

Dressing Avatars: Deep Photorealistic Appearance for Physically Simulated Clothing

DONGLAI XIANG, Carnegie Mellon University, USA and Meta Reality Labs Research, USA

TIMUR BAGAUTDINOV, TUUR STUYCK, and FABIAN PRADA, Meta Reality Labs Research, USA

JAVIER ROMERO, WEIPENG XU, and SHUNSUKE SAITO, Meta Reality Labs Research, USA

JINGFAN GUO, University of Minnesota, USA

BREANNAN SMITH, TAKAAKI SHIRATORI, and YASER SHEIKH, Meta Reality Labs Research, USA

JESSICA HODGINS, Carnegie Mellon University, USA and FAIR, Meta AI, USA

CHENGLEI WU, Meta Reality Labs Research, USA



Fig. 1. We develop pose-driven full-body avatars with photorealistic clothing by applying neural rendering to physically simulated garments. On the left, we show a skirt animation together with the body avatar built from the same captured sequence. We further retarget the skirt to a novel sequence with the same actor and two new actors. On the right, we animate the skirt and a T-shirt together.

Despite recent progress in developing animatable full-body avatars, realistic modeling of clothing - one of the core aspects of human self-expression - remains an open challenge. State-of-the-art physical simulation methods can generate realistically behaving clothing geometry at interactive rates. Modeling photorealistic appearance, however, usually requires physically-based rendering which is too expensive for interactive applications. On the other hand, data-driven deep appearance models are capable of efficiently

producing realistic appearance, but struggle at synthesizing geometry of highly dynamic clothing and handling challenging body-clothing configurations. To this end, we introduce pose-driven avatars with explicit modeling of clothing that exhibit both photorealistic appearance learned from real-world data and realistic clothing dynamics. The key idea is to introduce a neural clothing appearance model that operates on top of explicit geometry: at training time we use high-fidelity tracking, whereas at animation time we rely on physically simulated geometry. Our core contribution is a physically-inspired appearance network, capable of generating photorealistic appearance with view-dependent and dynamic shadowing effects even for unseen body-clothing configurations. We conduct a thorough evaluation of our model and demonstrate diverse animation results on several subjects and different types of clothing. Unlike previous work on photorealistic full-body avatars, our approach can produce much richer dynamics and more realistic deformations even for many examples of loose clothing. We also demonstrate that our formulation naturally allows clothing to be used with avatars of different people while staying fully animatable, thus enabling, for the first time, photorealistic avatars with novel clothing.

CCS Concepts: • **Computing methodologies** → **Animation; Rendering; Computer vision; Image-based rendering.**

Additional Key Words and Phrases: photorealistic avatars, cloth simulation, neural rendering

Authors' addresses: Donglai Xiang, Carnegie Mellon University, USA and Meta Reality Labs Research, USA, donglaix@cs.cmu.edu; Timur Bagautdinov, timurb@fb.com; Tuur Stuyck, tuur@fb.com; Fabian Prada, fabianprada@fb.com, Meta Reality Labs Research, USA; Javier Romero, javierromero1@fb.com; Weipeng Xu, xuweipeng@fb.com; Shunsuke Saito, shunsukesaito@fb.com, Meta Reality Labs Research, USA; Jingfan Guo, guo00109@umn.edu, University of Minnesota, USA; Breannan Smith, breannan@fb.com; Takaaki Shiratori, tshiratori@fb.com; Yaser Sheikh, yasers@fb.com, Meta Reality Labs Research, USA; Jessica Hodgins, Carnegie Mellon University, USA and FAIR, Meta AI, USA, jkh@cmu.edu; Chenglei Wu, chenglei@fb.com, Meta Reality Labs Research, USA.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

© 2022 Copyright held by the owner/author(s).

0730-0301/2022/12-ART222

<https://doi.org/10.1145/3550454.3555456>

ACM Reference Format:

Donglai Xiang, Timur Bagautdinov, Tuur Stuyck, Fabian Prada, Javier Romero, Weipeng Xu, Shunsuke Saito, Jingfan Guo, Breannan Smith, Takaaki Shiratori, Yaser Sheikh, Jessica Hodgins, and Chenglei Wu. 2022. Dressing Avatars: Deep Photorealistic Appearance for Physically Simulated Clothing. *ACM Trans. Graph.* 41, 6, Article 222 (December 2022), 15 pages. <https://doi.org/10.1145/3550454.3555456>

1 INTRODUCTION

Photorealistic avatars are an essential component of authentic social telepresence, which is one of the key applications for AR/VR. Recent work on full-body avatars [Bagautdinov et al. 2021; Habermann et al. 2021b; Xiang et al. 2021] introduced animatable models capable of producing photorealistic synthetic representations of humans from sparse signals such as body pose. However, synthesizing realistic clothing still remains a challenge in avatar modeling. At the same time, clothing is an essential form of self-expression, creating a fundamental need for high-quality photorealistic animation of clothing.

Existing work on avatars with animatable clothing can be categorized into two main streams. Cloth simulation creates realistic clothing deformations with dynamics [Bouaziz et al. 2014; Bridson et al. 2002; Ly et al. 2020; Macklin et al. 2016; Müller et al. 2007], but only focuses on modeling geometry. The other line of the work leverages real-world captures to build neural representations of clothing geometry [Bertiche et al. 2021a] and may include appearance [Habermann et al. 2021b; Liu et al. 2021; Xiang et al. 2021]. However, these systems usually damp the clothing dynamics, struggle at generalizing to unseen poses and cannot handle collisions well. Our key insight is that these two lines of work are actually complementary to each other, and combining them can help achieve the best of both worlds.

In this work, we propose to integrate physics-based cloth simulation into avatar modeling, so that the clothing on the avatar can be animated photorealistically with the body, while achieving high-quality dynamics, collision handling and capabilities to animate and render avatars with novel clothing. Our work builds upon full-body Codec Avatars [Bagautdinov et al. 2021; Xiang et al. 2021], which leverage a Variational Autoencoder (VAE) to model the geometry and appearance of a human body. In particular, we follow the multi-layer formulation of [Xiang et al. 2021], but redesign the clothing layer to integrate a physically-based simulator. Namely, at the training stage, we learn the clothing appearance model using real-world data, by processing raw captures with our dynamic clothing registration pipeline. At test time, we simulate the clothing geometry on top of the underlying body model with appropriate material parameters, and then apply the learned appearance model to synthesize the final output.

Unfortunately, there are two major issues with a naive implementation of this pipeline. First, there exists a gap between the simulator output and the tracking obtained from the real data. Estimating the full set of physical parameters for body and clothing to faithfully reproduce the clothed body configuration remains an unsolved problem, despite some progress in controlled settings [Miguel et al. 2012] or in estimating only the body parameters [Guo et al. 2021]. There are inevitable differences between the test-time simulation output

with manually selected parameters and the real-world clothing geometry used for training. Second, tracking clothing and underlying body geometry at high accuracy is still a challenging problem, especially for loose clothing such as skirts and dresses. Both of these issues, inconsistency between training and test scenarios and unreliable tracking, make learning a generalizable appearance model more challenging. Thus, a good design of the appearance model should avoid learning chance correlations between degenerated tracked cloth geometry and specific appearance. To this end, we design the model to be localized in terms of both architecture (U-Net [Ronneberger et al. 2015]) and input representation (normals). We also take inspiration from physically-based rendering and decompose appearance into local diffuse components, view-dependent and global illumination effects such as shadowing. In particular, we rely on an unsupervised shadow network conditioned on the ambient occlusion map explicitly computed from the body and clothing geometry, so that the dynamic shadowing can be effectively modeled even for a different underlying body model at test time.

Our approach generates physically realistic dynamics and photorealistic appearance that are robust to diverse body motion with complex body-clothing interactions. In addition, our formulation allows the transfer of clothing between different individuals' body avatars as shown in Fig. 1, as well as virtual garment resizing in Fig. 9. Our method opens up the possibility to dress photorealistic avatars with novel garments. To summarize, our contributions are as follows:

- We present animatable clothed human avatars with data-driven photorealistic appearance and physically realistic clothing dynamics from simulation;
- We develop a deep clothing appearance model to produce photorealistic clothing appearance that bridges the generalization gap between the tracked clothing geometry for training and the simulated clothing geometry at test time;
- Our animation system further enables transferring clothing between different subjects as well as editing of garment size.

In our experiments, we evaluate the effectiveness of our approach by animating multiple identities and clothing types, and provide a comprehensive qualitative and quantitative comparison to existing techniques.

2 RELATED WORK

Our goal in this work is to build photorealistic animatable avatars, and we review work in this area in Sec. 2.1. To collect data for training such avatars we use techniques for multi-view clothing capture (Sec. 2.2). We leverage physics-based clothing animation (Sec. 2.3) so that the animation output not only looks photorealistic in appearance, but also natural in terms of physical dynamics.

2.1 Photorealistic Animatable Avatars

Models for dynamic full-body humans have been widely studied because of their applications in the movie and gaming industries, as well as virtual and augmented reality. Human avatars are usually built to be animatable with joint angles and skinning to allow easy control of skeleton-level deformation. The classical approach

for modeling full-body appearance is to use a fixed texture associated with the template mesh [Carranza et al. 2003; de Aguiar et al. 2008; Starck and Hilton 2007], which enables the rendering of the avatars by various graphics techniques such as direct shading or ray tracing. However, it is impossible for simple rendering to account for the complicated appearance change under the variation of viewpoint, deformation, material and lighting. Improved realism has been achieved in classic methods with techniques related to image-based warping [Chabert et al. 2006; Eisemann et al. 2008; Jain et al. 2010; Volino et al. 2014], but they do not enable body reposing. To achieve photorealistic rendering of animatable full-body human appearance using classical techniques requires heavy computation and the work of digital artists. In recent years there has been a lot of interest in building photorealistic human avatars directly from captured images in a data-driven manner. In these approaches, a deep neural network takes as input a specific body configuration (for example pose or mesh vertices), and outputs a predicted image supervised by the ground truth capture from the same viewpoint [Habermann et al. 2021a; Prokudin et al. 2021; Shysheya et al. 2019; Zhi et al. 2020]. Data-driven approaches based on image translation networks have also attacked this problem purely from a monocular 2D perspective, using rendered skeletons [Chan et al. 2019] or 2D body correspondences [Wang et al. 2018] as input. Some work further utilizes recent breakthroughs in neural rendering, such as Neural Radiance Fields [Kwon et al. 2021; Liu et al. 2021; Mildenhall et al. 2020; Noguchi et al. 2021; Peng et al. 2021a,b; Su et al. 2021; Xu et al. 2021; Zheng et al. 2022] and Deferred Neural Rendering [Grigorev et al. 2021; Raj et al. 2021; Thies et al. 2019]. Notably, full-body Codec Avatars [Bagautdinov et al. 2021; Remelli et al. 2022; Xiang et al. 2021] have been proposed as a promising approach to achieving photorealistic telepresence. Most of these approaches consider clothing as rigidly attached to the human body, and the results are limited to tight clothing such as T-shirts and pants. One exception is the work by Habermann and colleagues [2021a], which can model the deformation of loose clothing such as skirts and dresses in a coarse-to-fine manner, but the predicted clothing deformation is not physically realistic because the approach lacks explicit physics priors.

2.2 Clothing Capture

Clothing capture has been explored as a source of geometry for garment modeling. Previous work on this topic can be divided into multi-view approaches and single-view approaches. Multi-view approaches use a multi-camera system or a 3D scanner to capture and reconstruct the shape of the clothing. In order for the captured clothing shape to be ready for use in downstream tasks, much effort has been devoted to registration, i.e. the representation of the garment geometry by a consistent mesh topology in a temporal sequence [Bhatnagar et al. 2019; Bradley et al. 2008; Hong et al. 2021; Pons-Moll et al. 2017]. Some work [Chen et al. 2021; Zhang et al. 2017] also estimates the underlying body pose and shape from the clothed human reconstruction. In this paper, multi-view clothing capture provides data to train our appearance modeling network.

Recent progress in computer vision makes it possible to reconstruct clothing shape from a single-view input, which allows a

much simpler capture setup. One line of work treats the problem as a monocular 3D shape regression task with explicit [Alldieck et al. 2019; Zheng et al. 2019] or implicit [Burov et al. 2021; Huang et al. 2020; Li et al. 2020; Saito et al. 2019, 2020; Xiu et al. 2022] shape representation. Alternatively, optimization-based [Habermann et al. 2019; Xiang et al. 2020; Xu et al. 2018] or learning-based approaches [Habermann et al. 2021b, 2020] are used to deform a personalized template shape to match the input image. Most of these methods only focus on estimating clothed body shape without photorealistic texture (except [Li et al. 2020]). In addition, due to the fundamental difficulty of estimating depth from monocular RGB images, the output accuracy of these approaches is deficient for high-quality telepresence. Higher depth accuracy can be achieved in a monocular setup by using depth cameras [Chen et al. 2015; Yu et al. 2018], although the quality of the results still does not match that of multi-view systems.

A notable line of work aims to capture clothing with the assistance of cloth simulation, both in the cases of monocular [Yang et al. 2017, 2018; Yu et al. 2019] and multiview input [Guo et al. 2021; Li et al. 2021a; Stoll et al. 2010]. However, these methods are still limited in terms of output accuracy. Estimating precise physical parameters solely from visual input is still a difficult problem.

2.3 Clothing Animation

Clothing is an integral part of human appearance, and thus the animation of clothing has long been studied as an important component in character animation. *Physics-based simulation* of cloth has been established as a standard tool in the animation industry. Various aspects of cloth simulation, including speed, robustness, and collision detection and response, have been thoroughly studied [Baraff and Witkin 1998; Bouaziz et al. 2014; Bridson et al. 2002; Li et al. 2018, 2021b; Ly et al. 2020; Macklin et al. 2016; Müller et al. 2007]. In order to further improve animation efficiency, there have been efforts to combine physics-based animation with *data-driven* approaches [Kim et al. 2013; Wang et al. 2010]. A notable paradigm is to use simulation to provide training data to supervise machine learning models that predict clothing dynamics given body pose and shape inputs [Bertiche et al. 2021b; Chentanez et al. 2020; Gundogdu et al. 2019; Holden et al. 2019; Jin et al. 2020; Patel et al. 2020; Pfaff et al. 2021; Santesteban et al. 2019, 2021; Vidaurre et al. 2020; Zhang et al. 2021b]. The recent emergence of publicly available 3D captured clothing datasets [Ma et al. 2020; Tiwari et al. 2020] has inspired work that directly learns clothing animation models from such real-world data instead of simulation [Corona et al. 2021; Ma et al. 2021a,b; Saito et al. 2021]. Additionally, the use of a self-supervised physics-based loss [Bertiche et al. 2021a; Santesteban et al. 2022] has been explored for the generation of 3D garment deformations.

Although the aforementioned work can produce reasonable clothing shape deformation with dynamics, photorealistic appearance modeling of the animation output is a non-trivial problem especially for telepresence applications. Classic methods developed specialized BRDF models [Yasuda et al. 1992], offline models for knitware like the lumislice [Xu et al. 2001], or spatially-varying BRDFs which use pattern geometry arrays to improve realism like [Daubert et al.

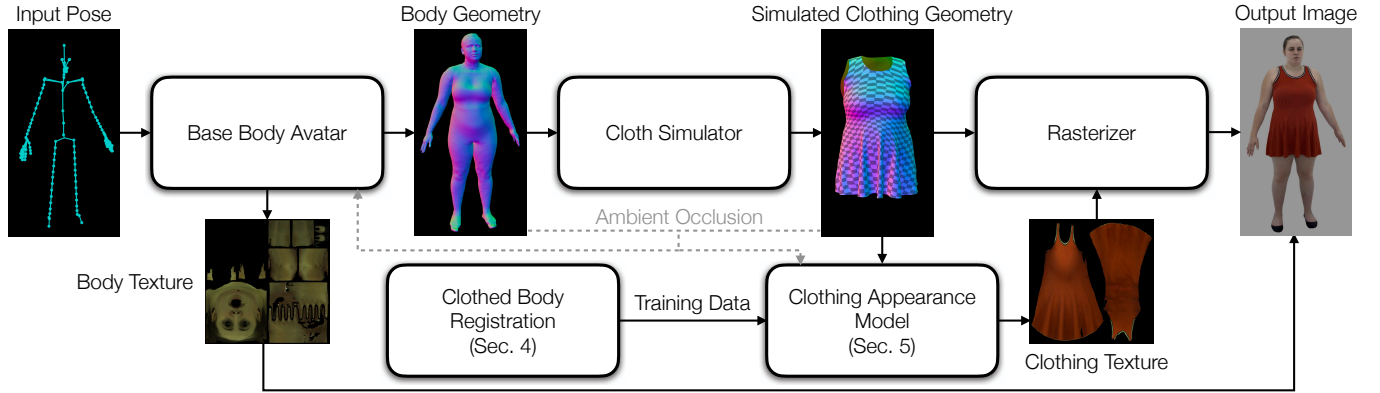


Fig. 2. Our animation pipeline includes three major modules: the base body avatar that predicts body geometry and texture given pose as input, the cloth simulator that generates clothing deformation on top of the body geometry, and the clothing appearance model that predicts photorealistic clothing texture. The appearance model is trained using real captured data with registered body and clothing geometry. The body avatar and clothing appearance model also takes in ambient occlusion between the body and clothing geometry for dynamic shadowing effects. The geometry and texture pairs are then rasterized together to produce the final output image.

2001]. More recently, classic graphic methods evolved to generate cloth fibers procedurally on the GPU in real-time, improving rendering quality [Wu and Yuksel 2017] although still not reaching photorealism. A more complete review of fabric rendering can be found in [Castillo et al. 2019]. Among the learning-based methods, Xiang and colleagues [Xiang et al. 2021] train a full-body avatar that models clothing as a separate layer for photorealistic appearance and temporal deformation, but they show only a single clothing style (T-shirt) and degree of dynamics. Zhang and colleagues [Zhang et al. 2021a] propose a neural rendering approach that produces output images from coarse garments generated by a temporal model. However, this work has several limitations. First, this work shows results mostly in the synthetic domain. The output rendering is detailed but limited in photorealism, because generating highly photorealistic training images for the neural renderer is non-trivial. Second, the method assumes input pairs of body-only images as background and clothed body images as ground truth. In the real-world setting, it is not clear how to obtain such paired data, especially the body-only images. Third, the screen-space garment renderer can have difficulty inferring the depth order between the body in the background layer and the clothing in the foreground layer. By contrast, our method builds photorealistic avatars with highly dynamic clothing using captured data in the real world. We track body and clothing geometry and photometric correspondences at high quality in the captured data, which facilitates the effective learning of the appearance function. Our physically inspired clothing appearance model operates in 3D space, and thus is free from the depth ordering issue even with complex body-clothing interactions.

3 METHOD OVERVIEW

Our goal is to build pose-driven full-body avatars with dynamic clothing and photorealistic appearance. Following [Bagautdinov et al. 2021; Xiang et al. 2021], we train our avatars on multi-view capture sequences of each subject wearing the clothing of interest. At test time, the avatars are animated by a sparse driving signal of

skeleton motion (including facial keypoints if available), and can be rendered in a novel camera viewpoint.

We aim to achieve high-fidelity animation that looks realistic both in terms of appearance of the human subject and temporal clothing dynamics. For this purpose, we develop an animation pipeline consisting of three modules: an underlying body avatar model, physics-based cloth simulation and a clothing appearance model. The underlying body avatar takes as input the skeleton pose (including face conditioning if animating expression) and outputs the body geometry. Given a sequence of body geometry, we then use cloth simulation [Stuyck 2018] to generate clothing geometry with natural and rich dynamics, physically consistent with the motion of the underlying body. Finally, we apply the clothing appearance model to the simulated geometry and generate photorealistic texture, which takes into account not only the clothing geometry but also shadows caused by the occlusion of the avatar body. The shadow cast by the clothing on the body is also modeled similarly in the underlying body avatar. Fig. 2 illustrates our overall pipeline.

Photorealistic full-body Codec Avatars [Bagautdinov et al. 2021; Xiang et al. 2021] and physics-based cloth simulation [Stuyck 2018] have been extensively explored in the existing literature. For these two modules we mostly follow previous work and provide implementation detail in the supplementary document¹. We find efficient photorealistic appearance modeling of clothing to be a critical missing component in such a pipeline, and we tackle two major technical challenges in order to build such a system. On the one hand, as explained in Sec. 5, we develop a deep clothing appearance model that can efficiently generate highly photorealistic clothing texture with dynamic view-dependent and shadowing effects. For the design of the model, we focus on the generalization of the produced appearance, because the input geometry from the cloth simulator can

¹Unless otherwise stated, we use a GPU-based XPBD [Macklin et al. 2016] cloth simulation of a mass-spring system for its superior runtime performance, but our method is not restricted to a specific choice of model or integration technique. Examples generated by a different simulator [Ly et al. 2020] are provided in the supplementary document.

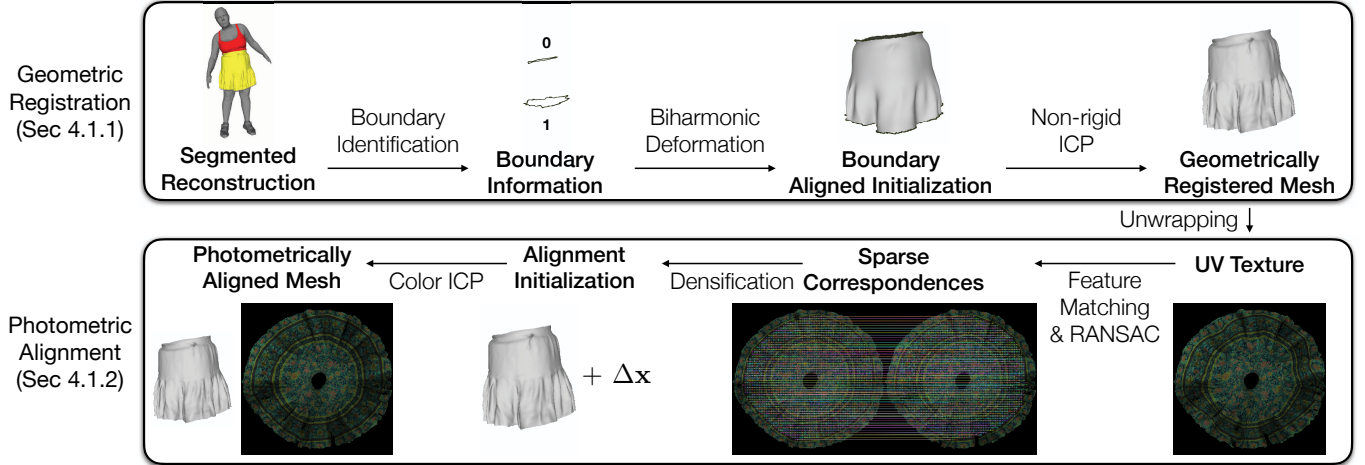


Fig. 3. Our full clothing registration method consists of two major steps: geometric registration and photometric alignment. The geometric registration step aligns the garment surface of the segmented reconstruction with a template mesh. We rely on the boundary information for initialization and then minimize the surface error using non-rigid ICP. In the photometric alignment step, we establish sparse correspondences between the salient region of the unwrapped texture and a template texture. The sparse correspondences are then densified as a displacement for each vertex. This step is followed by color ICP which outputs both a geometrically and photometrically aligned mesh in the template topology.

be different from the tracked clothing geometry used to train the appearance model. On the other hand, in order to generate training data for the appearance model, we extend the previous clothing registration algorithm [Pons-Moll et al. 2017; Xiang et al. 2021] to handle highly dynamic clothing types including a skirt and a dress. In addition, we track photometric correspondences on the garments by matching salient features, which are essential for the modeling of highly textured clothing appearance. The clothing registration step is described in Sec. 4.

4 CLOTHED BODY REGISTRATION

In this section, we introduce our data processing pipeline to obtain training data for the clothing appearance model described in Sec. 5. Because clothing registration is not the core contribution of this paper, we focus on challenges posed by large clothing dynamics and rich texture, which must be addressed to achieve high-fidelity animation.

Our data capture setup is similar to previous work [Xiang et al. 2021]. The pipeline takes multi-view image sequences of a subject as input, and outputs registered meshes of the garment and the underlying body in two separate layers. We follow [Xiang et al. 2021] for raw geometry reconstruction and mesh segmentation into body and clothing regions. We similarly estimate the kinematic body pose and inner-layer body surface.

4.1 Dynamic Clothing Registration

The goal of clothing registration is to represent the clothing geometry in a single mesh topology with consistent correspondences. Our clothing registration method consists of two major steps, geometric registration and photometric alignment. Fig. 3 illustrates the overview of the clothing registration method.

4.1.1 Geometric Registration. In this step, we fit a clothing template to the segmented clothing region of the reconstructed mesh using the non-rigid Iterative Closest Point (ICP) algorithm. The non-rigid ICP algorithm is similar to previous work [Pons-Moll et al. 2017; Xiang et al. 2021] and thus we omit the details here.

In order to track loose and dynamic clothing types (e.g. skirt and dress), it is important to provide good initialization for non-rigid ICP. We observe that the clothing boundaries provide useful information about the overall orientation and deformation of the garment. Therefore, we start by estimating coarse boundary correspondences. Utilizing the fact that each garment has a fixed number of boundaries (for example, two for the skirt and four for the dress), we associate each point on the the target clothing boundary with the template mesh boundary by querying the nearest vertex in the tracked inner-body surface. Given this coarse boundary correspondence, we use Biharmonic Deformation Fields [Jacobson et al. 2010] to solve for per-vertex deformations that satisfy the boundary alignment constraints while minimizing the interior distortion. We use the output of the Biharmonic Deformation Fields as the initialization for the non-rigid ICP algorithm.

4.1.2 Photometric Alignment. The geometric registration method in [Pons-Moll et al. 2017; Xiang et al. 2021] aligns the clothing geometry with a single template topology by minimizing the surface distance, but does not explicitly solve for interior correspondences. In order to effectively model clothing appearance, it is necessary to make sure that each vertex in the template consistently tracks the same color (essentially reflectance), which we call photometric correspondences in this paper. We observe that the chunk-based inverse rendering algorithm in [Xiang et al. 2021] can correct small deviations of the photometric correspondences in the geometric registration step but cannot recover from large errors in the initialization.

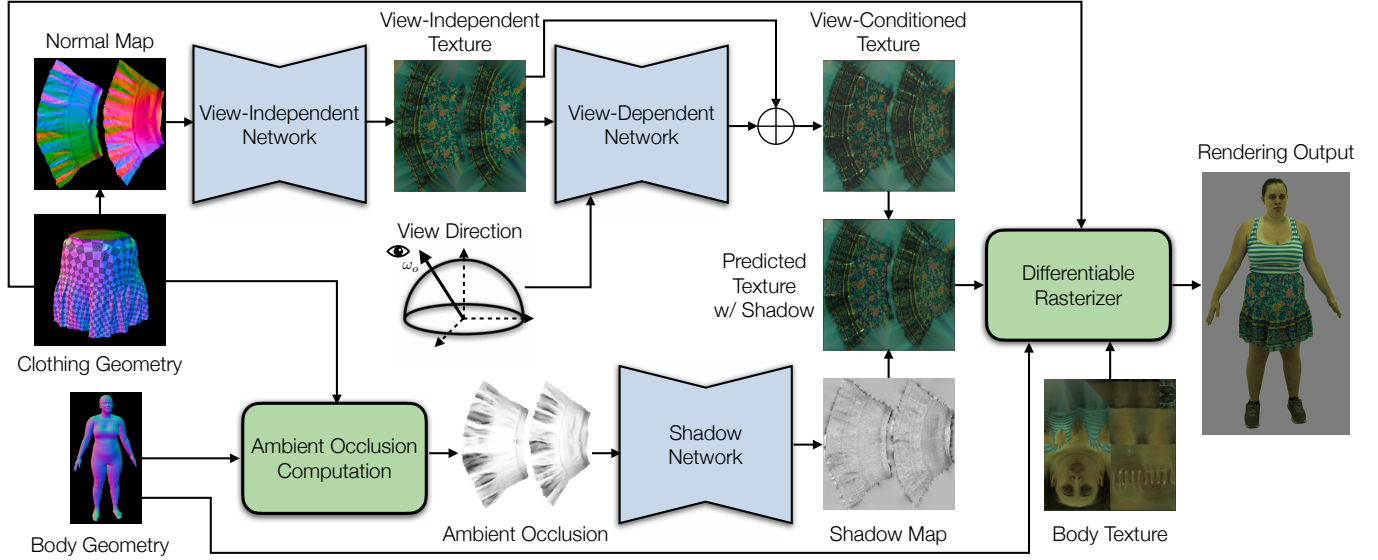


Fig. 4. The architecture of our deep clothing appearance model. We model local diffuse appearance with a view-independent network conditioned on surface normals. The view-dependent network additionally takes in view direction and produces a view-conditioned offset. We explicitly model shadowing using a shadow network that predicts a multiplicative shadow map given ambient occlusion between the body and clothing as input.

Here, we explicitly solve for photometric correspondences by matching salient features in highly textured regions of the garments. For each frame in the sequence, we first unwrap a mean texture from multi-view images to the UV space using the geometrically aligned mesh from the previous step. Then, following [Bogo et al. 2017], we use DeepMatching [Revaud et al. 2016] to establish sparse correspondence pairs between the unwrapped texture and the template texture. We also use RANSAC to prune erroneous correspondences. The sparse correspondences are then densified to each vertex by solving a Laplace’s equation, similar to Biharmonic Deformation Fields in the previous step. Finally, we run color ICP ([Park et al. 2017], our own implementation) between the template and the target mesh to photometrically align all the vertices.

5 DEEP DYNAMIC CLOTHING APPEARANCE MODEL

In this section, we introduce our clothing appearance model which is the key technical component that enables our photorealistic clothing animation system. Given clothing geometry as input, the goal of the clothing appearance model is to generate clothing texture that can be used for rasterization together with the input geometry to produce photorealistic appearance.

We build a data-driven clothing appearance model in order to learn complex photorealistic appearance from real captured image sequences. Several factors need to be taken into account when designing such a model. First, the appearance model is trained on tracked geometry from the previous registration step, but at test time, it takes simulated clothing geometry as input. Therefore, it is essential for the model to bridge the generalization gap between the training and testing data. Second, the generated texture should include various aspects of photorealistic appearance, such as view-dependent effects and dynamic shadowing. Third, for the sake of

efficiency, the model should only involve basic quantities that can be easily derived from the clothing geometry, without computationally expensive operations such as multi-bounce Monte-Carlo ray tracing.

5.1 Background: Rendering Equation

For a particular point \mathbf{x} in the scene, the classical rendering equation [Kajiya 1986] is written as

$$L_o(\mathbf{x}, \omega_o) = \int_{\Omega(\mathbf{n})} f(\mathbf{x}, \omega_i, \omega_o) L_i(\mathbf{x}, \omega_i) (\omega_i \cdot \mathbf{n}) d\omega_i, \quad (1)$$

where ω_i, ω_o denotes the incident and outgoing direction, L_i, L_o denotes the incident and output radiance, f denotes the BRDF function, and Ω denotes the hemisphere where the integral is computed, determined by the normal direction \mathbf{n} . Here the emission term is omitted because clothing usually does not emit radiance.

A common strategy in physics-based rendering is to decompose the appearance into a diffuse component and a specular component. For the diffuse component, the BRDF function f is assumed to be a constant per point. Then the diffuse radiance can be written as

$$L_d(\mathbf{x}) = f_d(\mathbf{x}) \int_{\Omega(\mathbf{n})} L_i(\omega_i) (\omega_i \cdot \mathbf{n}) d\omega_i. \quad (2)$$

When assuming distant lighting and ignoring cast shadows, L_i can be considered as an environment lighting map which represents the consistent illumination condition for the whole captured sequence. Under this circumstance, the integral term is only determined by the normal direction \mathbf{n} at each point, and thus

$$L_d(\mathbf{x}) = f_d(\mathbf{x}) E(\mathbf{n}), \quad E(\mathbf{n}) = \int_{\Omega(\mathbf{n})} L_i(\omega_i) (\omega_i \cdot \mathbf{n}) d\omega_i. \quad (3)$$

Besides the local diffuse component, we need to additionally account for specular (view-dependent) effects and global illumination

effects especially cast shadows. The decomposition of appearance into these three components has direct implication for the design of network architecture described in the following section.

5.2 Model Formulation

From our capture setup we obtain a sequence of registered clothing meshes $\{V_i\}$ and underlying body meshes $\{B_i\}$, paired with multi-view images I_i^c , where i denotes the frame number and c denotes the camera ID. We learn an appearance function F from geometry V_i , B_i and viewpoint v^c to a photorealistic UV texture. Fig. 4 shows our clothing appearance model.

From the analysis in the previous section, the function F should be able to model local diffuse (view-independent) appearance, view-dependent effects and cast shadows. Under the assumption in Eq. 3, the diffuse component at each location is determined by the normal direction and diffuse albedo. Therefore, the view-independent part of F is conditioned on the normal direction \mathbf{n} to encode the illumination direction. We use 2D convolution with untied biases to encode the spatially varying reflectance over the clothing. In our experiments, we show that the normal direction \mathbf{n} , as a local geometric property, is more effective for generalization when used as input conditioning than the absolute vertex positions V_i .

The view-dependent part of the network additionally takes the viewpoint information v_c as input, and predicts an additive view-dependent appearance offset. The viewpoint is represented by the viewing direction vector in the local Tangent-Binormal-Normal (TBN) coordinate at each position on the mesh.

For dynamic shadowing effects, previous work [Bagautdinov et al. 2021; Xiang et al. 2021] demonstrates that shadow maps predicted by a shadow network from ambient occlusion can achieve a good tradeoff between output quality and computation efficiency. Therefore, we follow previous work to use a shadow network to predict a quasi-shadow map, which is multiplied with the view-dependent texture to obtain the final predicted texture. The input ambient occlusion is computed by ray-mesh intersection with both body and clothing geometry.

We use a 2D convolutional neural network (CNN) for the architecture of the appearance model F . All the inputs to the network are converted to 2D feature maps according to a fixed UV mapping. Instead of the autoencoder structure of Codec Avatars [Bagautdinov et al. 2021; Xiang et al. 2021], we adopt the U-Net architecture [Ronneberger et al. 2015] in each of the three parts. This architecture can focus on local information around each input location to learn a generalizable appearance function.

Loss function. The appearance model produces a view-conditioned photorealistic texture for the clothing given the input clothing and underlying body geometry (used for ambient occlusion computation), denoted by $F(V_i, B_i, v_c)$. A differentiable rasterizer R is then used to render body and clothing, denoted by $R(V_i, F(V_i, B_i, v_c))$. We penalize the difference between the output rendering and the raw captured image in the clothing region with the following differentiable rendering loss

$$l = \sum_{i,c} \|(R(V_i, F(V_i, B_i, v_c)) - I_i^c) \odot M_i^c\|_1, \quad (4)$$



Fig. 5. Illustration for ablation studies of the deep clothing appearance model. We compare our full model with representative baselines. Regions of particular interest are marked with black squares. (A) The mean-texture rendering has baked-in shadows. (B) Without the view-dependent component, brightness near the silhouette caused by Fresnel reflection cannot be modeled. (C) The model trained on data without photometric alignment produces blurry results.

where M_i^c denotes the mask of clothing region in I_i^c obtained from image segmentation, and \odot denotes element-wise multiplication. We also rasterize the underlying tracked body geometry with a mean unwrapped body texture together with the clothing for correct body-clothing occlusion, which is omitted in Eq. 4 for simplicity.

6 RESULTS

In this section we present experimental results. We first introduce the data capture setup in Sec. 6.1. Then we evaluate the clothing appearance model presented in Sec. 6.2, followed by full pose-driven animation results including body and clothing in Sec. 6.3. In Sec. 6.4, we show an application of dressing photorealistic avatars, i.e. the animation of clothing on top of novel subjects. Finally, in Sec. 6.5, we present a runtime analysis of our method.

6.1 Data Capture Setup

We capture and process a total of four garments: two T-shirts worn by two different male subjects, and a skirt and a dress worn by a female subject. The four garments span different types of texture patterns, including uniform color, a logo, and tiled floral texture, and thus provide a good testbed for our deep clothing appearance model. Following previous work [Xiang et al. 2021], we treat other clothing

Table 1. Quantitative evaluation of the clothing appearance model applied on tracked clothing geometry. Mean Squared Error (MSE, the lower the better) and Structural Similarity Index Measure (SSIM, the higher the better) are reported. We conduct ablation studies on the effectiveness of each network component, type of input conditioning, and photometric alignment for the training data. We also compare with similar modules in the previous work.

Method	MSE↓	SSIM↑
Network components: mean texture	316.48	0.67
Network components: mean texture + shadow	260.44	0.68
Network components: view-independent + shadow	189.29	0.73
Network components: view-independent + view-dependent	177.27	0.75
Our full model (network components: view-independent + view-dependent + shadow)	155.31	0.76
Geometry input conditioning: raw vertices	206.10	0.72
Geometry input conditioning: unposed normal	160.38	0.76
Geometry input conditioning: unposed vertices	161.08	0.76
Our full model (geometry input conditioning: normal)	155.31	0.76
W/o photometric alignment for training data (test on data w/o photometric alignment)	597.71	0.44
W/o photometric alignment for training data (test on data w/ photometric alignment)	483.73	0.46
Our full model (w/ photometric alignment for training data)	155.31	0.76
Previous method: Dynamic Neural Garments [Zhang et al. 2021a]	326.58	0.57
Previous method: Clothing Codec Avatars [Xiang et al. 2021] (texture only)	261.28	0.64
Previous method: Clothing Codec Avatars [Xiang et al. 2021] (geometry + texture)	379.87	0.58
Our full model	155.31	0.76

worn along with the four garments of interest in the body layer, including pants paired with the T-shirts and the tank top paired with the skirt. These garments move closely together with the body, and can be handled well in the same layer.

For the purpose of dressing avatars, we additionally capture three base body avatars with minimal clothing (tight outfits with greenish color in Fig. 8), including the subject from the skirt and dress capture and two new subjects. In Sec. 6.4, we generate synthetic clothing animation on these three base body avatars.

6.2 Evaluation of the Clothing Appearance Model

In this section, we evaluate the deep clothing appearance model, including ablation studies and comparisons with previous work. We leave out a segment of the captured skirt sequence as the test set, and apply the clothing appearance model on the tracked clothing geometry. In this way, we can directly compute the difference between the rendered clothing output and the original captured images as ground truth. We report the error in the clothing area of the images indicated by the segmentation masks.

6.2.1 Ablation studies: network components. In the first group of experiments in Table 1, we validate the effectiveness of different modules in our clothing appearance model. The simplest version of the appearance model is to directly rasterize a single mean unwrapped texture for all the frames. This extremely cheap rendering method can retain the pattern on the skirt, but no dynamic appearance effects from illumination and shadowing are modeled, which is exemplified by the unnatural baked-in shadows in Row (A) of Fig. 5. We further add view-independent, view-dependent and shadow networks, which correspond to the three dynamic effects we attempt to model. In Row (B) of Fig. 5, we compare the full method with the network without view-dependent effects, which corresponds to the “view-independent + shadow” case in Table 1. The most obvious

difference is the brightness pixel intensity near the silhouette of the clothing in the full method, due to Fresnel reflection when the incident angle is close to 90° . The results in Table 1 verify that the full model with all three modules performs the best.

6.2.2 Ablation studies: input conditioning. In Sec. 5.2, our appearance model is conditioned on surface normals as geometry input. In the second group of experiments in Table 1, we compare this formulation with the one from [Xiang et al. 2021], where the network takes the clothing vertex location after inverse Linear Blend Skinning (LBS) as input, also called “unposed vertices”. We additionally consider the other possibilities of using “raw vertices” (without inverse LBS) and the surface normals of unposed vertices (“unposed normal”). The quantitative results demonstrate the superiority of our proposed formulation. We believe there are two reasons for this performance improvement. First, the surface normals are a localized representation, and thus allow better generalization of the model than the absolute positions of vertices. Second, given the analysis in Sec. 5.1, we know that the surface normals encode the direction of incident illumination, thus following the principles of physics-based rendering.

6.2.3 Ablation studies: photometric alignment. In the third group of experiments, we verify the importance of photometric alignment for highly textured clothing. As a comparison, we also train an appearance model using clothing meshes generated by geometric registration without photometric alignment. Both the quantitative results in Table 1 and the qualitative evaluation in Row (C) of Fig. 5 show significant degradation when photometric alignment is removed. The output images are very blurred when the model is trained on such data. Different positions in the textured regions of the clothing have a different reflectance (BRDF), so the tracked

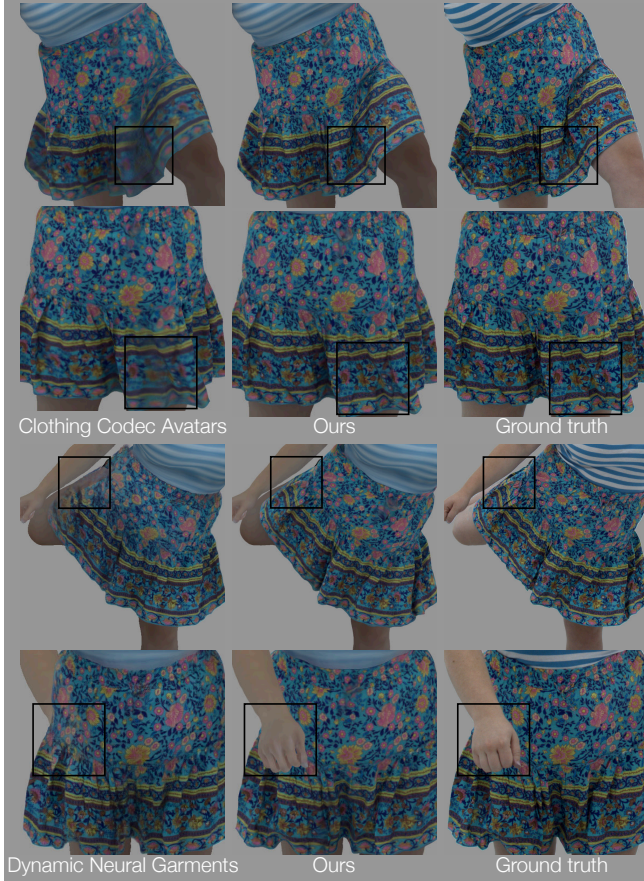


Fig. 6. Comparison between our clothing appearance model and similar modules in previous methods. The Clothing Codec Avatars [Xiang et al. 2021] have difficulties modeling the highly dynamic region of the skirt. A major issue with the screen-space garment renderer in Dynamic Neural Garments [Zhang et al. 2021a] is the incorrect depth ordering between body in the background layer and clothing in the foreground layer.

photometric correspondences must be consistently aligned with the captured images in order to learn the correct appearance function.

6.2.4 Comparison with previous methods. We compare the clothing appearance model with similar modules from two previous methods: Clothing Codec Avatars [Xiang et al. 2021] and Dynamic Neural Garments [Zhang et al. 2021a]. For these experiments, we use the same training data as our full model and only test the influence of model choice.

Clothing Codec Avatars [Xiang et al. 2021] are deep neural networks with an autoencoder architecture. The encoder takes unposed clothing geometry as input and outputs a latent code which is fed to the decoder to generate both geometry and photorealistic view-dependent texture, so the network can be used as clothing appearance model as well². We test the model in two ways: (1) rendering

²In the original work, the encoder takes both geometry and texture as input. We adapt the encoder input for this experiment. The Clothing Codec Avatars have the same shadow network that takes ambient occlusion as input.

tracked geometry (identical to input) with decoded texture (2) rendering decoded geometry with decoded texture. As shown in the last group of results in [Xiang et al. 2021] and in Fig. 6, our deep appearance model achieves the lowest rendering errors, which we attribute to three factors. First, the bottleneck structure of Clothing Codec Avatars is essential for controlling the avatar with a low-dimensional latent code but disadvantageous for learning detailed appearance compared with the U-Net architecture of our appearance model. This analysis is further supported by comparing Clothing Codec Avatars with the appearance model with the same input conditioning (“unposed vertices”) but a U-Net architecture among the second group of experiments in Table 1. Second, Clothing Codec Avatars are trained using a differentiable rendering loss on decoded geometry and texture. However, the high frequency dynamics of the skirt itself is hard for the network to learn, further jeopardizing the modeling of texture, manifested by the blurry regions at the bottom of the skirt in Fig. 6. Third, in the ablation studies, we verified that the input conditioning of unposed vertices is not as good as the surface normal used by our appearance model. All these factors demonstrate that for our goal, it is advantageous to adopt the novel formulation of the appearance model instead of that of Clothing Codec Avatars.

Dynamic Neural Garments (DNG) [Zhang et al. 2021a] models garment appearance with an improved version of Deferred Neural Rendering [Thies et al. 2019] that is more temporally coherent. The code for the garment renderer is released (not including the garment geometry prediction part), so we train the renderer using the same data as our appearance model for comparison. We generate the background images by rasterizing tracked body geometry with the fixed mean body texture, and composite it with the clothing region of the original image as ground truth. The results are shown in Table 1 and Fig. 6. As a screen-space method, the garment renderer in DNG has difficulty reasoning about the depth order between the body and clothing layer when there are complex interactions between the arms or legs and the garment. This limitation is acknowledged in the original paper [Zhang et al. 2021a] and contributes to the high quantitative error in Table 1 when skin color appears in the clothing region of ground truth images.

6.3 Pose-Driven Animation

In this section, we present pose-driven animation results for the four garments on top of the original body avatars that are captured together with the clothing. These results are generated by the complete animation pipeline shown in Fig. 2. We make modest efforts to select the physical parameters for the simulation to roughly resemble the material behavior in the training sequence.

We compare the results with the output of the full method of Clothing Codec Avatars [Xiang et al. 2021]. Some examples are shown in Fig. 7 and full animation results are shown in the supplementary video. The most obvious advantage of our method is the temporal dynamics, especially for the loose clothing that does not exactly follow the body motion. Although clothing is modeled explicitly in a separate layer in [Xiang et al. 2021], the mapping from body motion to clothing dynamics is a highly complex, temporally dependent function that is not easy to learn well. Our approach



Fig. 7. Pose-driven animation results in comparison with [Xiang et al. 2021]. In each row we show one result generated by both methods and a held out captured image for reference from two different views.

further separates the problem into dynamics and appearance to best capture both these aspects.

6.4 Application: Dressing Avatars

Our formulation naturally enables garment animations on top of base body avatars that are different from the original captured sequence. Because we model the clothing separately from the body layer, we simply replace the base body avatar and apply the physics-based simulation (with scaling if necessary) and clothing appearance model in exactly the same way as the default pipeline. We show dressed avatars of the same identity (wearing a tight capture suit)

and different identities. The animation results are shown in Fig. 8 and in the supplementary video.

6.4.1 Application: Editing garment size. We edit the garment size by changing the rest length of the garment template in the cloth simulator and keeping the same set of physical parameters. We maintain the same mesh topology, so that our appearance model can be directly applied to the edited garment. The animation results after size editing are shown in Fig. 9. This manipulation is only possible because both the physics-based simulator and our appearance model generalize well to unseen clothing configurations.



Fig. 8. Dressing novel avatars. In the top row, a minimally clothed body avatar of the same actor as the original capture is shown in a skirt and dress. The middle and bottom rows show dressed avatars created from other actors. The bottom left results are dressed with two pieces of clothing together, a skirt and T-shirt.

6.5 Runtime Analysis

In this section we report the runtime for the key modules of our algorithm: the physics-based cloth simulator, the base body avatar and the deep clothing appearance model.

We use a GPU-based cloth simulator with the eXtended Position Based Dynamics (XPBD) formulation [Macklin et al. 2016]. When simulating a garment of 35k vertices on top of a body sequence of 9k vertices, the average runtime per frame for the simulation solver



Fig. 9. We edit garment sizes by scaling the template shape used in the physics-based simulation. Although the clothing appearance models are trained only using captured data of a particular size, they generalize well to the edited garments of different sizes.

is between 8 ms ~ 10 ms when taking 20 steps per frame on a Nvidia Tesla V100 GPU.

The base body avatar and the clothing appearance models are implemented using the PyTorch deep learning framework [Paszke et al. 2019]. For deployment, we use PyTorch Just-In-Time (JIT) compilation to improve performance. The average runtime for those two modules combined together is 78 ms on a Nvidia RTX 3090 GPU on a Lambda workstation. We use multiple GPUs to parallelize the inference. For example, we can achieve > 30 fps with 3 GPUs in parallel. In the supplementary video, we show a demo of viewing an animated sequence in a VR headset, where the cloth simulation is precomputed and the body avatars and clothing appearance are rendered in real-time. Although we currently run the cloth simulator and the neural networks separately, given the modest computation for each of the three modules, our pipeline has the potential to be integrated into a complete real-time online system.

7 DISCUSSION

In this work, we present efficient full-body clothed avatars with physically plausible dynamics and photorealistic appearance. Our

method achieves high-fidelity registration of dynamic cloth geometry and learns a deep clothing appearance model based on the registered geometry for training. At test time, the clothing is deformed through physics-based simulation and is rendered by the appearance model. To bridge the gap between tracked cloth geometry at training time and simulated clothing geometry at test time, we design an input representation and an architecture that consists of three modules inspired by the rendering equation. Once a garment is modeled, our system can use it to *dress* a novel avatar.

Limitations. Clothing has a very large space of geometric and appearance variations. Although our method handles single-layer loose clothing much better than prior work, it may have difficulty in dealing with large folds of extremely loose clothing such as a Kimono or multi-layered clothing such as a coat and dress where the inner layers are significantly occluded. The photometric registration may fail when the feature matching is inaccurate due to bad initialization and/or repeated patterns. Some possibilities for improvement include using simulated clothing data with high-quality offline rendering to train the appearance model, or to utilize clothing with printed markers for registration [Halimi et al. 2022; White et al. 2007].

In this work we manually adjust the physical parameters of the simulated garments. Optimization-based parameter estimation may improve the accuracy by leveraging recent advancements in differentiable cloth simulators [Li et al. 2022; Liang et al. 2019]. In addition, the physics-based cloth simulation may face challenges in handling highly complicated hand-cloth interactions such as dragging and pinching, due to the real-time constraint imposed on the simulation.

Although the clothing appearance model is trained using the data captured for only one subject, the shadow branch shows reasonable generalization ability when used to dress novel avatars because it is influenced by the body shape only indirectly through the occlusion of rays. Nevertheless, how the model works for extreme body shapes remains untested due to the limited availability of minimally clothed data. A systematic investigation of shadowing across different garments and identities is an interesting future direction. Furthermore, we do not consider complicated interreflection of lighting between clothing and body and between the folds of clothing. We also assume consistent illumination conditions when dressing novel avatars. When this assumption does not hold, slight inconsistency can be observed, for example, in the two-cloth animation results (T-shirt and skirt). In the future, we would like to relax this constraint by incorporating a relighting capability [Bi et al. 2021] to our avatar and clothing models with an updated capture setup.

ACKNOWLEDGMENTS

We thank Anuj Pahuja for his contribution in exporting the clothing appearance model and building the VR demo. Donglai Xiang is supported by the Meta AI Mentorship (AIM) program.

REFERENCES

- Thiemo Aaldieck, Gerard Pons-Moll, Christian Theobalt, and Marcus Magnor. 2019. Tex2shape: Detailed full human body geometry from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.
- Timur Bagautdinov, Chenglei Wu, Tomas Simon, Fabian Prada, Takaaki Shiratori, Shih-En Wei, Weipeng Xu, Yaser Sheikh, and Jason Saragih. 2021. Driving-signal aware full-body avatars. *ACM Transactions on Graphics (TOG)* 40, 4 (2021).
- David Baraff and Andrew Witkin. 1998. Large steps in cloth simulation. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Proceedings*.
- Hugo Bertiche, Meysam Madadi, and Sergio Escalera. 2021a. PBNS: physically based neural simulation for unsupervised garment pose space deformation. *ACM Transactions on Graphics (TOG)* 40, 6 (2021).
- Hugo Bertiche, Meysam Madadi, Emilio Tylson, and Sergio Escalera. 2021b. DeePSD: Automatic deep skinning and pose space deformation for 3D garment animation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.
- Bharat Lal Bhatnagar, Garvita Tiwari, Christian Theobalt, and Gerard Pons-Moll. 2019. Multi-garment net: Learning to dress 3d people from images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.
- Sai Bi, Stephen Lombardi, Shunsuke Saito, Tomas Simon, Shih-En Wei, Kevyn Mcphail, Ravi Ramamoorthi, Yaser Sheikh, and Jason Saragih. 2021. Deep relightable appearance models for animatable faces. *ACM Transactions on Graphics (TOG)* 40, 4 (2021).
- Federica Bogo, Javier Romero, Gerard Pons-Moll, and Michael J Black. 2017. Dynamic FAUST: Registering human bodies in motion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
- Sofien Bouaziz, Sebastian Martin, Tiantian Liu, Ladislav Kavan, and Mark Pauly. 2014. Projective dynamics: Fusing constraint projections for fast simulation. *ACM Transactions on Graphics (TOG)* 33, 4 (2014).
- Derek Bradley, Tiberiu Popa, Alla Sheffer, Wolfgang Heidrich, and Tamy Boubekeur. 2008. Markerless garment capture. *ACM Transactions on Graphics (TOG)* 27, 3 (2008).
- Robert Bridson, Ronald Fedkiw, and John Anderson. 2002. Robust treatment of collisions, contact and friction for cloth animation. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Proceedings*.
- Andrei Burov, Matthias Nießner, and Justus Thies. 2021. Dynamic surface function networks for clothed human bodies. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.
- Joel Carranza, Christian Theobalt, Marcus A Magnor, and Hans-Peter Seidel. 2003. Free-viewpoint video of human actors. *ACM Transactions on Graphics (TOG)* 22, 3 (2003).
- Carlos Castillo, Jorge López-Moreno, and Carlos Aliaga. 2019. Recent advances in fabric appearance reproduction. *Computers & Graphics* 84 (2019).
- Charles-Félix Chabert, Per Einarsson, Andrew Jones, Bruce Lamond, Wan-Chun Ma, Sebastian Sylwan, Tim Hawkins, and Paul Debevec. 2006. Relighting human locomotion with flowed reflectance fields. In *Eurographics Symposium on Rendering*.
- 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*.
- Xin Chen, Anqi Pang, Wei Yang, Peihao Wang, Lan Xu, and Jingyi Yu. 2021. TightCap: 3D Human Shape Capture with Clothing Tightness Field. *ACM Transactions on Graphics (TOG)* 41, 1 (2021).
- Xiaowu Chen, Bin Zhou, Fei-Xiang Lu, Lin Wang, Lang Bi, and Ping Tan. 2015. Garment modeling with a depth camera. *ACM Transactions on Graphics (TOG)* 34, 6 (2015).
- 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. In *Computer Graphics Forum*, Vol. 39.
- Enric Corona, Albert Pumarola, Guillem Alenya, Gerard Pons-Moll, and Francesc Moreno-Noguer. 2021. SMPLicit: Topology-aware generative model for clothed people. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- Katja Daubert, Hendrik Lensch, Wolfgang Heidrich, and Hans-Peter Seidel. 2001. Efficient cloth modeling and rendering. In *Eurographics Workshop on Rendering Techniques*.
- Edilson de Aguiar, Carsten Stoll, Christian Theobalt, Naveed Ahmed, Hans-Peter Seidel, and Sebastian Thrun. 2008. Performance capture from sparse multi-view video. *ACM Transactions on Graphics (TOG)* 27, 3 (2008).
- Martin Eisemann, Bert De Decker, Marcus Magnor, Philippe Bekaert, Edilson De Aguiar, Naveed Ahmed, Christian Theobalt, and Anita Sellent. 2008. Floating textures. In *Computer Graphics Forum*, Vol. 27.
- Artur Grigorev, Karim Isakov, Anastasia Ianina, Renat Bashirov, Ilya Zakharkin, Alexander Vakhitov, and Victor Lempitsky. 2021. Stylepeople: A generative model of fullbody human avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- 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*.
- Jingfan Guo, Jie Li, Rahul Narain, and Hyun Soo Park. 2021. Inverse Simulation: Reconstructing Dynamic Geometry of Clothed Humans via Optimal Control. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- Marc Habermann, Lingjie Liu, Weipeng Xu, Michael Zollhoefer, Gerard Pons-Moll, and Christian Theobalt. 2021a. Real-time deep dynamic characters. *ACM Transactions on Graphics (TOG)* 40, 4 (2021).
- Marc Habermann, Weipeng Xu, Michael Zollhoefer, Gerard Pons-Moll, and Christian Theobalt. 2019. LiveCap: Real-time human performance capture from monocular video. *ACM Transactions On Graphics (TOG)* 38, 2 (2019).
- Marc Habermann, Weipeng Xu, Michael Zollhoefer, Gerard Pons-Moll, and Christian Theobalt. 2021b. A Deeper Look into DeepCap. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021).
- Marc Habermann, Weipeng Xu, Michael Zollhofer, Gerard Pons-Moll, and Christian Theobalt. 2020. Deepcap: Monocular human performance capture using weak supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- Oshri Halimi, Tuur Stuyck, Donglai Xiang, Timur Bagautdinov, He Wen, Ron Kimmel, Takaaki Shiratori, Chenglei Wu, Yaser Sheikh, and Fabian Prada. 2022. Pattern-Based Cloth Registration and Sparse-View Animation. *ACM Transactions on Graphics (TOG)* 41, 6 (2022).
- Daniel Holden, Bang Chi Duong, Sayantan Datta, and Derek Nowrouzezahrai. 2019. Subspace neural physics: Fast data-driven interactive simulation. In *Proceedings of the 18th annual ACM SIGGRAPH/Eurographics Symposium on Computer Animation*.
- Fangzhou Hong, Liang Pan, Zhongang Cai, and Ziwei Liu. 2021. Garment4D: Garment Reconstruction from Point Cloud Sequences. *Advances in Neural Information Processing Systems* 34 (2021).
- Zeng Huang, Yuanlu Xu, Christoph Lassner, Hao Li, and Tony Tung. 2020. Arch: Animatable reconstruction of clothed humans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- Alec Jacobson, Elif Tosun, Olga Sorkine, and Denis Zorin. 2010. Mixed finite elements for variational surface modeling. In *Computer Graphics Forum*, Vol. 29.
- Arjun Jain, Thorsten Thormählen, Hans-Peter Seidel, and Christian Theobalt. 2010. Moviereshape: Tracking and reshaping of humans in videos. *ACM Transactions on Graphics (TOG)* 29, 6 (2010).

- Ning Jin, Yilin Zhu, Zhenglin Geng, and Ronald Fedkiw. 2020. A Pixel-Based Framework for Data-Driven Clothing. In *Computer Graphics Forum*, Vol. 39.
- James T Kajiya. 1986. The rendering equation. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Proceedings*.
- Doyub Kim, Wojong Koh, Rahul Narain, Kayvon Fatahalian, Adrien Treuille, and James F O'Brien. 2013. Near-exhaustive precomputation of secondary cloth effects. *ACM Transactions on Graphics (TOG)* 32, 4 (2013).
- Youngjoong Kwon, Dahun Kim, Duygu Ceylan, and Henry Fuchs. 2021. Neural human performer: Learning generalizable radiance fields for human performance rendering. *Advances in Neural Information Processing Systems* 34 (2021).
- Jie Li, Gilles Daviet, Rahul Narain, Florence Bertails-Descoubes, Matthew Overby, George E Brown, and Laurence Boissieux. 2018. An implicit frictional contact solver for adaptive cloth simulation. *ACM Transactions on Graphics (TOG)* 37, 4 (2018).
- Minchen Li, Danny M Kaufman, and Chenfanfu Jiang. 2021b. Codimensional incremental potential contact. *ACM Transactions on Graphics (TOG)* 40, 4 (2021).
- Ruilong Li, Yuliang Xiu, Shunsuke Saito, Zeng Huang, Kyle Olszewski, and Hao Li. 2020. Monocular real-time volumetric performance capture. In *European Conference on Computer Vision*.
- Yifei Li, Tao Du, Kui Wu, Jie Xu, and Wojciech Matusik. 2022. DiffCloth: Differentiable cloth simulation with dry frictional contact. *ACM Transactions on Graphics (TOG)* (2022).
- Yue Li, Marc Habermann, Bernhard Thomaszewski, Stelian Coros, Thabo Beeler, and Christian Theobalt. 2021a. Deep physics-aware inference of cloth deformation for monocular human performance capture. In *2021 International Conference on 3D Vision (3DV)*. IEEE.
- Junbang Liang, Ming Lin, and Vladlen Koltun. 2019. Differentiable cloth simulation for inverse problems. *Advances in Neural Information Processing Systems* 32 (2019).
- Lingjie Liu, Marc Habermann, Viktor Rudnev, Kripasindhu Sarkar, Jiatao Gu, and Christian Theobalt. 2021. Neural actor: Neural free-view synthesis of human actors with pose control. *ACM Transactions on Graphics (TOG)* 40, 6 (2021).
- Mickaël Ly, Jean Jouve, Laurence Boissieux, and Florence Bertails-Descoubes. 2020. Projective dynamics with dry frictional contact. *ACM Transactions on Graphics (TOG)* 39, 4 (2020).
- Qianli Ma, Shunsuke Saito, Jinlong Yang, Siyu Tang, and Michael J Black. 2021a. SCALE: Modeling clothed humans with a surface codec of articulated local elements. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- Qianli Ma, Jinlong Yang, Siyu Tang, and Michael J Black. 2021b. The power of points for modeling humans in clothing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.
- Miles Macklin, Matthias Müller, and Nuttapon Chentanez. 2016. XPBD: position-based simulation of compliant constrained dynamics. In *Proceedings of the 9th International Conference on Motion in Games*.
- Eder Miguel, Derek Bradley, Bernhard Thomaszewski, Bernd Bickel, Wojciech Matusik, Miguel A Otaduy, and Steve Marschner. 2012. Data-driven estimation of cloth simulation models. In *Computer Graphics Forum*, Vol. 31.
- Ben Mildenhall, Pratul P Srinivasan, Matthew Tatencik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. 2020. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European Conference on Computer Vision*.
- Matthias Müller, Bruno Heidelberger, Marcus Hennix, and John Ratcliff. 2007. Position based dynamics. *Journal of Visual Communication and Image Representation* 18, 2 (2007), 109–118.
- Atsuhiko Noguchi, Xiao Sun, Stephen Lin, and Tatsuya Harada. 2021. Neural articulated radiance field. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.
- Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. 2017. Colored point cloud registration revisited. In *Proceedings of the IEEE International Conference on Computer Vision*.
- Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems* 32 (2019).
- 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- Sida Peng, Junting Dong, Qianqian Wang, Shangzhan Zhang, Qing Shuai, Xiaowei Zhou, and Hujun Bao. 2021a. Animatable neural radiance fields for modeling dynamic human bodies. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.
- Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. 2021b. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- Tobias Pfaff, Meire Fortunato, Alvaro Sanchez-Gonzalez, and Peter Battaglia. 2021. Learning Mesh-Based Simulation with Graph Networks. In *International Conference on Learning Representations*.
- 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).
- Sergey Prokudin, Michael J Black, and Javier Romero. 2021. SmpIpx: Neural avatars from 3d human models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*.
- Amit Raj, Julian Tanke, James Hays, Minh Vo, Carsten Stoll, and Christoph Lassner. 2021. Anr: Articulated neural rendering for virtual avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- Edoardo Remelli, Timur Bagautdinov, Shunsuke Saito, Chenglei Wu, Tomas Simon, Shih-En Wei, Kaiwen Guo, Zhe Cao, Fabian Prada, Jason Saragih, et al. 2022. Drivable Volumetric Avatars using Texel-Aligned Features. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Proceedings*.
- Jerome Revaud, Philippe Weinzaepfel, Zaid Harchaoui, and Cordelia Schmid. 2016. Deepmatching: Hierarchical deformable dense matching. *International Journal of Computer Vision* 120, 3 (2016).
- Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*. Springer, 234–241.
- Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. 2019. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.
- Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. 2020. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- Shunsuke Saito, Jinlong Yang, Qianli Ma, and Michael J Black. 2021. SCANimate: Weakly supervised learning of skinned clothed avatar networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- Igor Santesteban, Miguel A Otaduy, and Dan Casas. 2019. Learning-based animation of clothing for virtual try-on. In *Computer Graphics Forum*, Vol. 38.
- Igor Santesteban, Miguel A Otaduy, and Dan Casas. 2022. SNUG: Self-Supervised Neural Dynamic Garments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- Igor Santesteban, Nils Thuerey, Miguel A Otaduy, and Dan Casas. 2021. Self-Supervised Collision Handling via Generative 3D Garment Models for Virtual Try-On. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- Aliaksandra Shysheya, Egor Zakharov, Kara-Ali Aliev, Renat Bashirov, Egor Burkov, Karim Isakov, Aleksei Ivakhnenko, Yury Malkov, Igor Pasechnik, Dmitry Ulyanov, et al. 2019. Textured neural avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- Jonathan Starck and Adrian Hilton. 2007. Surface capture for performance-based animation. *IEEE Computer Graphics and Applications* 27, 3 (2007).
- Carsten Stoll, Juergen Gall, Edison De Aguiar, Sebastian Thrun, and Christian Theobalt. 2010. Video-based reconstruction of animatable human characters. *ACM Transactions on Graphics (TOG)* 29, 6 (2010).
- Tuur Stuyck. 2018. Cloth Simulation for Computer Graphics. *Synthesis Lectures on Visual Computing: Computer Graphics, Animation, Computational Photography, and Imaging* 10, 3 (2018), 1–121.
- 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* 34 (2021).
- Justus Thies, Michael Zollhöfer, and Matthias Nießner. 2019. Deferred neural rendering: Image synthesis using neural textures. *ACM Transactions on Graphics (TOG)* 38, 4 (2019).
- 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*.
- Raquel Vidas, Igor Santesteban, Elena Garces, and Dan Casas. 2020. Fully Convolutional Graph Neural Networks for Parametric Virtual Try-On. In *Computer Graphics Forum*, Vol. 39.
- Marco Volino, Dan Casas, John P Collomosse, and Adrian Hilton. 2014. Optimal representation of multi-view video. In *Proceedings of British Machine Vision Conference*.
- Huamin Wang, Florian Hecht, Ravi Ramamoorthi, and James F O'Brien. 2010. Example-based wrinkle synthesis for clothing animation. *ACM Transactions on Graphics (TOG)* 29, 4 (2010).
- Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. 2018. Video-to-Video Synthesis. *Advances in Neural Information Processing Systems* 31 (2018).
- Ryan White, Keenan Crane, and David A Forsyth. 2007. Capturing and animating occluded cloth. *ACM Transactions on Graphics (TOG)* 26, 3 (2007).
- Kui Wu and Cem Yuksel. 2017. Real-time cloth rendering with fiber-level detail. *IEEE Transactions on Visualization and Computer Graphics* 25, 2 (2017).
- Donglai Xiang, Fabian Prada, Timur Bagautdinov, Weipeng Xu, Yuan Dong, He Wen, Jessica Hodgins, and Chenglei Wu. 2021. Modeling clothing as a separate layer for

- an animatable human avatar. *ACM Transactions on Graphics (TOG)* 40, 6 (2021).
- Donglai Xiang, Fabian Prada, Chenglei Wu, and Jessica Hodgins. 2020. Monoclothcap: Towards temporally coherent clothing capture from monocular rgb video. In *2020 International Conference on 3D Vision (3DV)*. IEEE.
- Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J Black. 2022. Icon: Implicit clothed humans obtained from normals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- Hongyi Xu, Thiemo Alldieck, and Cristian Sminchisescu. 2021. H-nerf: Neural radiance fields for rendering and temporal reconstruction of humans in motion. *Advances in Neural Information Processing Systems* 34 (2021).
- 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 Transactions on Graphics (TOG)* 37, 2 (2018).
- Ying-Qing Xu, Yanyun Chen, Stephen Lin, Hua Zhong, Enhua Wu, Baining Guo, and Heung-Yeung Shum. 2001. Photorealistic rendering of knitwear using the lumislice. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Proceedings*.
- Shan Yang, Junbang Liang, and Ming C Lin. 2017. Learning-based cloth material recovery from video. In *Proceedings of the IEEE International Conference on Computer Vision*.
- Shan Yang, Zherong Pan, Tanya Amert, Ke Wang, Licheng Yu, Tamara Berg, and Ming C Lin. 2018. Physics-inspired garment recovery from a single-view image. *ACM Transactions on Graphics (TOG)* 37, 5 (2018).
- Takami Yasuda, Shigeki Yokoi, and Jun-ichiro Toriwaki. 1992. A shading model for cloth objects. *IEEE Computer Graphics and Applications* 12, 6 (1992).
- Tao Yu, Zerong Zheng, Kaiwen Guo, Jianhui Zhao, Qionghai Dai, Hao Li, Gerard Pons-Moll, and Yebin Liu. 2018. Doublefusion: Real-time capture of human performances with inner body shapes from a single depth sensor. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- 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*.
- 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*.
- Meng Zhang, Duygu Ceylan, Tuanfeng Y. Wang, and Niloy J. Mitra. 2021a. Dynamic Neural Garments. *ACM Transactions on Graphics (TOG)* 40, 6 (2021).
- Meng Zhang, Tuanfeng Wang, Duygu Ceylan, and Niloy J Mitra. 2021b. Deep detail enhancement for any garment. In *Computer Graphics Forum*, Vol. 40.
- Zerong Zheng, Han Huang, Tao Yu, Hongwen Zhang, Yandong Guo, and Yebin Liu. 2022. Structured local radiance fields for human avatar modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- Zerong Zheng, Tao Yu, Yixuan Wei, Qionghai Dai, and Yebin Liu. 2019. Deephuman: 3d human reconstruction from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.
- Tiancheng Zhi, Christoph Lassner, Tony Tung, Carsten Stoll, Srinivasa G Narasimhan, and Minh Vo. 2020. Texmesh: Reconstructing detailed human texture and geometry from rgb-d video. In *European Conference on Computer Vision*.