Explicit Clothing Modeling for an Animatable Full-Body Avatar

DONGLAI XIANG, Carnegie Mellon University, USA and Facebook Reality Labs Research, USA
FABIAN PRADA, Facebook Reality Labs Research, USA
TIMUR BAGAUTDINOV, Facebook Reality Labs Research, USA
WEIPENG XU, Facebook Reality Labs Research, USA
YUAN DONG, Facebook Reality Labs Research, USA
HE WEN, Facebook Reality Labs Research, USA
JESSICA HODGINS, Carnegie Mellon University, USA and Facebook AI Research, USA
CHENGLEI WU, Facebook Reality Labs Research, USA

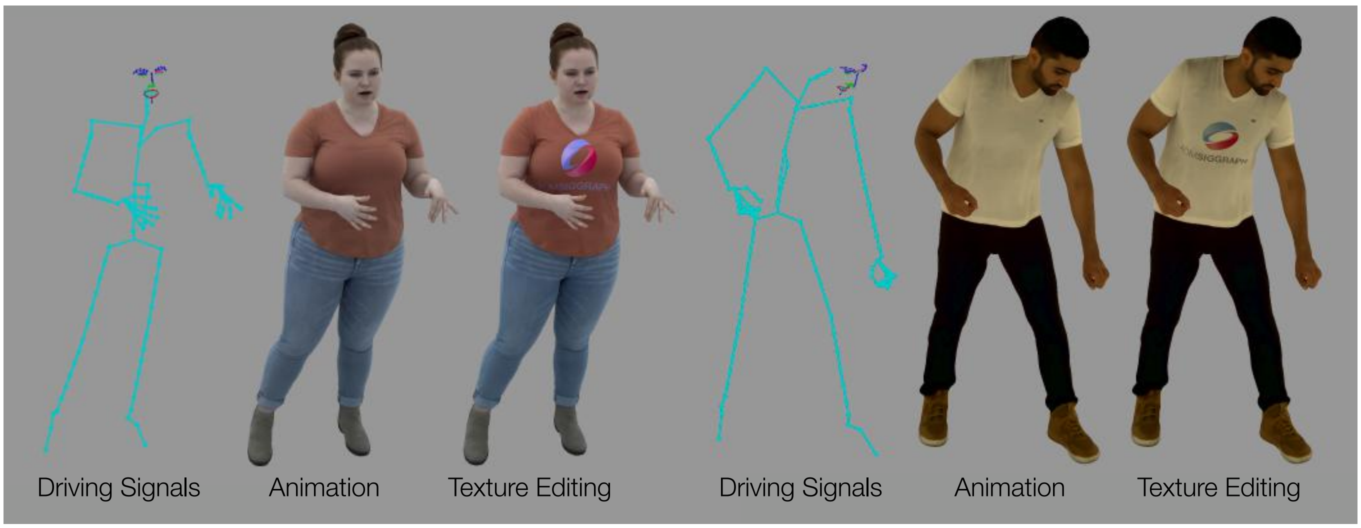

Fig. 1. Given a novel sequence of skeletal poses and facial keypoints as input, our proposed two-layer codec avatars produce photorealistic animation output, where the clothing texture can be consistently edited. From left to right, we show driving signals, animation output and editing results for two subjects.

Recent work has shown great progress in building photorealistic animatable full-body codec avatars, but these avatars still face difficulties in generating high-fidelity animation of clothing. To address the difficulties, we propose a method to build an animatable clothed body avatar with an explicit representation of the clothing on the upper body from multi-view captured videos. We use a two-layer mesh representation to separately register the 3D scans with templates. In order to improve the photometric correspondence across different frames, texture alignment is then performed through inverse rendering of the clothing geometry and texture predicted by a variational autoencoder. We then train a new two-layer codec avatar with separate modeling of the upper clothing and the inner body layer. To learn the interaction between the body dynamics and clothing states, we use a temporal convolution network to predict the clothing latent code based on a sequence of input skeletal poses. We show photorealistic animation output for three different actors, and demonstrate the advantage of our clothed-body avatars over single-layer avatars in the previous work. We also show the benefit of an explicit clothing model which allows the clothing texture to be edited in the animation output.

Work done during Donglai Xiang's internship at Facebook Reality Labs Research.

Authors' addresses: Donglai Xiang, Carnegie Mellon University, USA, Facebook Reality Labs Research, USA, donglaix@cs.cmu.edu; Fabian Prada, Facebook Reality Labs Research, USA, fabianprada@fb.com; Timur Bagautdinov, Facebook Reality Labs Research, USA, timurb@fb.com; Weipeng Xu, Facebook Reality Labs Research, USA, xuweipeng@fb.com; Yuan Dong, Facebook Reality Labs Research, USA, ydong142857@fb.com; He Wen, Facebook Reality Labs Research, USA, hewen@fb.com; Jessica Hodgins, Carnegie Mellon University, USA, Facebook AI Research, USA, jkh@cs.cmu.edu; Chenglei Wu, Facebook Reality Labs Research, USA, chenglei@fb.com.

0730-0301/2021/7-ART $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

ACM Reference Format:
Donglai Xiang, Fabian Prada, Timur Bagautdinov, Weipeng Xu, Yuan Dong, He Wen, Jessica Hodgins, and Chenglei Wu. 2021. Explicit Clothing Modeling for an Animatable Full-Body Avatar. *ACM Trans. Graph.* 1, 1 (July 2021), 14 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Animatable photorealistic digital humans are a key capability for enabling social telepresence, with the potential to open up a new way for people to remain connected without the constraint of space.

The problem of building digital human avatars has long been studied in the computer graphics community. Early work on human body modeling built low-dimensional geometric representations of the body surface with minimal clothing [Loper et al. 2015; Osman et al. 2020; Romero et al. 2017]. As a separate field of work, cloth simulation has been studied and used to create large clothing deformation that does not conform tightly to the human body [Baraff and Witkin 1998; Kavan et al. 2011; Narain et al. 2012]. However, both these lines of work focus on modeling only the geometry, and cannot directly produce photorealistic rendering output. Even with the recent data-driven methods using neural networks [Lahner et al. 2018], animating photorealistic clothed human is still far from a solved problem.

In this work, we seek to build photorealistic full-body clothed avatars that can be animated with driving signals that can be easily accessible, for example, 3D body pose and facial keypoints. Simultaneously modeling both geometry and texture with a deep generative model, like Variational Autoencoders (VAE), has been demonstrated to be an effective way to create photorealistic face avatars [Lombardi et al. 2018]. Recently, [Bagautdinov et al. 2021] extends this approach to model full-body avatars with VAE, conditioned on body pose and facial keypoints. Because these conditional signals cannot uniquely describe all the body states, like clothing, hair and gaze, the VAE latent code is used distinguish between the multiple possible body stages. In addition, it is essential to impose disentanglement between driving signals and the latent code, in order to reduce the spurious correlations between them.

Despite the progress in previous work [Bagautdinov et al. 2021], challenges still remain in building high-fidelity animatable full-body avatars, and we identify the modeling of clothing as one major difficulty. Artifacts can be observed including the imperfect correlation between body pose and clothing state, ghosting effects along the boundary between clothing and skin, as well as severe loss of wrinkle details and dynamics. These artifacts become more noticeable when the captured clothing is loose and exhibits high dynamics. On one hand, due to the registration challenge, the network may underfit the data, unable to reproduce high-frequency clothing detail; on the other hand, in spite of the disentanglement, the network may still overfit, capturing unwanted chance correlation between the driving signal and the clothing state.

In this work, we explicitly represent the body and clothing as separate layers of meshes in a codec avatar. We notice that this leads to several benefits. First, this allows us to accurately register both body and clothing, especially with our newly developed photometric tracking approach that uses inverse rendering to align clothing texture to a reference. Second, modeling the body and clothing in separate layers alleviates the aforementioned problem of chance correlation for a single-layer avatar, as the separate layers are naturally disentangled from each other. With our two-layer VAE, a single frame of joint angle can well describe the body state, while the clothing dynamics can be inferred from the sequences of poses with a Temporal Convolutional Network (TCN), which evolves the clothing state in a way that is consistent with the body motion. Third, thanks to the explicit modeling of clothing, our animation output can be further edited, for example, by changing the clothing texture. This opens up the potential to change the appearance of full-body avatars.

To summarize, our contributions are as follows:

- We present an animatable two-layer codec avatar model for photorealistic full-body telepresence; our proposed avatar can produce more temporally coherent animation with sharper boundaries and fewer ghosting artifacts compared to the single-layer avatar;
- Based on inverse rendering with our proposed two-layer codec avatar, we present a photometric tracking algorithm that aligns the salient clothing texture, significantly improving correspondence in the registered clothing meshes;
- We demonstrate the application of our two-layer codec avatar through editing of the clothing texture that is hard to achieve by the single-layer model used in previous work.

We evaluate the proposed pipeline on the captured clothed body sequences of three different actors. We demonstrate the effectiveness and benefit of our proposed method against alternative approaches. We show that our model, with only a sequence of poses and facial keypoints as input, achieves high-quality body animation and rendering with photorealistic clothing that can be viewed from arbitrary viewpoints.

2 RELATED WORK

Our goal in this paper is to build a realistic virtual avatar of a person which can be animated by driving signals of skeletal poses and facial keypoints to create a telepresence experience. The **classical pipeline** for modeling such an animatable avatar typically relies on building a textured template mesh from a 3D scan and rigging the template mesh to a parameterized skeleton model such that the deformation of the template mesh is associated with the skeletal pose according to the skinning weights. The most commonly used skinning method is the Linear Blend Skinning (LBS), which we also use to model the skeletal motion in our method. In the literature, many methods have been developed in order to reduce the unnatural skinning artifacts that occur with LBS, e.g., [Kavan et al. 2008; Kavan and Zara 2005]. However, a fundamental disadvantage of these approaches is that high-frequency deformations of skin and clothing, such as muscle bulging, folds, and wrinkles, cannot be precisely modeled. In order to solve this problem, pose dependent blend shapes [Lewis et al. 2000] have been proposed to reduce the skinning artifacts, which are corrective shapes that can be interpolated with respect to the pose and added to the skinned mesh. Although the blend shapes work well for skin and tight clothes, the non-rigid deformation of soft tissue and loose clothes usually lacks realism. Furthermore, this approach typically relies on a significant amount of manual sculpting from artists to build the blend shapes.

Physical simulation provides an automatic way to create secondary motion of the virtual characters, such as muscle bulging and cloth deformation. Cloth simulation is typically not real-time due to the computational complexity and therefore many of the earlier methods focus on the efficiency [Gillette et al. 2015; Goldenthal et al. 2007; Kavan et al. 2011; Kim et al. 2013; Wang et al. 2010]. More recent research tackles the efficiency problem by learning the mapping from body pose to the clothing deformation produced by physical simulations [Bertiche et al. 2020a,b; Chentanez et al. 2020; Gundogdu et al. 2019; Jin et al. 2020; Patel et al. 2020; Santesteban et al. 2019; Vidaurre et al. 2020; Zhang et al. 2020]. Cloth simulation has been leveraged in human performance capture to produce more realistic dynamic deformation of the clothing. Stoll and colleagues reconstruct a time-varying surface geometry of the clothing from multiview video recordings and then estimate the parameters of a physical simulation model of the clothing [Stoll et al. 2010]. SimulCap contributes a monocular human performance capture system that not only captures the skeleton motion but also simulates cloth dynamics and the cloth-body interactions [Yu et al. 2019].

Data-driven human modeling has been leveraged very effectively in recent years. The seminal work, SCAPE [Anguelov et al. 2005], learns parametrized human body shape from a large-scale dataset of 3D scans. A variation of SCAPE that integrates the learned pose dependent blend shapes, SMPL [Loper et al. 2015], has been widely used for human modeling and pose estimation. However, these models can only model a human body dressed in perfectly skin tight clothing. In order to synthesize the deformation of clothing, apart from the aforementioned simulation-based learning approaches, many methods resort to learning the deformation from real 4D capture data. DeepWrinkle [Lahner et al. 2018] consists of two modules that learn the global cloth deformation in a PCA subspace as well as high frequency details, such as finer wrinkles, on a normal texture. Similarly, Ma and colleagues also learn the pose-dependent clothing shape from 4D scans, but their graph convolution-based method directly operates on the mesh instead of the texture space [Ma et al. 2020]. Another family of generative human modeling methods do not focus on the 3D geometry, but aim at synthesizing photo-realistic human images. These neural rendering approaches typically formulate the task as an image translation problem, and learn the mapping from joint heatmaps [Aberman et al. 2019], rendered skeleton [Chan et al. 2019; Esser et al. 2018; Pumarola et al. 2018; Si et al. 2018], or rendered meshes [Liu et al. 2019c,b; Prokudin et al. 2021; Raj et al. 2020; Sarkar et al. 2020; Wang et al. 2018], to real images. In contrast to these approaches, Deep Appearance Models [Lombardi et al. 2018] explicitly handle *both* facial geometry and appearance in the form of view-dependent texture, and is capable of producing view-dependent effects and correcting geometric artifacts. In recent work [Bagautdinov et al. 2021], Bagautdinov et al. extend deep appearance models to full bodies. However, as this method does not explicitly model clothing, it may struggle in settings where clothing is more loose and exhibits more dynamics.

Dynamic scene capture is an alternative yet less practical approach for telepresence, because it does not compress the dynamic information of the scene as a latent code like ours, and therefore requires a much larger communication bandwidth. That said, our method is still highly related to these methods, as we rely on dynamic scene capture to obtain training data. Most of the existing approaches rely on multi-camera systems to recover the detailed geometry using silhouettes or photometric stereo. They reconstruct either the shapes of each individual time step [Matusik et al. 2000; Starck and Hilton 2007; Waschbüsch et al. 2005], or a temporally coherent shape by deforming a template to match the multi-view constraints [Carranza et al. 2003; De Aguiar et al. 2008]. While

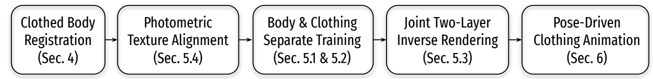

Fig. 2. An overview of our proposed method in procedural order.

some of the methods work for general scenes, many of them are dedicated to human bodies [Bray et al. 2006; Brox et al. 2010; Gall et al. 2009; Liu et al. 2011; Mustafa et al. 2015; Vlasic et al. 2008; Wu et al. 2013, 2012]. In recent years, many attempts have been made to alleviate the requirement of multi-camera systems by using depth sensors [Bogo et al. 2015; Guo et al. 2015; Helten et al. 2013; Li et al. 2009; Zhang et al. 2014] or even a monocular RGB camera [Habermann et al. 2019; Huang et al. 2017; Xu et al. 2018]. Although compelling results have been demonstrated, these approaches are fundamentally ill-posed and suffer from occlusion and depth ambiguities. Furthermore, in contrast to our method, they typically treat the character as a topologically connected template, and therefore are not able to handle movement of the clothing, such as sliding of the sleeves on the arms. Another line of work specifically focuses on capturing clothing deformations [Bradley et al. 2008; Chen et al. 2015; Pons-Moll et al. 2017; Xiang et al. 2020; Zhou et al. 2013]. For instance, ClothCap [Pons-Moll et al. 2017] automatically segments the different pieces of clothing and tracks the deformation of the clothing over time from 4D scans. Zhang and colleagues recover the detailed body shape under the clothing [Zhang et al. 2017]. Our approach highly relies on these two methods for the generation of training data. More recently, multiple approaches have been proposed to capture human appearance by modeling the radiance field with a deep neural network [Park et al. 2020; Peng et al. 2020; Pumarola et al. 2020; Wang et al. 2020]. These methods can synthesize photorealistic novel views of the captured scene or human subject, but unlike our work, cannot be used as animatable virtual avatars.

3 METHOD OVERVIEW

Our goal in this paper is to build full-body clothed digital avatars that enable photorealistic rendering from any view-point. To make the avatars useful, they should be animatable given some driving signals that can be obtained at modest cost. We choose 3D skeletal joint angles and facial keypoints as the input conditioning, similar to the previous work [Bagautdinov et al. 2021].

The central idea of our method is to explicitly represent body and clothing as two separate layers. This approach is motivated by several reasons. First, we notice that the deformation of body and clothing follows very different pattern of dynamics. A single frame of joint angles in the driving signal can majorly determine the body state through Linear Blending Skinning (LBS) and pose-dependent deformation; by contrast, the dynamics of clothing can vary too much to be described only by current body pose without considering the temporal information. Thus the body and clothing layers need to be controlled by different input conditioning. Second, in the single-layer registration of body with clothing, a specific vertex along the clothing boundary can be inaccurate. In different frames, it can belong to either the body region or the clothing

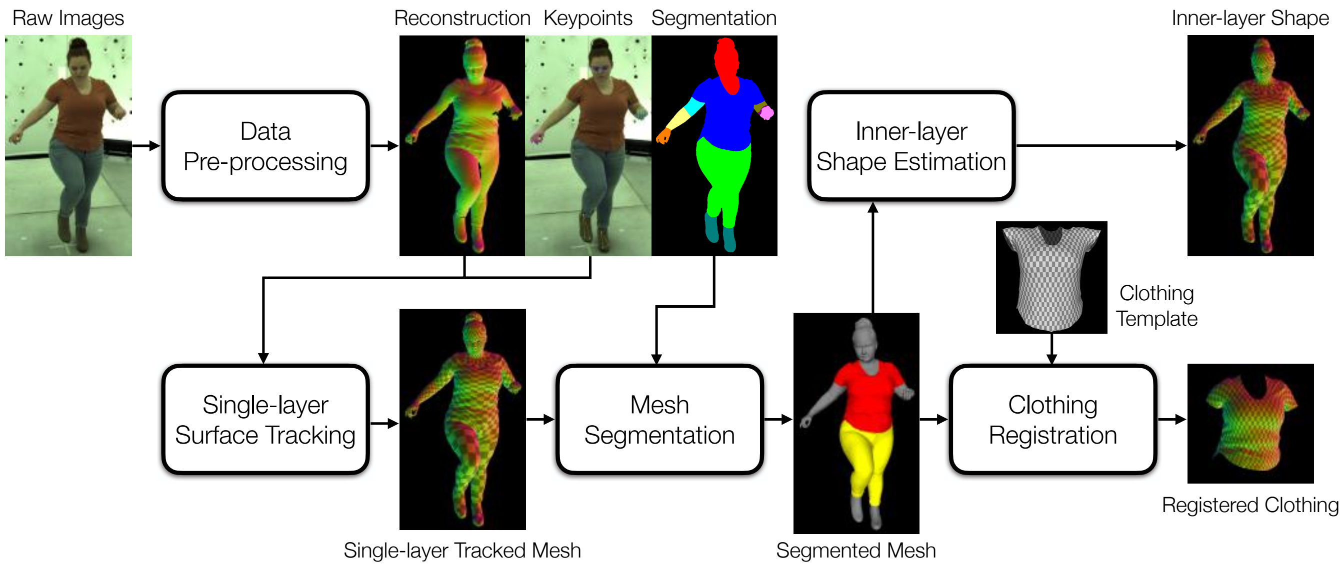

Fig. 3. The clothed body registration pipeline that we use to generate necessary training data for our two-layer codec avatars.

region due to the relative sliding motion of the clothing, which violates the single layer assumption. A codec avatar trained with such data often has blended color of clothing and skin in such a region, leading to ghosting effects around the sleeves and neck of the garment. Although disentanglement could alleviate this kind of artifacts, it cannot guarantee due to limited training data capturing the interaction between clothing and the body. In our work, with the registration of both and clothing in separate layers, such artifacts can be avoided because each vertex plays a consistent role of body or clothing across all different frames. Third, the representation of body and clothing in separate layers opens up opportunities for further changing the appearance of the avatar, such as temporally consistent editing of clothing texture of the clothing without interfering with the body appearance. This capability also points to the possible direction of altering the clothing style through physical simulation, which we leave for future work.

In this work, we assume that the subject to be modeled wears a T-shirt on the upper body and pants on the lower body. In the outer layer, we only model the T-shirt, which exhibits most of the dynamics accounting for variations in both geometry and texture. In the inner layer, we model the invisible body region covered by the outer layer and the rest of human surface, including the torso, head, arms, pants[1] and shoes.

The rest of method sections are arranged as follows. In Section 4, we briefly describe our two-layer geometry-based surface registration method to generate the necessary training data for our proposed codec avatars. In Section 5, we present our proposed two-layer codec avatars. We describe the architecture of the body branch in Section 5.1 and clothing branch in Section 5.2 respectively, as well as the joint training of both branches through inverse rendering in Section 5.3. In Section 5.4, we propose a method for texture alignment to improve the photometric correspondences between registered clothing meshes across different frames. In Section 6, we present the temporal model used to animate our clothed avatars using sequence of joint angles from the driving signal as input. A visualization of the whole method is shown in Fig. 2.

4 CLOTHED BODY REGISTRATION

In this section, we briefly describe our pipeline to generate the data required to train our two-layer codec avatars as illustrated in Fig. 3. Our goal here is to register the body and clothing geometry in two separate layers. A more detailed description of this pipeline can be found in the supplementary document.

Data preprocessing. The raw input to our pipeline is a sequence of RGB images of the subject captured by a synchronized multi-camera system. The raw RGB images are then used for dense 3D reconstruction of the human surface by a multi-view patchmatch reconstruction algorithm [Galliani et al. 2015]. One example of the reconstructed mesh can be seen in Fig. 3. In addition, we obtain part segmentation of different body and clothing region for each captured image. We also run 2D keypoint detection for body, face and hands, which are triangulated to obtain 3D keypoints.

Single-Layer Surface Tracking. We non-rigidly register the reconstructed meshes with a kinematic body model, similar to [Zhang et al. 2017] and [Walsman et al. 2017]. We estimate a personalized rest-state shape and a set of of joint angles for each frame by minimizing the difference between the LBS output and the reconstructed surface, as well as the 3D keypoints in the previous step. We further perform free-form ICP registration using the skinned kinematic model as initialization.

Mesh Segmentation. In this step, we segment the single-layer tracked meshes into body and different clothing parts. We unproject

[1] The pants of the captured subjects in this work are tight and thus not worth the effort of modeling as a separate layer. We demonstrate in the results that the advantage of clothing modeling as a separate layer is obvious when the garment is loose.

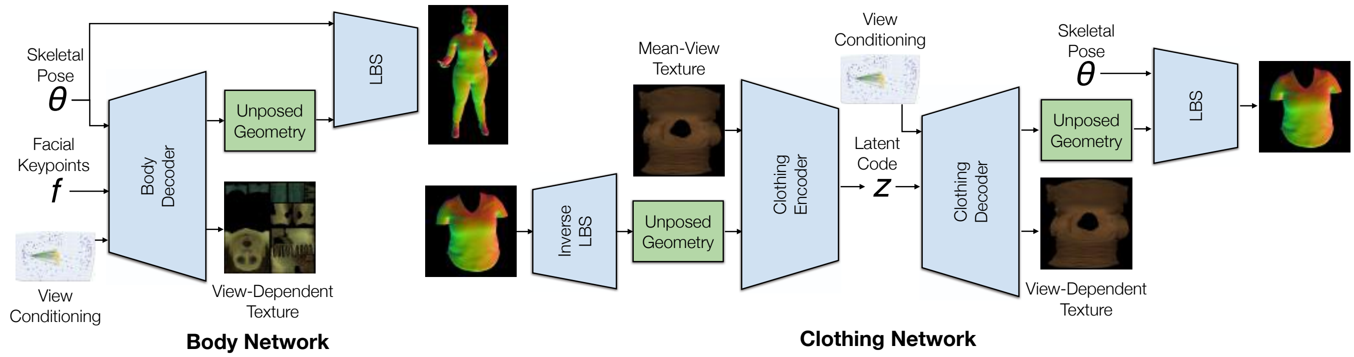

Fig. 4. Network architecture of our two-layer full-body codec avatar. We show the body network on the left and the clothing network on the right, including the input and output of each network.

the image segmentation labels onto the mesh and for each vertex take the majority of votes across different views. Similar to [Pons-Moll et al. 2017], we also run Markov Random Field (MRF) to remove noisy segmentation labels.

Clothing Registration. Our clothing registration step is similar to [Pons-Moll et al. 2017]. We manually create a template clothing mesh and use it to register the clothing region of the single-layer tracked mesh for each frame. Essentially we run a non-rigid Iterative Closest Points (ICP) algorithm that aligns the the template and target clothing region. To provide good initialization for the optimization, we find it useful to apply biharmonic deformation fields [Jacobson et al. 2010] which generate a deformed template mesh whose boundary is directly aligned with the target clothing boundary with the lowest possible interior distortion.

Inner-Layer Shape Estimation. The inner layer geometry consists of two parts: the invisible body region covered by the upper clothing, which we estimate using the method in [Zhang et al. 2017], and other visible region of the human surface, which can be directly obtained by matching with the single-layer tracking results. Different from [Zhang et al. 2017], we only need to estimate the underlying body shape of the upper body, since the pants are treated as part of the inner layer in this work.

5 CLOTHED BODY MODELING

With the registered meshes for body and clothing, we now present our two-layer codec avatars with explicit clothing modeling. Similar to [Lombardi et al. 2018] and [Bagautdinov et al. 2021], we employ a Variational Autoencoder (VAE) as our generative model. With our two-layer formation, we learn a separate decoder for body and clothing respectively. In our two-layer formulation, the two decoders learn the deformation space for body and clothing separately, while the correlation between body and clothing can be learned afterwards with a temporal model for animation. To this end, we train a body decoder which takes the skeletal pose as input, and predicts geometry and view-conditioned texture for the inner body layer, as shown on the left of Fig. 4. Similarly, we train a clothing decoder with VAE, whose network structure can be found in the right part of Fig. 4. Notice that, similar to many existing body modeling work [Loper et al. 2015; Osman et al. 2020], we only learn the geometry in the canonical pose space for both the body layer and the clothing layer by applying the inverse LBS transform. This effectively reduces the deformation space that needs to be learned. In the following sections, we introduce the detailed structure for the body decoder, the clothing decoder, and how we train them.

5.1 Body Decoder

As shown on the left of Fig.4, our body network is similar to the decoder structure in [Bagautdinov et al. 2021], without have an encoder part. We observe that once clothing is decoupled from body, the skeletal pose and facial keypoints contain sufficient information to describe the body states (including pants that are relatively tight). We do not use a latent code as conditioning for the body network to avoid the difficult problem of disentanglement between latent space and driving signal, as described in [Bagautdinov et al. 2021]. Our body decoder takes in skeletal pose, facial keypoints and view-conditioning as input, produces unposed geometry in a UV positional map and view-dependent texture for the body as output. LBS transformation is then applied to the unposed mesh restored from the UV map to produce the final output mesh.

Our loss function to train the body network is defined as:

$$\begin{aligned} E^B_{\text{train}} &= \lambda_g \|\mathrm{V}^{\mathrm{p}}_B - \mathrm{V}^{\mathrm{r}}_B\|^2 + \lambda_{lap}\|\mathrm{L}(\mathrm{V}^{\mathrm{p}}_B) - \mathrm{L}(\mathrm{V}^{\mathrm{r}}_B)\|^2 \\ &+ \lambda_t \|(\mathrm{T}^{\mathrm{p}}_B - \mathrm{T}^{\mathrm{t}}_B) \odot M^{\mathrm{V}}_B\|^2, \end{aligned} \tag{1}$$

where $\mathrm{V}^{\mathrm{p}}_B$ is the vertex position interpolated from the predicted position map in UV, and $\mathrm{V}^{\mathrm{r}}_B$ is the vertex from inner layer registration from Sec. 4, $L(\cdot)$ is the Laplacian operator, $\mathrm{T}^{\mathrm{p}}_B$ is the predicted texture, $\mathrm{T}^{\mathrm{t}}_B$ is the reconstructed texture per-view, and M^{V}_B is the mask indicating the valid UV region.

5.2 Clothing Network

As shown on the right of Fig. 4, we model the clothing appearance with a conditional variational autoencoder (cVAE). The encoder takes as input the unposed clothing geometry and mean-view texture, and produces parameters of a Gaussian distribution, from which a latent code z is sampled. Besides the latent code, the decoder also takes spatial-varying view conditioning as input, and predicts geometry and texture for the clothing. Then, the training loss is described as:

$$E^{C}_{\text{train}} = \lambda_g \|\mathrm{V}^{\mathrm{p}}_C - \mathrm{V}^{\mathrm{r}}_C\|^2 + \lambda_{lap}\|\mathrm{L}(\mathrm{V}^{\mathrm{p}}_C) - \mathrm{L}(\mathrm{V}^{\mathrm{r}}_C)\|^2 + \lambda_t \|(\mathrm{T}^{\mathrm{p}}_C - \mathrm{T}^{\mathrm{t}}_C) \odot M^{\mathrm{V}}_C\|^2 + \lambda_{\mathrm{kl}} E_{\mathrm{kl}}, \tag{2}$$

where $\mathrm{V}^{\mathrm{p}}_C$, $\mathrm{V}^{\mathrm{t}}_C$, $\mathrm{T}^{\mathrm{p}}_B$, $\mathrm{T}^{\mathrm{t}}_B$, and M^{V}_C are all defined similarly to the parameters in the body decoder but with respect to clothing, E_{kl} is a conventional KL divergence loss.

5.3 Inverse Rendering with Two-layer Representation

ICP-based clothing registration algorithm in Section 4 and previous work [Pons-Moll et al. 2017] aims to align the boundary of clothing template with the target area, while there is no explicit constraint for the interior correspondences except for the mesh regularization. Therefore, the registered meshes from Sec. 4 may suffer from correspondence errors in the interior (see the first column of Fig. 7), which significantly influences the decoder quality, especially for the clothing because of its dynamics. In order to correct the correspondences in the training stage, we need to link the predicted geometry and texture to the input multi-view images in a differentiable way. To this end, after the body and clothing networks are separately trained as described in Sec. 5.1 and 5.2, we jointly train the body and clothing networks by rendering the output with a differentiable renderer. We use the following loss functions:

$$E^{\mathrm{inv}}_{\text{train}} = \lambda_i \|\mathrm{I}^{\mathrm{R}} - \mathrm{I}^{\mathrm{C}}\| + \lambda_m \|\mathrm{M}^{\mathrm{R}} - \mathrm{M}^{\mathrm{C}}\| + \lambda_v E_{\text{softvisi}} + \lambda_{lap} E_{\text{lap}}, \tag{3}$$

where I^{R} and I^{C} are the rendered image and the captured image, M^{R} and M^{C} are the rendered foreground mask and the captured foreground mask, and E_{lap} is the laplacian geometry loss similar to that defined in Eqn. 1 and Eqn. 2. E_{softvisi} is a soft visibility loss, similar to [Liu et al. 2019a], which is specifically designed to handle the depth reasoning between body and clothing so that the gradient can be back-propagated through if the depth order is wrong. In detail, we define the soft visibility for a specific pixel as

$$S = \sigma\left(\frac{D^{\mathrm{C}} - D^{\mathrm{B}}}{c}\right), \tag{4}$$

where $\sigma(\cdot)$ is the sigmoid function, D^{C} and D^{B} are the depth rendered from the current viewpoint for the clothing and body layer, and c is a scaling constant. Then the soft visibility loss is defined as:

$$E_{\text{softvisi}} = S^2, \tag{5}$$

when $S > 0.5$ and also the current pixel is assigned to be clothing according to the 2D cloth segmentation. Otherwise, E_{softvisi} is set to be zero. If the pixel is labeled as clothing but the body layer comes out of the clothing layer in this viewpoint, the soft visiblity loss will back-propagate the information to update the surfaces towards the right depth order.

Following [Bagautdinov et al. 2021], in this inverse rendering stage, we also use a shadow network that computes quasi-shadow maps for body and clothing given the ambient occlusion maps. Different from [Bagautdinov et al. 2021] where the ambient occlusion is approximated with body template after LBS tansformation, we compute exact ambient occlusion using the output geometry from the body and clothing decoders since clothing deformation does not following the LBS transformation. The quasi-shadow map is then multiplied with the view-dependent texture before applying the differentiable renderer.

5.4 Texture Alignment with Inverse Rendering

The inverse rendering method mentioned in Sec. 5.3 already has the capability to improve photometric correspondences to some extent, since the network tends to predict texture with less variance across frames, along with deformed geometry to align the rendering output with the ground truth images. Ideally we only need to train the two decoders simultaneously with the inverse rendering loss to correct the correspondences while creating the generative model we need for driving the animation. However, we find that doing this alone would not correct all the correspondence error. Two reasons might prevent the decoder from obtaining a better minimum. First, the variation in photometric correspondences in our initial registration may be too large for the network to fix. Secondly, our VAE model with view conditioning may allow the decoder to cheat with the view-dependent texture rather than pushing the geometry.

These reasons motivate us to propose a new scheme to use inverse rendering for correspondence improvement. First, we separate the registered meshes into chunks of 50 neighboring frames. Then, we select the first chunk as the anchor frames, and train an anchor network for this chunk using inverse rendering model described in Sec. 5.3. After convergence, we use the trained network parameters to initialize the training of other chunks. To make sure that the alignment of the other chunks does not drift from the anchor frames, we set a small learning rate (1e-4 for the AdamW optimizer), and mix the anchor frames with each other chunk during training. Compared with the network used in Sec. 5.3, we remove the view conditioning in the texture branch of our decoder, and use a single texture prediction for inverse rendering in all the camera views. The output geometry predicted by the network of each chunk after training has more consistent correspondences across frames compared with the input, which is manifested by the consistent projected texture pattern in the UV space shown in Fig. 7.

Notice that in the whole pipeline of this paper, the method described here is immediately applied after the two-layer registration is obtained in Section 4. For each frame, we use the output geometry predicted by the network of its chunk as a new registered mesh with the improved correspondences. We use these data to officially train the body and the clothing networks, which are described in Section 5.1-5.3.

6 TEMPORAL MODELING FOR POSE-DRIVEN CLOTHING ANIMATION

In our two-layer codec avatars, the body output is conditioned on a single frame of skeletal pose and facial keypoints, while the clothing state is determined by the latent code. In order to animate the clothing from the driving signal, we use a Temporal Convolution Network (TCN) to learn the correlation between body dynamics and clothing deformation. Motivated by the work of cloth simulation, our TCN takes in the sequence of history and current skeletal pose and infers the latent clothing state. For cloth simulation, the clothing states (vertex position and velocity) in the previous frame is also

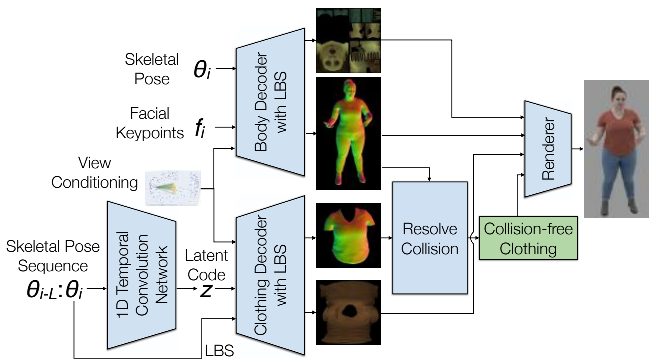

Fig. 5. Our clothed body animation pipeline.

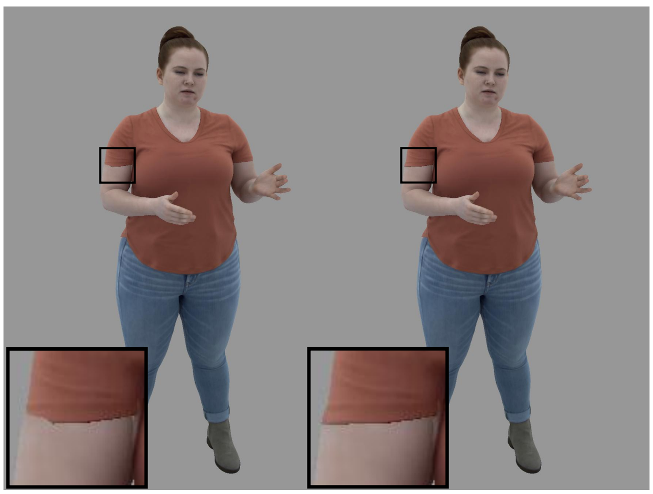

Fig. 6. An example of collision resolution. The collision area is highlighted in the zoomed-in boxes.

needed to calculate the current clothing state. In our data-driven formulation, we find that an auto-regressive model that takes in history clothing states is hard to train and does not outperform the non-autoregressive model given the limited amount of data (25 min). Therefore, the input to our TCN to is a temporal window of skeletal poses, not including the history clothing states.

An illustration of our animation pipeline is shown in Fig. 5. The temporal convolution network takes as input the joint angles in a window of L previous and current frames, and passes through several 1D temporal convolution layers to predict the clothing latent code for the current frame z. To train the TCN, we minimize the following loss function:

$$E_{\text{train}}^{TCN} = \|\mathrm{z} - \mathrm{z}^{\mathrm{c}}\|^2, \tag{6}$$

where z^{c} is the ground truth latent code obtained from the trained clothing VAE.

Resolving Collision. One solution is to add a training loss for TCN to make sure that the predicted clothing does not collide with the body. However, even without a loss to penalize collision, the clothing states predicted by our TCN model already well match the body shape, resulting in only minimal collisions. Thus we only need to resolve collision as a post processing step. In practice, we project the colliding clothing back onto the body surface with an additional margin in the body normal direction. This operation will solve most of collisions and make sure that the clothing and body are in the right depth order for rendering. One example of collision solving results can be seen in Fig. 6.

7 RESULTS

In this section, we first introduce our capture system and captured data. Then we show the results of our photometric texture alignment method to demonstrate its effectiveness in achieving better photometric correspondence in the UV space. After that, we show the animation output of our two-layer codec avatars with explicit clothing modeling. In particular, we demonstrate the advantage of our two-layer formulation against the single-layer model in previous work. This is followed by a demonstration of clothing texture editing in the animation output.

7.1 Data Capture

The training data for our codec avatars are captured by a multi-view capture system consisting of around 140 cameras that are distributed uniformly on a half dome above the ground. All the cameras run with hardware synchronization, capturing at the resolution of 4096×2668 and 30 fps. Three identities, one female (*Subject 1*) and two males (*Subject 2* and *Subject 3*), are captured with a pre-defined acting script. The script is designed to capture peak poses with the activation going through all body joints, followed by a 10-minute conversation to capture social behavior. For each subject, we collect sequences of 40k-50k frames in total and intentionally leave out approximately a chunk of 4-5k frames for testing.

7.2 Texture Alignment with Inverse Rendering

In this section, we show the results of texture alignment based on inverse rendering (Section 5.4) on the sequence of *Subject 2*. Textures are projected from the raw captured images to the registered meshes before and after the texture alignment procedure, and then unwrapped into the UV space for comparison. Example results for several frames are shown in the first and third column of Fig. 7. To assess the quality of alignment, we compare the mean UV texture of the anchor frames with the unwrapped texture of each individual frame. The error map is then visualized by the Jet colormap, shown in the second and fourth column of Fig. 7 respectively.

The visible pattern in the heatmap before texture alignment (the second column) verifies the lack of accurate interior correspondences in the registered clothing meshes from the ICP algorithm (Section 4). After the texture alignment (the fourth column), the error between UV texture of those frames and the mean of anchor frames is significantly reduced. This suggests that the correspondences in the mesh interior are improved in the inverse rendering process, and demonstrates the effectiveness of our texture alignment method.

To statistically evaluate the quality of photometric correspondence in the UV space, we compute the mean and standard deviation

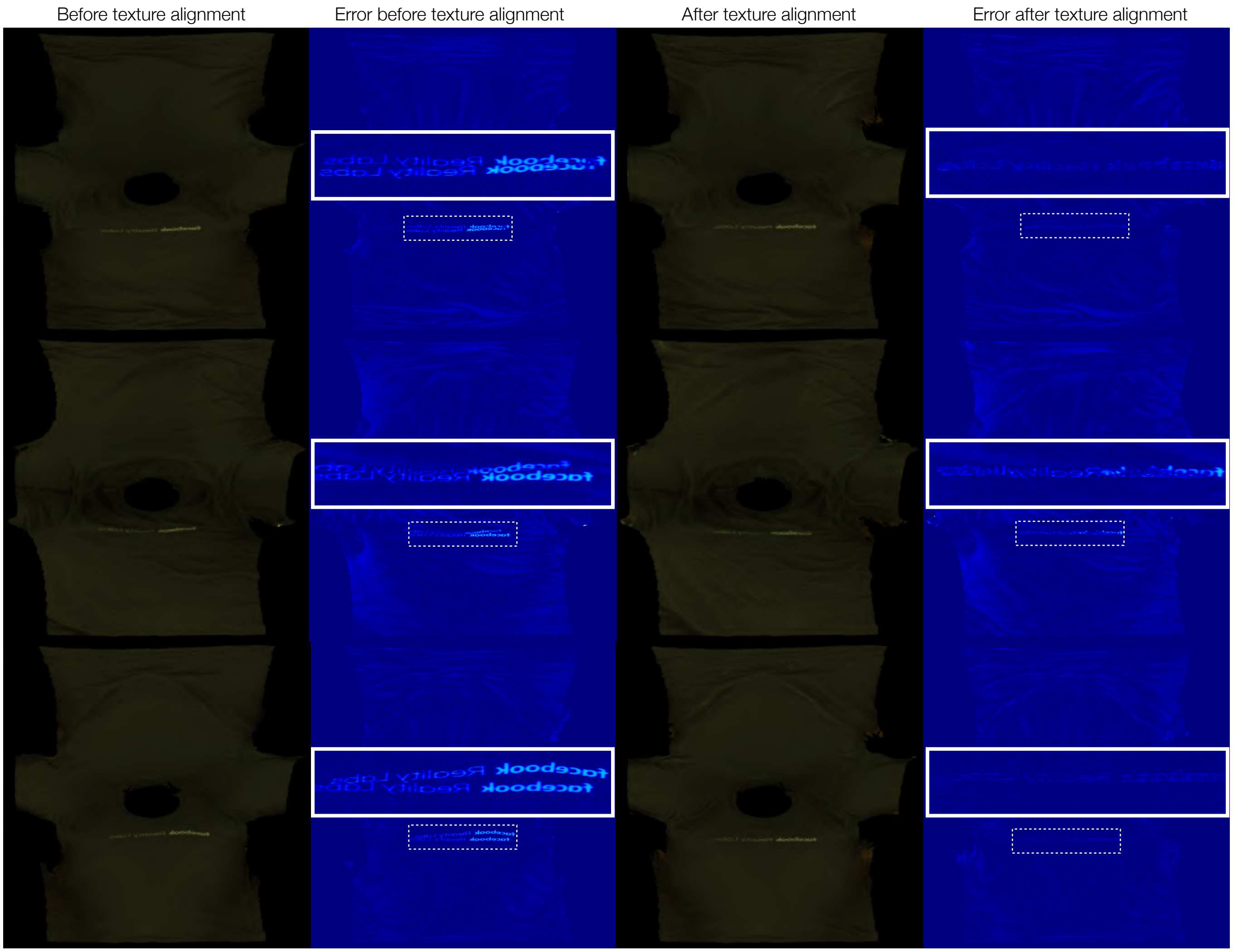

Fig. 7. Inverse-rendering-based texture alignment results. From left to right, we show (1) projected texture on the clothing mesh before texture alignment, (2) error map between the first column and the mean texture of anchor frames, (3) projected clothing texture after texture alignment, and (4) the difference between the third column and the mean texture of anchor frames. The error maps are visualized with the Jet colormap; lighter color represents larger error. We also show a zoom-in of the text region to highlight the difference.

of the unwrapped texture across different frames, as visualized in Fig. 8. Comparing the mean texture images, we observe a much sharper text pattern after texture alignment than before. Similarly, the standard deviation after texture alignment becomes smaller and more concentrated in the spatial domain. This result also verifies the improvement of photometric correspondence thanks to our proposed texture alignment approach.

7.3 Pose-Driven Animation

In this section, we present animation results produced by our two-layer codec avatars driven by the 3D skeletal pose and facial keypoints. In our animation pipeline, the body decoder is directly fed with skeletal pose and facial keypoints of the current frame; on the other hand, the clothing decoder is driven by latent clothing code generated by the temporal clothing model in Section 6, which takes a temporal window of history and current poses as input. We compare the quality of our animation with previous work [Bagautdinov et al. 2021] that uses a single-layer codec avatar. We follow the method described in the original paper [Bagautdinov et al. 2021] to animate the single-layer codec avatar: we randomly sample a noise from the unit Gaussian distribution, which is used for imputation of the latent code. The sampled latent code, the skeletal pose and facial keypoints are fed together into decoder network. We present qualitative animation results on all three testing sequences, shown in Fig. 9, 10, and 11 respectively. Our animation results are better seen in the supplementary video.

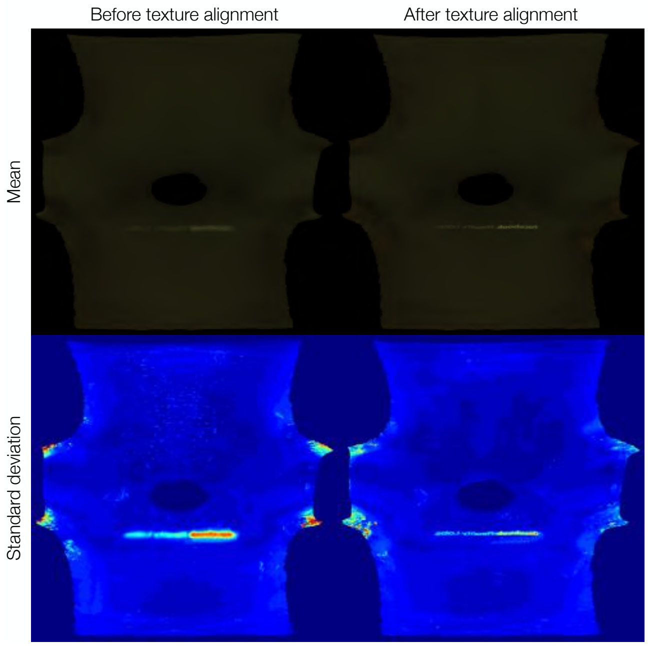

Fig. 8. Mean (top row) and standard deviation (bottom row, converted to jet colormap) of unwrapped texture before (left column) and after (right column) texture alignment on the sequence of *Subject 2*.

Sequence	[Bagautdinov et al. 2021]		Ours	
	MSE↓	SSIM↑	MSE↓	SSIM↑
Subject 1	100.57	0.8720	**74.73**	**0.8816**
Subject 2	81.95	0.8804	**58.14**	**0.8917**
Subject 3	456.20	0.8159	**356.52**	**0.8230**

Table 1. Quantitative comparison between our proposed method and the previous work. We report Mean Square Error (lower better) and the Structural Similarity Index Measure (higher better) on all three testing sequences.

Noticeably, our two-layer formulation helps remove the severe artifacts in the clothing regions in the animation output of [Bagautdinov et al. 2021], especially around the clothing boundary of Fig. 9, and 11. Indeed, as the body and clothing are modeled together, the single-layer avatars rely on the latent code to describe the many possible clothing states corresponding to the same body pose. During animation, however, the absence of a ground truth latent code leads to degradation of the output, despite the efforts in the [Bagautdinov et al. 2021] to disentangle the latent space from the driving signal. By contrast, our animation model achieves better animation quality by separating body and clothing into different modules: the body decoder can well determine the body states given the driving signal of the current frame; the temporal model learns to infer the most plausible clothing states from body dynamics for a longer period; the clothing VAE ensures a reasonable clothing output given its learned smooth latent manifold. In addition, our two-layer avatars show results with a sharper clothing boundary and clearer wrinkle patterns in these qualitative images.

Fig. 9. Comparison of animation output between our proposed method and baseline [Bagautdinov et al. 2021] on the *Subject 1* sequence.

We also quantitatively compare the animation output of our two-layer codec avatars with the baseline method [Bagautdinov et al. 2021] by evaluating the output images against the captured ground truth images. We report the evaluation metrics of Mean Square Error (MSE) and Structural Similarity Index Measure (SSIM) over the foreground pixels. The results are shown in Tab. 1. Our method consistently outperforms [Bagautdinov et al. 2021] on all three sequences and both evaluation metrics. In particular, it is worth

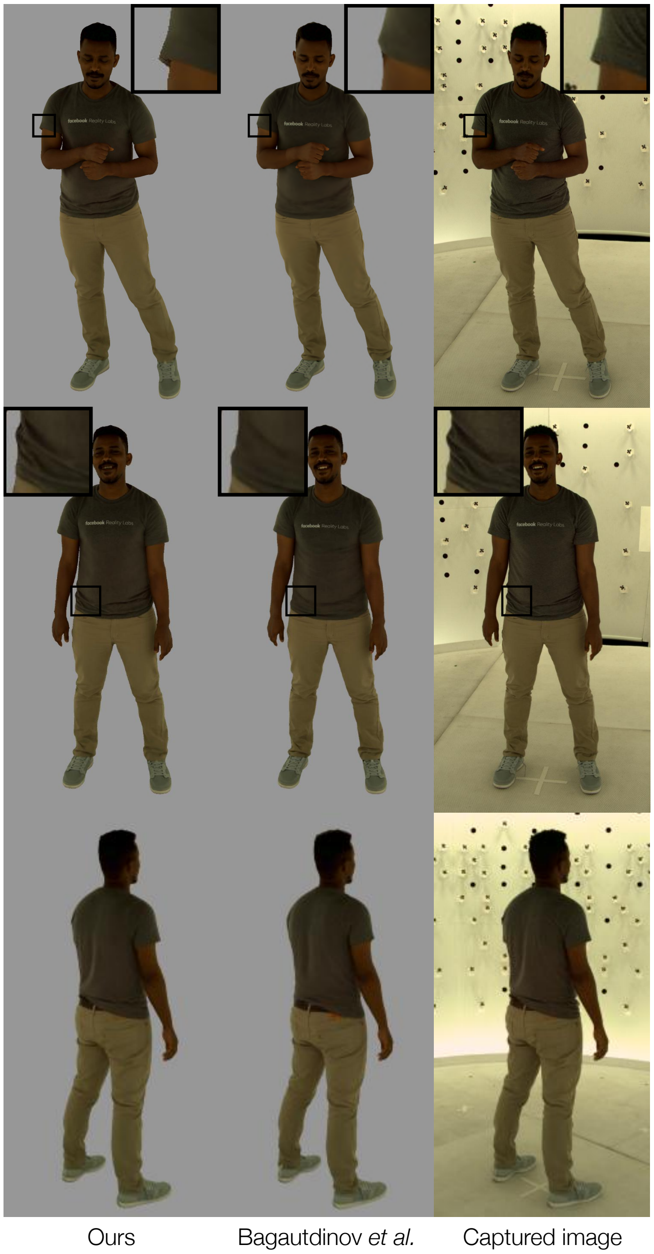

Fig. 10. Comparison of animation output between our proposed method and baseline [Bagautdinov et al. 2021] on the *Subject 2* sequence.

Fig. 11. Comparison of animation output between our proposed method and baseline [Bagautdinov et al. 2021] on the *Subject 3* sequence.

noting that our advantage on MSE is most obvious on the sequence of *Subject 3*, who is wearing a loose T-shirt that is hard for the single-layer avatar to model. This agrees with our observation in the qualitative images as well.

7.4 Application: Clothing Texture Editing

In this section, we demonstrate the application of clothing texture editing. On top of our photorealistic animation output, we further edit the clothing pattern in four different styles. First, we multiply the RGB channels of the clothing UV texture with different scaling factor to modify the color of the clothing. Second, we apply a checkerboard pattern on our clothing layer. Third, we ask an artist to create a stylistic pattern and then apply it to our clothing animation output. Fourth, we add the ACM SIGGRAPH Logo and text to the front side of the clothing. The results are shown in Fig. 12. Notice that once the desired pattern is determined, our model can produce

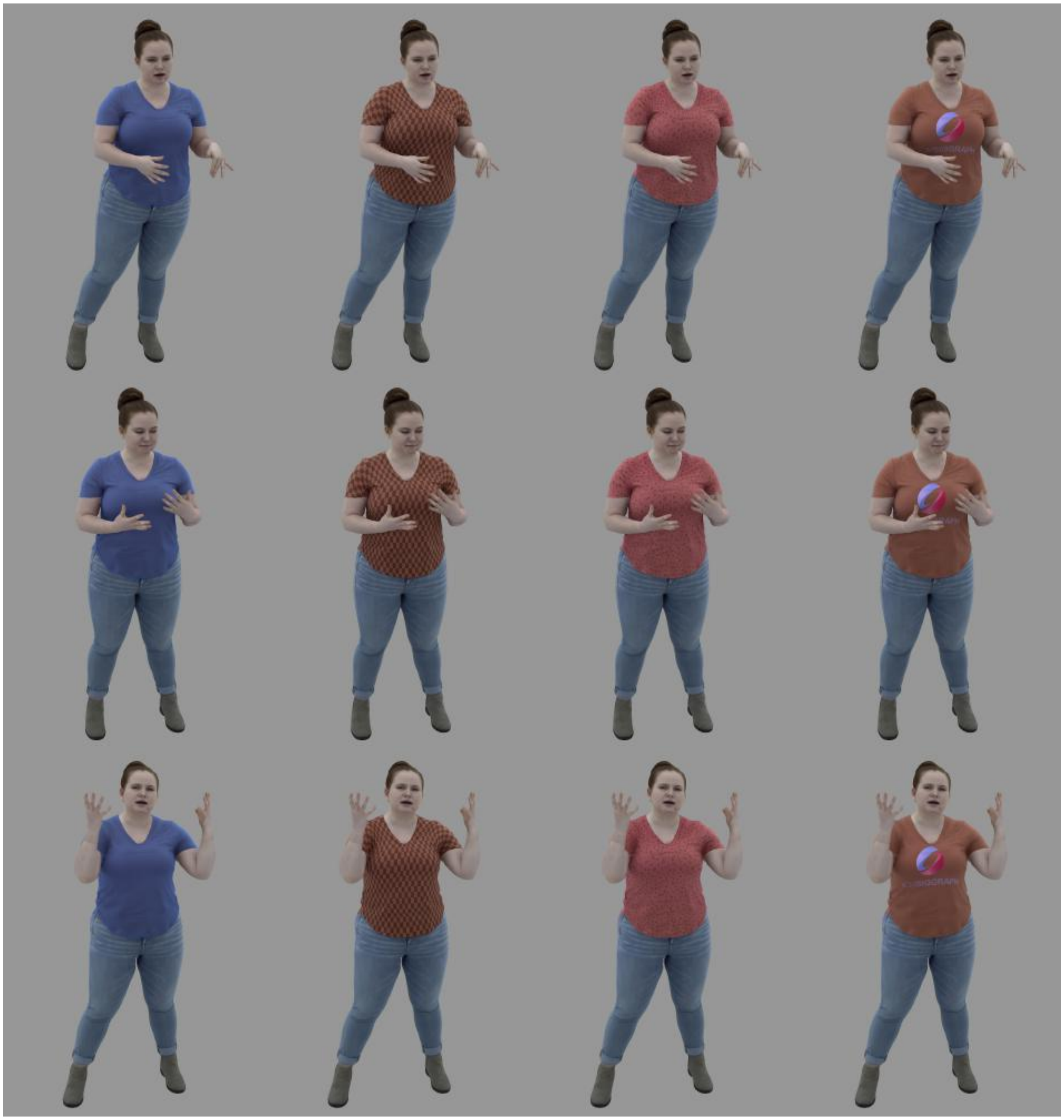

Fig. 12. Texture editing results of our two-layer codec avatars. From left to right, we show application of color transformation, checkerboard pattern, random artist-created pattern, and an ACM SIGGRAPH logo, respectively, for three different frames.

animation with the edited texture given any motion sequence along with the facial keypoints as input.

Compared with the single-layer model, our two-layer structure naturally allows us to easily manipulate the clothing texture in the UV space without interfering with the inner layer in a temporally coherent manner. For comparison, we apply the same blue color transformation to the single-layer output. For this purpose, we manually segment out the clothing region for the first frame in the sequence in the UV space, and apply the color transformation in the segmented region to all the following frames. This produces reasonable results for the first frame (shown on the first column of Fig. 13); for the following frames, however, applying the color transformation in the same UV region will suffer from misalignment of

Fig. 13. Comparison between the single-layer model (bottom row) and the two-layer model (top row) on texture editing in three different frames. In particular, the first column shows the frame where we manually segment out the upper clothing region in the UV space for the single-layer model.

the edited area and actual clothing region, as shown in the right two columns of Fig. 13. The visual artifact caused by this misalignment is highlighted in the zoomed-in boxes in the figure.

8 DISCUSSION

We have proposed a two-layer mesh representation for building an animatable avatar for clothed body. Results have demonstrated that the explicit clothing modeling not only improves the rendered clothing quality in animation, but also enables the editability of the clothing texture, opening up new possibilities with codec avatars. The two-layer avatar models cannot be obtained without the success of two-layer registration of the clothed body. We thus have presented a new clothed body registration method along with a texture alignment method to improve the photometric correspondences using inverse rendering.

Our clothed body model is trained person-specifically and also can only be animated person-specifically. All the driving signals have been captured from the same identity performing some normal but different social interactions. The animatable model may not be able to generalize to some poses deviating largely from the training pose distribution. In another word, major artifacts may appear if our model is used for arbitrary motion retargeting. Besides, in this work we are only focusing on T-shirts. To extend the work to lower body clothing, like short pants with the boundary shifting on the legs, we need to extend the current two-layer work to handle multiple layers, potentially with occlusion between layers, which poses additional challenges to both registration and modeling. Another common lower-body clothing is skirt, which could be even more difficult due to its large motion and deformation. We cannot handle topology-changing clothing, like opening a zipped jacket.

Even with the current two-layer framework, our clothing registration method would fail if heavy interaction happens between hands and clothing, for example, hands dragging the clothing or hands put inside the clothing. The current non-physical interaction modeling between clothing and body may not easily extend to handle these challenges. One possibility is to integrate more physics constraint into registration and learning for animation. Another limitation we want to address is to handle degraded animation signals. Right now, if the input pose quality drops, the animation quality would also decrease. One potential solution is to pre-filter the animation signal or project to a learned pose prior.

REFERENCES

Kfir Aberman, Mingyi Shi, Jing Liao, Dani Lischinski, Baoquan Chen, and Daniel Cohen-Or. 2019. Deep video-based performance cloning. In *Computer Graphics Forum*, Vol. 38. Wiley Online Library, 219–233.

Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. 2005. SCAPE: Shape Completion and Animation of People. *ACM Trans. Graph.* 24, 3 (July 2005), 408–416. https://doi.org/10.1145/1073204.1073207

Timur Bagautdinov, Chenglei Wu, Tomas Simon, Takaaki Shiratori, Shih-En Wei, Fabian Prada, Weipeng Xu, Yaser Sheikh, and Jason Saraghi. 2021. Driving-Signal Aware Full-Body Avatars. *ACM Trans. on Graphics (Proc. SIGGRAPH)* (July 2021).

David Baraff and Andrew Witkin. 1998. Large steps in cloth simulation. In *Proceedings of the 25th annual conference on Computer graphics and interactive techniques*. 43–54.

Hugo Bertiche, Meysam Madadi, and Sergio Escalera. 2020a. CLOTH3D: Clothed 3D Humans. In *Computer Vision – ECCV 2020*, Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (Eds.). Springer International Publishing, Cham, 344–359.

Hugo Bertiche, Meysam Madadi, and Sergio Escalera. 2020b. PBNS: Physically Based Neural Simulator for Unsupervised Garment Pose Space Deformation. *arXiv preprint arXiv:2012.11310* (2020).

Federica Bogo, Michael J. Black, Matthew Loper, and Javier Romero. 2015. Detailed Full-Body Reconstructions of Moving People from Monocular RGB-D Sequences. In *International Conference on Computer Vision (ICCV)*. 2300–2308.

Derek Bradley, Tiberiu Popa, Alla Sheffer, Wolfgang Heidrich, and Tamy Boubekeur. 2008. Markerless garment capture. In *ACM Transactions on Graphics (TOG)*, Vol. 27. ACM, 99.

Matthieu Bray, Pushmeet Kohli, and Philip HS Torr. 2006. Posecut: Simultaneous segmentation and 3d pose estimation of humans using dynamic graph-cuts. In *European conference on computer vision*. Springer, 642–655.

Thomas Brox, Bodo Rosenhahn, Juergen Gall, and Daniel Cremers. 2010. Combined region and motion-based 3D tracking of rigid and articulated objects. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32, 3 (2010), 402–415.

Joel Carranza, Christian Theobalt, Marcus A. Magnor, and Hans-Peter Seidel. 2003. Free-viewpoint Video of Human Actors. *ACM Trans. Graph.* 22, 3 (July 2003).

Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A Efros. 2019. Everybody dance now. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 5933–5942.

Xiaowu Chen, Bin Zhou, Feixiang Lu, Lin Wang, Lang Bi, and Ping Tan. 2015. Garment Modeling with a Depth Camera. *ACM Trans. Graph.* 34, 6, Article 203 (Oct. 2015), 12 pages. https://doi.org/10.1145/2816795.2818059

Nuttapong Chentanez, Miles Macklin, Matthias Müller, Stefan Jeschke, and Tae-Yong Kim. 2020. *Cloth and Skin Deformation with a Triangle Mesh Based Convolutional Neural Network*. Eurographics Association, Goslar, DEU. https://doi.org/10.1111/cgf.14107

Edilson De Aguiar, Carsten Stoll, Christian Theobalt, Naveed Ahmed, Hans-Peter Seidel, and Sebastian Thrun. 2008. Performance capture from sparse multi-view video. In *ACM Transactions on Graphics (TOG)*, Vol. 27. ACM, 98.

Patrick Esser, Johannes Haux, Timo Milbich, et al. 2018. Towards learning a realistic rendering of human behavior. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*. 0–0.

Juergen Gall, Carsten Stoll, Edilson De Aguiar, Christian Theobalt, Bodo Rosenhahn, and Hans-Peter Seidel. 2009. Motion capture using joint skeleton tracking and surface estimation. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 1746–1753.

Silvano Galliani, Katrin Lasinger, and Konrad Schindler. 2015. Massively parallel multiview stereopsis by surface normal diffusion. In *Proceedings of the IEEE International Conference on Computer Vision*. 873–881.

Russell Gillette, Craig Peters, Nicholas Vining, Essex Edwards, and Alla Sheffer. 2015. Real-Time Dynamic Wrinkling of Coarse Animated Cloth. In *Proceedings of the 14th ACM SIGGRAPH / Eurographics Symposium on Computer Animation* (Los Angeles, California) *(SCA '15)*. Association for Computing Machinery, New York, NY, USA, 17–26. https://doi.org/10.1145/2786784.2786789

Rony Goldenthal, David Harmon, Raanan Fattal, Michel Bercovier, and Eitan Grinspun. 2007. Efficient Simulation of Inextensible Cloth. *ACM Trans. Graph.* 26, 3 (July 2007), 49–es. https://doi.org/10.1145/1276377.1276438

Erhan Gundogdu, Victor Constantin, Amrollah Seifoddini, Minh Dang, Mathieu Salzmann, and Pascal Fua. 2019. GarNet: A Two-Stream Network for Fast and Accurate 3D Cloth Draping. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Kaiwen Guo, Feng Xu, Yangang Wang, Yebin Liu, and Qionghai Dai. 2015. Robust Non-rigid Motion Tracking and Surface Reconstruction Using L0 Regularization. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV) (ICCV '15)*. 3083–3091.

Marc Habermann, Weipeng Xu, Michael Zollhöfer, Gerard Pons-Moll, and Christian Theobalt. 2019. LiveCap: Real-Time Human Performance Capture From Monocular Video. *ACM Trans. Graph.* 38, 2, Article 14 (March 2019), 17 pages. https://doi.org/10.1145/3311970

Thomas Helten, Meinard Muller, Hans-Peter Seidel, and Christian Theobalt. 2013. Real-Time Body Tracking with One Depth Camera and Inertial Sensors. In *The IEEE International Conference on Computer Vision (ICCV)*.

Yinghao Huang, Federica Bogo, Christoph Lassner, Angjoo Kanazawa, Peter V. Gehler, Javier Romero, Ijaz Akhter, and Michael J. Black. 2017. Towards Accurate Markerless Human Shape and Pose Estimation over Time. In *International Conference on 3D Vision (3DV)*.

Alec Jacobson, Elif Tosun, Olga Sorkine, and Denis Zorin. 2010. Mixed finite elements for variational surface modeling. In *Computer graphics forum*, Vol. 29. Wiley Online Library, 1565–1574.

Ning Jin, Yilin Zhu, Zhenglin Geng, and Ronald Fedkiw. 2020. A Pixel-Based Framework for Data-Driven Clothing. In *Computer Graphics Forum*, Vol. 39. Wiley Online Library, 135–144.

Ladislav Kavan, Steven Collins, Jiri Zara, and Carol O'Sullivan. 2008. Geometric Skinning with Approximate Dual Quaternion Blending. *ACM Trans. Graph.* 27, 4 (2008), 105.

Ladislav Kavan, Dan Gerszewski, Adam W. Bargteil, and Peter-Pike Sloan. 2011. Physics-Inspired Upsampling for Cloth Simulation in Games. *ACM Trans. Graph.* 30, 4, Article 93 (July 2011), 10 pages. https://doi.org/10.1145/2010324.1964988

Ladislav Kavan and Jiri Zara. 2005. Spherical Blend Skinning: A Real-time Deformation of Articulated Models. In *2005 ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games*. ACM Press, 9–16.

Doyub Kim, Woojong Koh, Rahul Narain, Kayvon Fatahalian, Adrien Treuille, and James F. O'Brien. 2013. Near-Exhaustive Precomputation of Secondary Cloth Effects. *ACM Trans. Graph.* 32, 4, Article 87 (July 2013), 8 pages. https://doi.org/10.1145/2461912.2462020

Zorah Lahner, Daniel Cremers, and Tony Tung. 2018. Deepwrinkles: Accurate and realistic clothing modeling. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 667–684.

John P Lewis, Matt Cordner, and Nickson Fong. 2000. Pose space deformation: a unified approach to shape interpolation and skeleton-driven deformation. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*. 165–172.

Hao Li, Bart Adams, Leonidas J. Guibas, and Mark Pauly. 2009. Robust single-view geometry and motion reconstruction. https://doi.org/10.1145/1618452.1618521

Lingjie Liu, Weipeng Xu, Michael Zollhöfer, Hyeongwoo Kim, Florian Bernard, Marc Habermann, Wenping Wang, and Christian Theobalt. 2019c. Neural Rendering and Reenactment of Human Actor Videos. *ACM Trans. Graph.* 38, 5, Article 139 (Oct. 2019), 14 pages. https://doi.org/10.1145/3333002

Shichen Liu, Tianye Li, Weikai Chen, and Hao Li. 2019a. Soft Rasterizer: A Differentiable Renderer for Image-Based 3D Reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Wen Liu, Zhixin Piao, Jie Min, Wenhan Luo, Lin Ma, and Shenghua Gao. 2019b. Liquid warping gan: A unified framework for human motion imitation, appearance transfer and novel view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 5904–5913.

Yebin Liu, Carsten Stoll, Juergen Gall, Hans-Peter Seidel, and Christian Theobalt. 2011. Markerless motion capture of interacting characters using multi-view image segmentation. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 1249–1256.

Stephen Lombardi, Jason Saragih, Tomas Simon, and Yaser Sheikh. 2018. Deep appearance models for face rendering. *ACM Transactions on Graphics (TOG)* 37, 4 (2018), 1–13.

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. 2015. SMPL: A skinned multi-person linear model. *ACM transactions on graphics (TOG)* 34, 6 (2015), 1–16.

Qianli Ma, Jinlong Yang, Anurag Ranjan, Sergi Pujades, Gerard Pons-Moll, Siyu Tang, and Michael J. Black. 2020. Learning to Dress 3D People in Generative Clothing. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 6468–6477.

Wojciech Matusik, Chris Buehler, Ramesh Raskar, Steven J Gortler, and Leonard McMillan. 2000. Image-based visual hulls. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*. ACM Press/Addison-Wesley Publishing Co., 369–374.

Armin Mustafa, Hansung Kim, Jean-Yves Guillemaut, and Adrian Hilton. 2015. General Dynamic Scene Reconstruction from Multiple View Video. In *ICCV*.

Rahul Narain, Armin Samii, and James F O'brien. 2012. Adaptive anisotropic remeshing for cloth simulation. *ACM transactions on graphics (TOG)* 31, 6 (2012), 1–10.

Ahmed A A Osman, Timo Bolkart, and Michael J. Black. 2020. STAR: A Sparse Trained Articulated Human Body Regressor. In *European Conference on Computer Vision (ECCV)*. https://star.is.tue.mpg.de

Keunhong Park, Utkarsh Sinha, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Steven M Seitz, and Ricardo-Martin Brualla. 2020. Deformable Neural Radiance Fields. *arXiv preprint arXiv:2011.12948* (2020).

Chaitanya Patel, Zhouyingcheng Liao, and Gerard Pons-Moll. 2020. TailorNet: Predicting Clothing in 3D as a Function of Human Pose, Shape and Garment Style. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE.

Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. 2020. Neural Body: Implicit Neural Representations with Structured Latent Codes for Novel View Synthesis of Dynamic Humans. *arXiv preprint arXiv:2012.15838* (2020).

Gerard Pons-Moll, Sergi Pujades, Sonny Hu, and Michael J Black. 2017. ClothCap: Seamless 4D clothing capture and retargeting. *ACM Transactions on Graphics (TOG)* 36, 4 (2017), 1–15.

Sergey Prokudin, Michael J Black, and Javier Romero. 2021. SMPLpix: Neural Avatars from 3D Human Models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 1810–1819.

Albert Pumarola, Antonio Agudo, Alberto Sanfeliu, and Francesc Moreno-Noguer. 2018. Unsupervised person image synthesis in arbitrary poses. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 8620–8628.

Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. 2020. D-NeRF: Neural Radiance Fields for Dynamic Scenes. *arXiv preprint arXiv:2011.13961* (2020).

Amit Raj, Julian Tanke, James Hays, Minh Vo, Carsten Stoll, and Christoph Lassner. 2020. ANR: Articulated Neural Rendering for Virtual Avatars. *arXiv preprint arXiv:2012.12890* (2020).

Javier Romero, Dimitrios Tzionas, and Michael J. Black. 2017. Embodied Hands: Modeling and Capturing Hands and Bodies Together. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)* 36, 6 (Nov. 2017).

Igor Santesteban, Miguel A. Otaduy, and Dan Casas. 2019. Learning-Based Animation of Clothing for Virtual Try-On. *Computer Graphics Forum (Proc. Eurographics)* 38, 2 (2019), 355–366. https://doi.org/10.1111/cgf.13643

Kripasindhu Sarkar, Dushyant Mehta, Weipeng Xu, Vladislav Golyanik, and Christian Theobalt. 2020. Neural Re-Rendering of Humans from a Single Image. In *European Conference on Computer Vision (ECCV)*.

Chenyang Si, Wei Wang, Liang Wang, and Tieniu Tan. 2018. Multistage adversarial losses for pose-based human image synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 118–126.

Jonathan Starck and Adrian Hilton. 2007. Surface capture for performance-based animation. *IEEE Computer Graphics and Applications* 27, 3 (2007), 21–31.

Carsten Stoll, Juergen Gall, Edilson de Aguiar, Sebastian Thrun, and Christian Theobalt. 2010. Video-Based Reconstruction of Animatable Human Characters. In *ACM SIGGRAPH Asia 2010 Papers* (Seoul, South Korea) *(SIGGRAPH ASIA '10)*. Association for Computing Machinery, New York, NY, USA, Article 139, 10 pages. https://doi.org/10.1145/1866158.1866161

Raquel Vidaurre, Igor Santesteban, Elena Garces, and Dan Casas. 2020. Fully Convolutional Graph Neural Networks for Parametric Virtual Try-On. *Computer Graphics Forum (Proc. SCA)* (2020).

Daniel Vlasic, Ilya Baran, Wojciech Matusik, and Jovan Popović. 2008. Articulated mesh animation from multi-view silhouettes. In *ACM Transactions on Graphics (TOG)*, Vol. 27. ACM, 97.

Aaron Walsman, Weilin Wan, Tanner Schmidt, and Dieter Fox. 2017. Dynamic High Resolution Deformable Articulated Tracking. (11 2017).

Huamin Wang, Florian Hecht, Ravi Ramamoorthi, and James F. O'Brien. 2010. Example-Based Wrinkle Synthesis for Clothing Animation. *ACM Trans. Graph.* 29, 4, Article 107 (July 2010), 8 pages. https://doi.org/10.1145/1778765.1778844

Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. 2018. Video-to-video synthesis. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. 1152–1164.

Ziyan Wang, Timur Bagautdinov, Stephen Lombardi, Tomas Simon, Jason Saragih, Jessica Hodgins, and Michael Zollhöfer. 2020. Learning Compositional Radiance Fields of Dynamic Human Heads. *arXiv preprint arXiv:2012.09955* (2020).

Michael Waschbüsch, Stephan Würmlin, Daniel Cotting, Filip Sadlo, and Markus Gross. 2005. Scalable 3D video of dynamic scenes. *The Visual Computer* 21, 8-10 (2005), 629–638.

Chenglei Wu, Carsten Stoll, Levi Valgaerts, and Christian Theobalt. 2013. On-set Performance Capture of Multiple Actors With A Stereo Camera. In *ACM Transactions on Graphics (Proceedings of SIGGRAPH Asia 2013)*, Vol. 32. 161:1–161:11. https://doi.org/10.1145/2508363.2508418

Chenglei Wu, Kiran Varanasi, and Christian Theobalt. 2012. Full body performance capture under uncontrolled and varying illumination: A shading-based approach. In *ECCV*. 757–770.

Donglai Xiang, Fabian Prada, Chenglei Wu, and Jessica K. Hodgins. 2020. MonoClothCap: Towards Temporally Coherent Clothing Capture from Monocular RGB Video. In *3DV*. IEEE, 322–332.

Weipeng Xu, Avishek Chatterjee, Michael Zollhöfer, Helge Rhodin, Dushyant Mehta, Hans-Peter Seidel, and Christian Theobalt. 2018. MonoPerfCap: Human Performance Capture From Monocular Video. *ACM Trans. Graph.* 37, 2, Article 27 (May 2018), 15 pages. https://doi.org/10.1145/3181973

Tao Yu, Zerong Zheng, Yuan Zhong, Jianhui Zhao, Qionghai Dai, Gerard Pons-Moll, and Yebin Liu. 2019. SimulCap : Single-View Human Performance Capture With Cloth Simulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Chao Zhang, Sergi Pujades, Michael J Black, and Gerard Pons-Moll. 2017. Detailed, accurate, human shape estimation from clothed 3D scan sequences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4191–4200.

Meng Zhang, Tuanfeng Wang, Duygu Ceylan, and Niloy J Mitra. 2020. Deep Detail Enhancement for Any Garment. *arXiv preprint arXiv:2008.04367* (2020).

Qing Zhang, Bo Fu, Mao Ye, and Ruigang Yang. 2014. Quality Dynamic Human Body Modeling Using a Single Low-cost Depth Camera. In *CVPR*. IEEE, 676–683.

Bin Zhou, Xiaowu Chen, Qiang Fu, Kan Guo, and Ping Tan. 2013. Garment Modeling from a Single Image. *Computer Graphics Forum* 32, 7 (2013), 85–91. https://doi.org/10.1111/cgf.12215 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/cgf.12215