Modeling Dynamic Clothing for Data-Driven Photorealistic Avatars

Donglai Xiang
donglaix@cs.cmu.edu
Carnegie Mellon University
USA

ABSTRACT

This thesis presents research on building photorealistic avatars of humans wearing complex clothing in a data-driven manner. Such avatars will be a critical technology to enable future applications such as VR/AR and virtual content creation. Loose-fitting clothing poses a significant challenge for avatar modeling due to its large deformation space. We address the challenge by unifying three components of avatar modeling: model-based statistical prior from pre-captured data, physics-based prior from simulation, and real-time measurement from sparse sensor input. First, we introduce a separate two-layer representation that allows us to disentangle the dynamics between the pose-driven body part and temporally-dependent clothing part. Second, we further combine physics-based cloth simulation with a physics-inspired neural rendering model to generate rich and natural dynamics and appearance even for challenging clothing such as a skirt and a dress. Last, we go beyond pose-driven animation and incorporate online sensor input into the avatars to achieve more faithful telepresence of clothing.

CCS CONCEPTS

• **Computing methodologies** → **Computer graphics**; **Computer vision**.

KEYWORDS

Digital avatars, garment animation

ACM Reference Format:
Donglai Xiang. 2023. Modeling Dynamic Clothing for Data-Driven Photorealistic Avatars. In *SIGGRAPH Asia 2023 Doctoral Consortium (SA Doctoral Consortium '23), December 12–15, 2023, Sydney, NSW, Australia.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3623053.3623373

1 INTRODUCTION

Photorealistic digital humans are a key capability for enabling social telepresence, which is one of the key applications for Virtual Reality (VR) or Augmented Reality (AR). Such a technology would allow people wearing VR/AR devices to communicate and interact with friends and colleagues in an immersive way that is natural and compelling, because their partners are represented in a way that is indistinguishable from reality. If we are successful in developing this technology, it will open up a new way for people around the world to remain connected without geographic constraints.

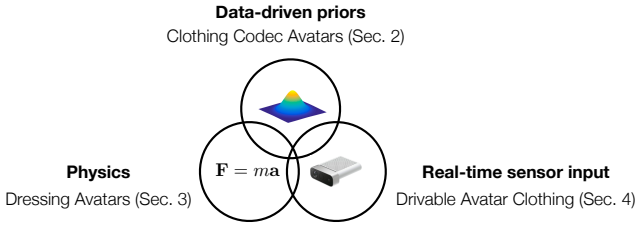

Figure 1: In this thesis, the three major perspectives for modeling photorealistic clothed avatars are data-driven priors, physics, and real-time sensor input.

One key question is how to create high-fidelity digital avatars that are photorealistic and resemble the appearance of individual subjects. One milestone is the development of the Codec Avatar [Lombardi et al. 2018], where the geometry and photorealistic appearance of human heads can be compressed into a low-dimensional latent space and then decoded for display efficiently by a Variational Autoencoder (VAE) [Kingma and Welling 2013] trained with captured human imagery. The absence of other body parts such as torso and hands is one of the major limitations of the first iteration of this technology. Recently, photorealistic full-body avatars [Bagautdinov et al. 2021; Habermann et al. 2021a; Liu et al. 2021] have been developed so that the communication signals conveyed by body and hands can also be represented. The central idea behind these avatars is to model large, skeleton-level deformation with skinning techniques to allow control through body joint angles.

The wide variety of clothing, however, poses a significant challenge in the modeling of geometry and appearance. The root of this challenge is the huge deformation space and rapid dynamics of clothing as it is driven by the underlying human body. Loose-fitting garments present particular challenges because they do not tightly follow the motion of the underlying body, and their deformation can go far beyond what the skeleton-level transformation can describe. Clothing does not contain universal identifiable keypoints to assist tracking of deformation, such as those commonly used for the tracking of the face [Wu et al. 2018] and body [Bogo et al. 2016].

The challenge is manifest in various aspects of the modeling of clothing, including tracking, animation and rendering, which are all required to enable high-fidelity clothing for photorealistic avatars. The efficient learning of dynamic appearance in NN-based avatars often requires tracking, or registration, of the clothing geometry, which is inherently difficult due to its rapid motion and abundance of folds and wrinkles. Existing data-driven approaches

SA Doctoral Consortium '23, December 12–15, 2023, Sydney, NSW, Australia

ACM ISBN 979-8-4007-0392-8/23/12.
https://doi.org/10.1145/3623053.3623373

also struggle to generate high-quality *animation* of clothing, because the mapping from body pose to the dynamics of clothing is a complex, nonlinear and history-dependent function. Furthermore, the sophisticated deformation of clothing also leads to complicated illumination and shadowing effects. Taken together with the wide variety of possible textures on the clothing, the *rendering* of photorealistic appearance of clothing is also a challenge.

In this thesis, we aim to build photorealistic full-body avatars with dynamic clothing that are useful for social telepresence. In order to address the aforementioned challenges and to achieve the highest possible fidelity, we believe that the system should include the following three components, as illustrated in Fig. 1.

- **Data-driven priors.** Modern learning-based approaches provide a powerful way to build statistical models from a large amount of data. We build the avatars using deep neural networks from real-world images of clothed humans, which are captured in a multi-view system of more than one hundred cameras, and cover a variety of body poses and clothing states. The priors learned from the data allow the avatars to be animated by sparse input signals such as body poses, and perform free viewpoint rendering by interpolating between a discrete set of capture views. In Sec. 2 [Xiang et al. 2021], we explore how such data-driven priors can be applied to dynamic clothing. In particular, we demonstrate the benefit of modeling clothing as a separate layer.
- **Physics.** Physics governs the motion of dynamic clothing on top of human body, and cloth simulation can generate clothing animation with natural and rich dynamics. In Sec. 3 [Xiang et al. 2022], we investigate how physics-based cloth simulation can serve as a complement for the learning-based approach, which faces difficulties modeling complicated dynamics of loose clothing. We combine physically natural clothing animation with photorealistic appearance from neural rendering to achieve dynamically moving, photorealistic clothing.
- **Real-time sensor input.** For certain applications, we may want the rendered clothing to look not only realistic, but also match the posture and motion of the subject in the real world. This goal may require system to extract more information than just skeleton poses [Bagautdinov et al. 2021; Habermann et al. 2021a] from the sparse sensor input, e.g. one or few RGB(-D) cameras. In other words, we would like to infer clothing states and dynamics from the sparse driving views, and use the data-driven priors to fill in the missing information. Such a method is presented in Sec. 4 [Xiang et al. 2023].

2 CLOTHING CODEC AVATARS: MODELING CLOTHING AS A SEPARATE LAYER

We seek to build photorealistic full-body clothed avatars that can be animated with driving signals that can be easily accessed, for example, 3D body pose and facial keypoints. Despite the progress in previous work [Bagautdinov et al. 2021], challenges still remain, and we identify the modeling of clothing as one major difficulty. Artifacts include the imperfect correlation between body pose and

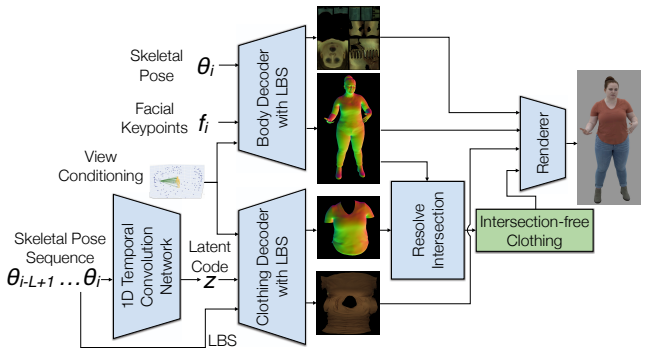

Figure 2: The clothed body animation pipeline of Clothing Codec Avatars.

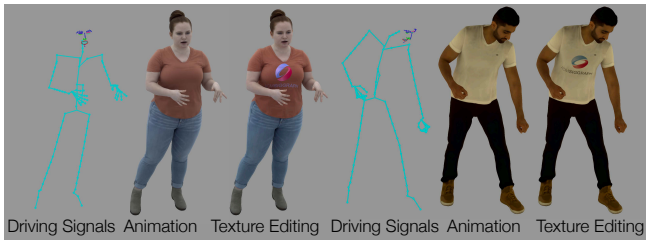

Figure 3: Results of Clothing Codec Avatars. From left to right, we show driving signals, animation output and editing results for two subjects.

clothing state, ghosting effects along the boundary between clothing and skin, as well as loss of wrinkle details and dynamics in the clothing. These artifacts become more noticeable when the captured clothing is loose and the performer moves more dynamically. On the one hand, due to registration error, the network may underfit the data, making it 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. The separation leads to several benefits. First, it 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 angles 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, the animation output can be further edited by changing the clothing texture. The animation pipeline of our method is shown in Fig. 2, and the results are shown in Fig. 3. Given a sequence of skeletal poses and facial keypoints as input, our Clothing Codec Avatars can produce

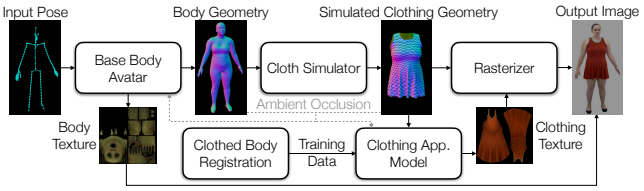

Figure 4: The pipeline of Dressing Avatars includes three major modules: the body avatar that predicts body geometry and texture given pose as input, the cloth simulator that generates clothing deformation on top of the body geometry, and the clothing appearance model that predicts photorealistic clothing 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.

photorealistic animation output with consistent clothing dynamics, and the clothing texture can be consistently edited. For more detail please refer to [Xiang et al. 2021].

3 DRESSING AVATARS: DEEP PHOTOREALISTIC APPEARANCE FOR SIMULATED CLOTHING

Existing work on avatars with animatable clothing can be categorized into two main streams. Cloth simulation creates realistic clothing deformations with dynamics [Macklin et al. 2016], 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. 2021] and may include appearance [Habermann et al. 2021b; Liu 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 the capability to animate and render avatars with novel clothing. Our work builds upon Full-Body Codec Avatars [Bagautdinov 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 in Sec. 2, 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. The overall pipeline is illustrated in Fig. 4.

Figure 5: Results of Dressing Avatars. 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.

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 degenerate tracked clothing geometry and specific appearance. To this end, we design the model to be localized in terms of both architecture (U-Net) 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. 5. Our method opens up the possibility to dress photorealistic avatars with novel garments. For more detail, please refer to [Xiang et al. 2022].

4 DRIVABLE AVATAR CLOTHING: FAITHFUL CLOTHING TELEPRESENCE DRIVEN BY SPARSE RGB-D INPUT

Clothing Codec Avatars (Sec. 2) and and Dressing Avatars (Sec. 3) can work well for pose-driven animation, i.e., synthesizing plausible clothing deformation and photorealistic appearance that are perceptually compatible with the input pose signal. However, there

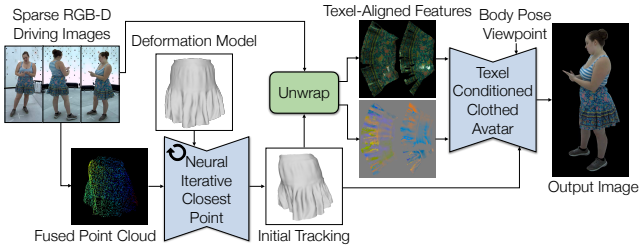

Figure 6: An overview of Drivable Avatar Clothing. First, the Neural Iterative Closest Point module efficiently tracks the clothing surface from the input point cloud with a clothing deformation model; then, the initial tracking result is used to unwrap the driving images and depth maps into texel-aligned features, which are then fed into the texel-conditioned avatar to predict the output image.

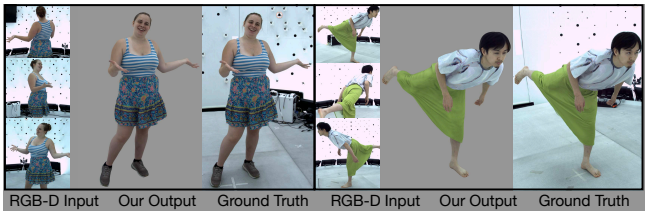

Figure 7: Results of Drivable Avatar Clothing. We show the input views, our output and the ground truth reference images in each group of results.

is no guarantee that the animation output will faithfully reproduce the actual states of clothing, and potentially distorting the conveyed social signals. An alternate approach for telepresence relies heavily on the availability of sensory inputs without a strong human prior, including those based on volumetric fusion methods [Newcombe et al. 2015], neural implicit functions [Yu et al. 2021], or neural radiance fields [Lin et al. 2022]. In theory, these methods are flexible enough to be able to reconstruct arbitrary shape from the given input streams. However, due to a lack of model constraints, it is generally more challenging for these methods to achieve high-fidelity temporal coherency especially with noisy or incomplete input, and the output quality is heavily tied with the sensory input.

To leverage the benefits of both families of approaches, we can rely on explicit avatar models as a prior, but expand the driving signal to include the denser input in addition to the body pose. We build avatars with dynamic clothing that can be driven from a sparse set of RGB-D cameras (usually three unless otherwise stated). This formulation allows for more faithful resynthesis of the human appearance, including clothing details. We build on top of DVA [Remelli et al. 2022], which proposes the texel-conditioned avatar, an encoder-decoder model that takes in UV-aligned driving features and predicts geometry and appearance for rendering. However, DVA only works well for tight clothing that closely follows the underlying body, due to the limitation in relying on a generic body shape prior.

To better handle loose clothing, our insight is to introduce a tracking stage that coarsely aligns the loose clothing surface with the input depth. More specifically, we propose a simple-yet-effective Neural Iterative Closest Points (N-ICP) algorithm to iteratively update a clothing deformation model given the feedback from surface error in a data-driven manner. N-ICP enjoys the flexibility of the classical ICP methods which allows us to handle large clothing deformations, while relying on learning for more efficient inference and reliable geometry estimates. In contrast to DVA, which uses coarse body geometry to extract features, the N-ICP tracking allows us to extract more accurate and meaningful texel-aligned features. It also eases the burden on the encoder-decoder model, since large deformations and misalignments are handled by the coarse tracking, and ultimately leads to better quality and generalization. In addition, several technical components have been leveraged to further improve the texel-conditioned avatars. To aid geometry prediction, we expand texel-aligned features with geometry features computed from depth and coarsely tracked geometry. To improve appearance, we adopt a specific perceptual loss to encourage high-frequency texture detail on the predicted clothing. An overview of this method is shown in Fig. 6, and results are shown in Fig. 7. Our method is driven by sparse RGB-D views (along with body pose and facial keypoints) and can faithfully reproduce the appearance and dynamics of challenging loose clothing from the input views. For more detail of this work, please refer to [Xiang et al. 2023].

5 CONCLUSION

In this thesis, I have presented a unified framework for modeling dynamic clothing in photorealistic avatars that involves data-driven prior, physics, and sensing. For future work, I would like to extend this framework to universal models that can handle multiple garment instances and identities. Another interesting direction is to incorporate differentiable simulation to estimate accurate physical parameters of garments, so that their synthesized motion can match the real world. Finally, I am interested in personalizing avatars from universal priors given sparse input, such as a monocular video, to make the technology accessible to general users. For more detail, please refer to the full thesis [Xiang 2023].

ACKNOWLEDGMENTS

I would like to thank my Ph.D. advisor Prof. Jessica Hodgins for her support as well as all co-authors at Meta Reality Labs Research.

REFERENCES

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).

Hugo Bertiche, Meysam Madadi, and Sergio Escalera. 2021. PBNS: physically based neural simulation for unsupervised garment pose space deformation. *ACM Transactions on Graphics (TOG)* 40, 6 (2021).

Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. 2016. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *European 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. 2021b. A Deeper Look into DeepCap. *IEEE Transactions on Pattern*

Analysis and Machine Intelligence (2021).

Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).

Haotong Lin, Sida Peng, Zhen Xu, Yunzhi Yan, Qing Shuai, Hujun Bao, and Xiaowei Zhou. 2022. Efficient Neural Radiance Fields for Interactive Free-viewpoint Video. In *SIGGRAPH Asia 2022 Conference Papers*.

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).

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

Miles Macklin, Matthias Müller, and Nuttapong 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.

Richard A Newcombe, Dieter Fox, and Steven M Seitz. 2015. Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time. In *Proceedings of the IEEE 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*.

Chenglei Wu, Takaaki Shiratori, and Yaser Sheikh. 2018. Deep incremental learning for efficient high-fidelity face tracking. *ACM Transactions on Graphics (TOG)* 37, 6 (2018).

Donglai Xiang. 2023. *Modeling Dynamic Clothing for Data-Driven Photorealistic Avatars*. Ph. D. Dissertation. Carnegie Mellon University.

Donglai Xiang, Timur Bagautdinov, Tuur Stuyck, Fabian Prada, Javier Romero, Weipeng Xu, Shunsuke Saito, Jingfan Guo, Breannan Smith, Takaaki Shiratori, et al. 2022. Dressing Avatars: Deep Photorealistic Appearance for Physically Simulated Clothing. *ACM Transactions on Graphics (TOG)* 41, 6 (2022).

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, Zhe Cao, Kaiwen Guo, Chenglei Wu, Jessica Hodgins, and Timur Bagautdinov. 2023. Drivable Avatar Clothing: Faithful Full-Body Telepresence with Dynamic Clothing Driven by Sparse RGB-D Input. In *SIGGRAPH Asia 2023 Conference Papers*.

Tao Yu, Zerong Zheng, Kaiwen Guo, Pengpeng Liu, Qionghai Dai, and Yebin Liu. 2021. Function4d: Real-time human volumetric capture from very sparse consumer rgbd sensors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5746–5756.