upsampling the blend shape bases and skinning weights using barycentric coordinates obtained from the upsampled template. As the upsampling does not increase the variability of the model, we add additional learnable vertex offsets $O$ for each sequence. Similar to Grassal et al. [2022], we use implicit models $F_d : t \rightarrow o$ to describe the offset from every vertex $t$ of $T$, and $F_e : t \rightarrow e$ to predict the RGB color of every vertex $t$.

**Clothing representation.** Due to the large variety of clothing in in-the-wild videos, we represent clothing using NeRF [Mildenhall et al. 2020] due to its ability to handle diverse topologies and transparent cloth materials. Following previous work (e.g., [Chen et al. 2021b; Peng et al. 2021a]), we define the NeRF model in canonical space as $F_c : x^c \rightarrow (c, \sigma)$ to predict RGB color $c$ and density $\sigma$ for each query point $x^c \in \mathbb{R}^3$. Note that unlike previous work that models entire clothed bodies with a NeRF (e.g., [Chen et al. 2021b; Liu et al. 2021b; Peng et al. 2021a,b; Weng et al. 2022]), we only represent clothing with a NeRF combined with an explicit surface representation for the underlying human body. The skinning articulation of a body model like SMPL-X is not sufficient to model pose-dependent clothing deformations. Following previous work ([Liu et al. 2021b; Peng et al. 2022; Weng et al. 2022]), to model pose-dependent effects, we learn a deformation function $F_m : \mathbb{R}^6 \rightarrow \mathbb{R}^3$ in the canonical space to model the residual non-rigid deformation. Specifically, given a body mesh $M(\beta, \theta, \psi, O) \rightarrow V$ and a point $x$ and $x^c$ in observation space and canonical space, respectively, we optimize the weights of an MLP $F_m : (x^c, \psi^c_{nn(x)}) \rightarrow d^c$, where $nn(x)$ is the index of the nearest neighbor vertex of $x$ in $V$. This MLP conditions $x^c$ on a vertex $v^c$ from the posed mesh $M(0, 0, 0) \rightarrow V$. Instead of $x^c$, the displaced point $x^c + d^c$ in canonical space is then input to $F_c$.

### 3.2 Canonicalization

To model the body and clothing in canonical space, we need to transfer points in observation space to the canonical space. Following Chen et al. [2021b], we use the inverse transformation of the underlying SMPL-X model to transform the query points from the body pose $\theta$ in the observation frame to the “star-like” body pose $\theta^*$ (Fig. 2) in the canonical space.

As the transformation between canonical space and observation space (Eq. 4) is only defined for surface vertices of the body model, Zheng et al. [2021] and Chen et al. [2021b] generalize the model transformation to the entire space. Formally, given a body mesh $M(\beta, \theta, \psi, O) \rightarrow V$ and a point $x$ (in homogeneous coordinates) in observation space, $x$ is transferred to canonical space with

$$\sum_{v_i \in N(x)} \frac{\omega_i(x)}{\omega_0(x)} M_i(0, \theta^*, 0, 0) (M_i(\beta, \theta, \psi, O))^{-1} x \rightarrow x^c,$$

where $N(x)$ is the set of nearest neighbor vertices of $x$ in $V$. Further, the transformations are weighted with

$$\omega_i(x) = \exp \left(-\frac{\|x - v_i\|_2}{\sigma^2} \|w_{nn(x)} - w_i\|_2\right),$$

and

$$\omega_0(x) = \sum_{v_i \in N(x)} \omega_i(x),$$

where $nn(x)$ is the index of the nearest neighbor vertex of $x$ in $V$, $w_i \in \mathbb{R}^{n_v}$ are the blend weights of $v_i$, and $\sigma$ is a constant weight.
3.3 Mesh Integrated Volume Rendering

Camera. To reconstruct SMPL-X from images, we use a scaled-orthographic camera model \( p = [s, t^T]^T \) with isotropic scale \( s \in \mathbb{R} \) and 2D translation \( t \in \mathbb{R}^2 \).

Mesh rendering. Given geometry parameters \((\beta, \theta, \varphi)\), vertex offsets \( \mathbf{O} \), colors \( \mathbf{C}_t : \mathbf{t}_i \rightarrow c_i \) for every vertex in the upsampled SMPL-X template, and camera information \( p \), we render the colored mesh into an image as \( R_m(M(\beta, \theta, \varphi, \mathbf{O}), p, c) \), where \( R_m \) denotes the differentiable rasterizer function.

Volume rendering: We follow Mildenhall et al. [2020] to use volumetric rendering. Given a camera ray \( R(t) = o + td \) with center \( o \in \mathbb{R}^3 \) and direction \( d \in \mathbb{R}^3 \), the rendering interval \( \mathbf{t} \in [t_n, t_f] \subset \mathbb{R} \) (near and far bounds) is evenly split into \( n_t \) bins. A random sample \( t_i (1 \leq i \leq n_t) \) from every bin is taken and the colors are aggregated across the ray samples \( R(t_i) \rightarrow \mathbf{r}_i \). Unlike previous work, we integrate the body model, \( M(\beta, \theta, \varphi, \mathbf{O}) \), into the volumetric rendering. Specifically, if \( R(t) \) intersects \( M \), we set the \( t_i \) such that \( R(t_i) \) is the intersection point with \( M \). In this case, we use the mesh color instead of the NeRF color \( c_{n_t} \) (see Fig. 3). Formally, the aggregated color is

\[
C(R) = \sum_{i=1}^{n_t-1} \alpha_i c_i + \mathbf{r}_c, \quad \text{with} \quad \alpha_i = \gamma_i(1 - \exp(-\sigma_i \delta_i)),
\]

where

\[
\gamma_i = \prod_{j=1}^{i-1} \exp(-\sigma_j \delta_j), \quad \text{and} \quad \tau = 1 - \sum_{i=1}^{n_t-1} \alpha_i,
\]

\[
c = \begin{cases} F_r(r^c_{n_t}), & \text{if } R(t) \text{ intersects } M, \\ c_{n_t}, & \text{otherwise.} \end{cases}
\]

Here, \( \delta_i = t_{i+1} - t_i \) is the distance between adjacent samples, \( R = \{r_1, \cdots, r_{n_t}\} \), \( F_r(r_i) \rightarrow (c_i, \sigma_i) \), and \( r^c_{n_t} \) is the canonicalized \( r_{n_t} \). For the scaled-orthographic camera, we use \( o = [o_x, o_y, 0] \) and \( d = [0, 0, 1] \) to compute the color of the pixel \( [o_x, o_y] \). We denote the image rendered by sampling rays for all image pixels as \( R_o \).

3.4 Objectives

Given a sequence of \( n_f \) images, \( I_f (1 \leq f \leq n_f) \), we optimize \( \beta \) and the weights of the MLPs \( F_d, F_c, F_t, F_m \) jointly across the entire sequence, and \( \theta_f \) and \( p_f \) per frame. The objective is

\[
L = L_{\text{recon}} + L_{\text{clothing}} + L_{\text{body}},
\]

with reconstruction loss \( L_{\text{recon}} \), clothing segmentation loss \( L_{\text{clothing}} \), and body loss \( L_{\text{body}} \). For simplicity, we omit the frame index \( f \) and the optimized parameters whenever possible. The sequence objective function is the sum over all frames.

Reconstruction loss. We minimize the difference between the rendered image and the input image as

\[
L_{\text{recon}} = \lambda_{\text{vol}} L_{\delta}(R_o - I) + \lambda_{\text{me}} L_{\text{mrf}}(R_o - I),
\]

where \( L_{\delta} \) is the Huber loss [Huber 1964], and \( L_{\text{mrf}} \) is an ID-MRF loss [Wang et al. 2018]. While the Huber loss focuses on the overall reconstruction, the ID-MRF loss allows us to reconstruct more details as previously shown by Feng et al. [2021b]. Solely minimizing \( L_{\text{recon}} \) results in a NeRF that models the entire clothed body including the non-clothing regions.

Cloth segmentation loss. Our goal is to only capture clothing with \( F_c \) instead of modeling the entire clothed body. This requires us to disentangle body and clothing. Given a clothing mask \( S_c \), which is 1 for every clothing pixel and 0 elsewhere, we minimize the clothing segmentation loss as

\[
L_{\text{clothing}} = \lambda\delta L(\|S_0 - S_c\|_{1,1}),
\]

with the rendered NeRF mask \( S_o \), which is obtained by sampling rays for all image pixels and computing per ray

\[
S(R) = \sum_{i=1}^{n_t-1} \prod_{j=1}^{i-1} \exp(-\sigma_j \delta_j) \left(1 - \exp(-\sigma_i \delta_i)\right).
\]

Minimizing \( L_{\text{clothing}} \) ensures that the aggregated density across rays (excluding the far bound) outside of clothing is 0 and therefore nothing outside of the clothing mask is modeled by the NeRF.

Human body loss. To further disentangle body and clothing, we must ensure that the body model does not catch clothing variations. For this purpose, we define different losses based on four observations.

First, the body mesh should match the masked image. Given a binary mask \( S \) of the clothed body (1 for inside, 0 elsewhere), we minimize the difference between the silhouette of the rendered body \( R_m^c(M, p) \) and the given mask as

\[
L_{\text{silhouette}} = \lambda_{\text{silhouette}} L_{\delta}(R_m^c(M, p) - S) - S).
\]

Second, the body mesh should match visible body parts. Optimizing \( L_{\text{silhouette}} \) only results in meshes that also fit the clothing, which is undesired especially for loose clothing (i.e., this leads to visible artifacts when transferring clothing between subjects). Instead, given a binary mask \( S_b \) of the visible body parts (1 for body parts, 0 elsewhere), we minimize a part-based silhouette loss

\[
L_{\text{bodymask}} = \lambda_{\text{bodymask}} L_{\delta}(S_b \odot R_m^c(M, p) - S_b),
\]

and a part-based photometric loss

\[
L_{\text{skin}} = \lambda_{\text{skin}} L_{\delta}(S_b \odot (R_m^c(M, \mathbf{c}, p) - I)),
\]

to put special emphasis on fitting visible body parts.

Third, the body mesh should stay within clothing regions, as

\[
L_{\text{inside}} = \lambda_{\text{inside}} L_{\delta}(|\text{ReLU}(R_m^c(M, \mathbf{c}, p) - S_b)|).
\]

Fourth, the skin color of occluded body vertices should be similar to non-occluded regions. For this, we assume that hands are visible for some parts of the sequence, and minimize the difference between the colors of body parts in occluded regions and the hand color as

\[
L_{\text{skinside}} = \lambda_{\text{skinside}} L_{\delta}(S_c \odot (R_m^c(M, \mathbf{c}, p) - \mathbf{c}_{\text{hand}})),
\]

where \( \mathbf{C} = [\mathbf{c}_{\text{hand}}, \cdots, \mathbf{c}_{\text{hand}}]^T \in \mathbb{R}^{n_c \times 3} \) is the tiled average color \( \mathbf{c}_{\text{hand}} \) of the hand vertices.

Regularization. We regularize the reconstructed mesh surface as

\[
L_{\text{reg}} = \lambda_{\text{edge}} L_{\text{edge}}(M) + \lambda_{\text{offset}} \|O\|_{2,2},
\]

where \( L_{\text{edge}} \) is the relative edge loss [Hirshberg et al. 2012] between the optimized body mesh \( w/ \) and \( w/o \) applied offsets. For the offset loss, we apply different weights on the body, hands and face region. For more details see the Sup. Mat.

Overall, the body loss is

\[
L_{\text{body}} = L_{\text{recon}} + L_{\text{bodymask}} + L_{\text{skin}} + L_{\text{skinside}} + L_{\text{inside}} + L_{\text{reg}},
\]
3.5 Implementation

SCARF is implemented in PyTorch, with a built-in PyTorch3D rasterizer [Ravi et al. 2020], and optimized with Adam [Kingma and Ba 2015]. For each frame, we run PIXIE [Feng et al. 2021a] to initialize \((\mathbf{b}, \mathbf{\theta}, \mathbf{\phi})\) and \(p\). For datasets without provided silhouette masks, we compute \(S\) with [Lin et al. 2022], and [Dabhi 2022] for \(S_c\). Following Mildenhall et al. [2020], we optimize both a coarse and a fine MLP to represent the NeRF. Our optimization pipeline has two stages. We first jointly optimize the canonical NeRF to estimate the entire clothed body (i.e., without clothing segmentation) and refine the SMPL-X pose for 100k iterations with a learning rate of \(5 \times 10^{-4}\). Then, we optimize the full model for another 50k iterations with learning rates of \(1 \times 10^{-5}\) for the NeRF \((F_c\) and \(F_m)\) and pose refinement, and \(1 \times 10^{-5}\) for the mesh color model \((F_t)\) and the offset \((F_d)\). For more details about the implementation, please refer to the Sup. Mat.

4 EXPERIMENTS

4.1 Datasets

We evaluate SCARF on sequences from People Snapshot [Alldieck et al. 2018b], iPER [Liu et al. 2019], SelfRecon [Jiang et al. 2022], and self-captured data. For People Snapshot, we use the provided SMPL pose as initialization instead of running PIXIE [Feng et al. 2021a]. For each subject, we use around 100-150 images for optimization. See the Sup. Mat. for more details.

4.2 Comparisons

Our method can capture the body and clothing from image sequences, enabling novel view synthesis. Previous works either model whole clothed body from video or reconstruct cloth geometry from a single image after training with plentiful 3D scan data. So we compare our method with others on two tasks: novel view synthesis and separate body and garment reconstruction from images.

Body and garment reconstruction. Similar to SCARF, SMPLicit [Corona et al. 2021] and BCNet [Jiang et al. 2020] separately model the body and clothing. Note that these methods and SCARF follow a different strategy. While they learn generative models from scans [Corona et al. 2021] or synthetic 3D data [Jiang et al. 2020] and then reconstruct the clothed body from a single image, SCARF extracts a clothed avatar from a video without 3D supervision. Figure 4 shows that SCARF reconstructs different clothing types more faithfully.

Body and cloth modeling. We quantitatively compare to NeRF [Omran et al. 2018], SMPLpix [Prokudin et al. 2021], Neural Body [Peng et al. 2021b] and Anim-NeRF [Chen et al. 2021b], following the evaluation protocol of [Chen et al. 2021b]. Table 1 shows that SCARF is more accurate than other methods under most metrics. Figure 5 provides qualitative comparisons demonstrating that SCARF better reconstructs hand and face geometry compared to SelfRecon [Jiang et al. 2022] and Anim-NeRF [Chen et al. 2021b].

4.3 Applications

Animation. Unlike previous methods that represent clothed bodies holistically, SCARF offers more fine grained control over body pose. Figure 6 shows repositioning into novel poses.

Cloth transfer. Figures 1 and 6 and the Sup. Mat. show that our hybrid representation enables transfer of clothing between avatars.
Table 1: Quantitative comparison of novel view synthesis on People-Snapshot [Alldieck et al. 2018b].

| Subject ID             | NeRF | SMPLpix | NB | Anim-NeRF | Ours | NeRF | SMPLpix | NB | Anim-NeRF | Ours | NeRF | SMPLpix | NB | Anim-NeRF | Ours |
|------------------------|------|---------|----|-----------|------|------|---------|----|-----------|------|------|---------|----|-----------|------|
| male-3-casual          | 20.64| 23.74   | 24.94| 29.37     | **30.59** | 899  | .923    | .943| .970      | **.977** | 101 | .022    | .033| .017      | .024 |
| male-4-casual          | 20.29| 22.43   | 24.71| 28.37     | **28.99** | 880  | .910    | .947| .961      | **.970** | 145 | .031    | .042| .027      | **.025** |
| female-3-casual        | 17.43| 22.33   | 23.87| 28.91     | **30.14** | 861  | .929    | .950| .974      | **.977** | 170 | .027    | .035| .022      | .028 |
| female-4-casual        | 17.63| 23.35   | 24.37| 28.90     | **29.96** | 858  | .926    | .945| .968      | **.972** | 183 | .024    | .038| .017      | .026 |

Pose initialization. SCARF refines the body pose during optimization. However, it may fail if the initial pose is far from the right pose. Handling difficult poses where PIXIE [Feng et al. 2021a] fails requires more robust 3D body pose estimator.

Dynamics. SCARF handles non-rigid cloth deformation with the pose-conditioned deformation model. While the global pose accounts for some deformation, the modeling of clothing dynamics as a function of body movement is the subject of future work.

Lighting. As with other NeRF methods, we do not factor lighting and material properties. This results in baked-in shading and the averaging of specular reflections across frames. Factoring lighting from shape and material is a key next step to improve realism.

Facial expressions. SCARF uses the facial expressions estimated by PIXIE [Feng et al. 2021a] which is unable to capture the full spectrum of emotions (cf. [Danecek et al. 2022]). Also, we have not exploited NeRF to capture changes in facial appearance, e.g. due to the mouth opening. We believe this is a promising future direction.

6 CONCLUSION
SCARF automatically extracts an animatable clothed 3D human avatar from a monocular video. Our key novelty is a hybrid representation that combines a mesh-based body model with a neural radiance field to separately model the body and clothing. This factored representation enables SCARF to transfer clothing between avatars, animate the body pose of the avatars including finger articulation, alter their body shape (see Sup. Mat.) and facial expression, and visualize them from unseen viewing directions. This property makes SCARF well suited to VR and virtual try-on applications. Finally, SCARF outperforms existing avatar extraction methods from videos in terms of visual quality and generality.

ACKNOWLEDGMENTS
We thank Sergey Prokudin, Weiying Liu, Yuliang Xiu, Songyou Peng, Qianli Ma for fruitful discussions, and Peter Kulitz, Zhen Liu, Yandong Wen, Hongwei Yi, Xu Chen, Soubhik Sanyal, Omri Ben-Dov, Shashank Tripathi for proofreading. We also thank Betty Mohler, Sarah Danes, Natalia Marciniak, Tsvetelina Alexiadi, Claudia Gallatz, and Andres Camilo Mendoza Patino for their supports with data. This work was partially supported by the Max Planck ETH Center for Learning Systems.

Disclosure: MJB has received research gift funds from Adobe, Intel, Nvidia, Meta/Facebook, and Amazon. MJB has financial interests in Amazon, Datagen Technologies, and Meschapade GmbH. While TB is part-time employee of Amazon, this research was performed solely at, and funded solely by, MPI.
Capturing and Animation of Body and Clothing from Monocular Video

Yu Rong, Takaaki Shiratori, and Hanbyul Joo. 2021. FrankMocap: A Monocular 3D Whole-Body Pose Estimation System via Regression and Integration. In International Conference on Computer Vision Workshops (ICCV-W).

Shunsuke Suto, Zeng Huang, Ryoita Natsume, Shigeo Morishima, Angsuo Kanazawa, and Hao Li. 2019. PFu: Pixel-Aligned Implicit Function for High-Resolution Clothed Human Digitization. In International Conference on Computer Vision (ICCV).

Shunsuke Suto, Tomas Simon, Jason Saragih, and Hanbyul Joo. 2020. PFuID: Multi-Level Pixel-Aligned Implicit Function for High-Resolution 3D Human Digitization. In Conference on Computer Vision and Pattern Recognition (CVPR).

Igor Santesteban, Miguel A Otaduy, and Dan Casas. 2019. Learning-based animation of clothing for virtual try-on. In Computer Graphics Forum, Vol. 38. Wiley Online Library, 355–366.

Shih-Yang Su, Frank Yu, Michael Zollhöfer, and Helge Rhodin. 2021. A-NeRF: Articulated neural radiance fields for learning human shape, appearance, and pose. Advances in Neural Information Processing Systems (NeurIPS) 34 (2021).

Yating Tian, Hongwen Zhang, Yebin Liu, and Limin Wang. 2022. Recovering 3D Human Mesh from Monocular Images: A Survey. arXiv preprint arXiv:2203.01923 (2022).

Garvita Tiwari, Bharat Lal Bhatnagar, Tony Tung, and Gerard Pons-Moll. 2020. SIZER: A dataset and model for parsing 3D clothing and learning size sensitive 3D clothing. In European Conference on Computer Vision (ECCV). Springer, 1–18.

Raquel Vidaurre, Igor Santesteban, Elena Garces, and Dan Casas. 2020. Fully Convolutional Graph Neural Networks for Parametric Virtual Try-On. In Computer Graphics Forum, Vol. 39. Wiley Online Library, 145–156.

Yi Wang, Xin Tao, Xiaojuan Qi, Xiaoyong Shen, and Jiaya Jia. 2018. Image inpainting via generative multi-column convolutional neural networks. In Advances in Neural Information Processing Systems (NeurIPS). 331–340.

Chung-Yi Weng, Brian Curless, Pratul P. Srinivasan, Jonathan T. Barron, and Ira Kemelmacher-Shlizerman. 2022. HumanNeRF: Free-Viewpoint Rendering of Moving People From Monocular Video. In Conference on Computer Vision and Pattern Recognition (CVPR). 16219–16220.

Zongxuan Yang, Lu Yang, Chengjie Wu, and Zheng Zhang. 2021. GarmentFusion: A Dataset and Benchmark for Monocular Garment Parsing. In Conference on Computer Vision and Pattern Recognition (CVPR). 1515–1524.

Lu Yang, Qing Song, Zhihui Wang, Mengjie Hu, Chun Liu, Xuexi Xin, Wenhe Jia, and Songcen Xu. 2020. Renovating Parsing R-CNN for Accurate Multiple Human Parsing. In European Conference on Computer Vision (ECCV) (Lecture Notes in Computer Science, Vol. 12357). Springer, 421–437.

Ze Yang, Shenlong Wang, Sivabalan Manivasagam, Zeng Huang, Wei-Chiu Ma, Xinchen Yan, Erzin Yumer, and Raquel Urtasun. 2021. S3: Neural shape, skeleton, and skinning fields for 3D human modeling. In Conference on Computer Vision and Pattern Recognition (CVPR). 13284–13293.

Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. 2021. Volume rendering of neural implicit surfaces. Advances in Neural Information Processing Systems (NeurIPS) 34 (2021).

Lior Yariv, Yoni Kasten, Dror Moran, Meirav Galun, Matan Atzmon, Barri Ronen, and Yaron Lipman. 2020. Multiview neural surface reconstruction by disentangling geometry and appearance. Advances in Neural Information Processing Systems (NeurIPS) 33 (2020), 2492–2502.

Andrei Zanfir, Eduard Gabriel Bazavan, Mihai Zanfir, William T Freeman, Rahul Sukthankar, and Cristian Sminchisescu. 2021. Neural Descent for Visual 3D Human Pose and Shape. In Conference on Computer Vision and Pattern Recognition (CVPR). Computer Vision Foundation / IEEE, 14484–14493.

Lior Yariv, Yoni Kasten, Dror Moran, Meirav Galun, Matan Atzmon, Barri Ronen, and Yaron Lipman. 2020. Multiview neural surface reconstruction by disentangling geometry and appearance. Advances in Neural Information Processing Systems (NeurIPS) 33 (2020), 2492–2502.

Andrei Zanfir, Eduard Gabriel Bazavan, Mihai Zanfir, William T Freeman, Rahul Sukthankar, and Cristian Sminchisescu. 2021. Neural Descent for Visual 3D Human Pose and Shape. In Conference on Computer Vision and Pattern Recognition (CVPR). Computer Vision Foundation / IEEE, 14484–14493.

Ze Yang, Shunlong Wang, Sivabalan Manivasagam, Zeng Huang, Wei-Chiu Ma, Xinchen Yan, Erzin Yumer, and Raquel Urtasun. 2021. S3: Neural shape, skeleton, and skinning fields for 3D human modeling. In Conference on Computer Vision and Pattern Recognition (CVPR). 13284–13293.

Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. 2021. Volume rendering of neural implicit surfaces. Advances in Neural Information Processing Systems (NeurIPS) 34 (2021).

Lior Yariv, Yoni Kasten, Dror Moran, Meirav Galun, Matan Atzmon, Barri Ronen, and Yaron Lipman. 2020. Multiview neural surface reconstruction by disentangling geometry and appearance. Advances in Neural Information Processing Systems (NeurIPS) 33 (2020), 2492–2502.

Andrei Zanfir, Eduard Gabriel Bazavan, Mihai Zanfir, William T Freeman, Rahul Sukthankar, and Cristian Sminchisescu. 2021. Neural Descent for Visual 3D Human Pose and Shape. In Conference on Computer Vision and Pattern Recognition (CVPR). Computer Vision Foundation / IEEE, 14484–14493.

Zerong Zheng, Tao Yu, Yebin Liu, and Qinghui Dai. 2021. PaMIR: Parametric model-conditioned implicit representation for image-based human reconstruction. Transactions on Pattern Analysis and Machine Intelligence (PAMI) (2021).

Heming Zha, Yu Cao, Hong Jin, Weiikai Chen, Dong Du, Zhangye Wang, Shuguang Cui, and Xiaoqiang Han. 2020. Deep Fashion3D: dataset and benchmark for 3D garment reconstruction from single images. In European Conference on Computer Vision (ECCV). Springer, 512–530.

Heming Zha, Lingteng Qiu, Yuda Qu, and Xiaoqiang Han. 2022. Registering Explicit to Implicit: Towards High-Fidelity Garment mesh Reconstruction from Single Images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 3845–3854.