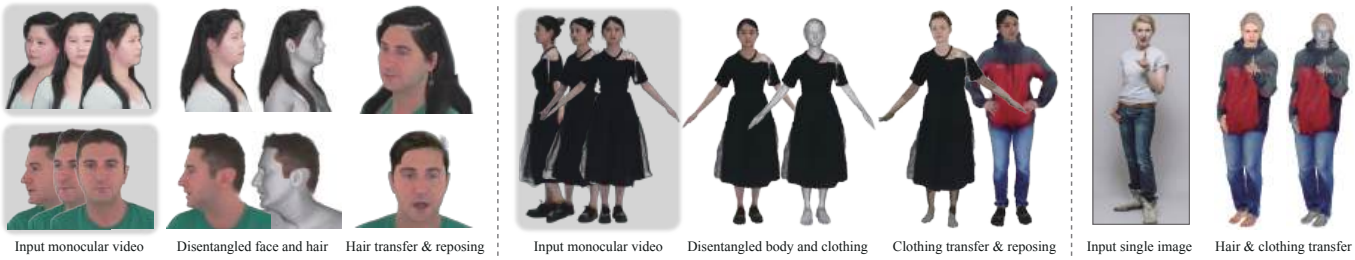
Learning Disentangled Avatars with Hybrid 3D Representations

YAO FENG, Max Planck Institute for Intelligent Systems, Germany & ETH Zürich, Switzerland
WEIYANG LIU, Max Planck Institute for Intelligent Systems, Germany & University of Cambridge, United Kingdom
TIMO BOLKART, Max Planck Institute for Intelligent Systems, Germany
JINLONG YANG, Max Planck Institute for Intelligent Systems, Germany
MARC POLLEFEYS, ETH Zürich, Switzerland
MICHAEL J. BLACK, Max Planck Institute for Intelligent Systems, Germany

Project Page: yfeng95.github.io/delta



(a) **Application I** (head): Disentangling face and hair from monocular video

(b) **Application II** (body): Disentangling body and clothing from monocular video

(c) **Application III** (head + body): Simultaneous hair and clothing transfer

Fig. 1. (a) Disentangled human head: DELTA outputs disentangled mesh-based face and NeRF-based hair given a monocular video input. (b) Disentangled human body: DELTA outputs disentangled mesh-based body and NeRF-based clothing given a monocular video input. (c) With the disentangled clothing and hair learned by DELTA, we can easily transfer any hair and clothing to a human body estimated from a single image.

Abstract: Tremendous efforts have been made to learn animatable and photorealistic human avatars. Towards this end, both explicit and implicit 3D representations are heavily studied for a holistic modeling and capture of the whole human (*e.g.*, body, clothing, face and hair), but neither representation is an optimal choice in terms of representation efficacy since different parts of the human avatar have different modeling desiderata. For example, meshes are generally not suitable for modeling clothing and hair. Motivated by this, we present Disentangled Avatars (DELTA), which models humans with hybrid explicit-implicit 3D representations. DELTA takes a monocular RGB video as input, and produces a human avatar with separate body and clothing/hair layers. Specifically, we demonstrate two important applications for DELTA. For the first one, we consider the disentanglement of the human body and clothing and in the second, we disentangle the face and hair. To do so, DELTA represents the body or face with an explicit mesh-based parametric 3D model and the clothing or hair with an implicit neural radiance field. To make this possible, we design an end-to-end differentiable renderer that integrates meshes into volumetric rendering, enabling DELTA to learn directly from monocular videos without any 3D supervision. Finally, we show that how these two applications can be easily combined to model full-body avatars, such that the hair, face, body and clothing can be fully disentangled yet jointly rendered. Such a disentanglement enables hair and clothing transfer to arbitrary body shapes. We empirically validate the effectiveness of DELTA's disentanglement by demonstrating its promising performance on disentangled reconstruction, virtual clothing try-on and hairstyle transfer. To facilitate future research, we also release an open-sourced pipeline for the study of hybrid human avatar modeling.

Technical Report, September 2023

1 INTRODUCTION

Recent years have witnessed an unparalleled surge in the utilization of 3D human reconstruction and reenactment in numerous applications such as virtual and augmented reality, telepresence, games, and movies. It is of broad interest to create personal avatars from readily available setups (*e.g.*, monocular videos). It is desirable in practice for the avatars to be photorealistic, 3D-consistent, animatable, easily editable and generalizable to novel poses. These characteristics call for a faithful disentanglement and modeling of different semantic components of the avatar (*e.g.*, face and hair for head, body and clothing for whole body). Therefore, how to disentangle human avatars while yielding accurate reconstructions is of great significance and remains an open challenge.

Existing methods for learning 3D human avatars can be roughly categorized into *explicit* ones and *implicit* ones. Explicit methods (*e.g.*, [Feng et al. 2021b; Grassal et al. 2022; Khakhulin et al. 2022; Sanyal et al. 2019] for head, [Choutas et al. 2020; Feng et al. 2021a; Kanazawa et al. 2018; Kolotouros et al. 2019; Pavlakos et al. 2019; Zanfir et al. 2021] for body) typically use triangular meshes as representation, and the reconstruction heavily relies on statistical shape priors, such as 3D morphable models for head [Blanz and Vetter 1999; Egger et al. 2020; Li et al. 2017] and 3D parametric models for body [Anguelov et al. 2005; Joo et al. 2018; Loper et al. 2015; Osman et al. 2020; Pavlakos et al. 2019; Xu et al. 2020]. Implicit methods usually encode the 3D geometry either with implicit surfaces (*e.g.*, signed distance fields (SDF)) [Jiang et al. 2022; Saito et al.

2019; Zheng et al. 2022] or with volumetric representation [Gafni et al. 2021; Gao et al. 2022; Peng et al. 2021b]. Both explicit and implicit methods use a single 3D representation to model different parts of the avatar, which ignores the representation efficacy and therefore can be sub-optimal. For example, triangular meshes are an efficient representation for faces and minimally clothed body, for which statistical template priors are available, but meshes are generally a poor representation for hair or clothing since they can be inefficient to capture the underlying geometry. On the other hand, implicit representation renders high-fidelity 2D views but it is nontrivial to animate and usually can not generalize to unseen poses and expressions. Since no single 3D representation is perfect, *why not use different one for different part of the avatar?* Motivated by this, we propose **DisEntangLed avaTAr** (DELTA), which models face and body with explicit triangular meshes, and models hair and clothing with an implicit neural radiance field (NeRF) [Mildenhall et al. 2020]. The intuition behind such a design is in two folds. First, both faces and bodies have regular topological structures and live in a low-dimensional subspace [Basri and Jacobs 2003; Li et al. 2009]. It is therefore a well-motivated choice to represent the face or body geometry with mesh templates. Second, hair consists of countless freely deformed thin strands, which hinders triangular meshes to be a suitable representation. Clothing (*e.g.*, dresses) also consists of complex topological structures and has a diverse set of styles. Due to the complex nature of hair and clothing, it is highly difficult to accurately model their surface geometry, which renders NeRF an arguably better choice of representation.

The effectiveness of hybrid 3D representation has already found its traces in human-scene reconstruction [Pavlakos et al. 2022], clothed body modeling [Feng et al. 2022], and human eye modeling [Li et al. 2022]. For example, [Pavlakos et al. 2022] reconstructs the static scene with a NeRF which excels at representing fine-grained scene details, and the people inside with a SMPL [Loper et al. 2015] representation which is good at body pose recovery. Despite modeling different subjects under different context, the essence of hybrid representation is the adoption of heterogeneous 3D representations such that each representation can be made the best use of. Extending our prior work [Feng et al. 2022], DELTA is the *first* method to demonstrate the power of hybrid representation for learning human avatars (including face, body, hair and clothing). Specifically, we instantiate the idea of DELTA in two capture settings. First, we consider the disentangled reconstruction of human head where the head (and upper shoulder) is represented by a parametric mesh model (*i.e.*, FLAME [Li et al. 2017] and SMPL-X [Pavlakos et al. 2019]) and the hair is represented by a NeRF. Unlike existing works [Gafni et al. 2021; Grassal et al. 2022; Zheng et al. 2022], DELTA additionally reconstruct the upper body (*e.g.*, shoulder), such that people with long hair can be better captured. Second, we consider the disentangled reconstruction of human body where the body is represented by a parametric mesh model (*i.e.*, SMPL-X) and the clothing is represented by a NeRF. Combining the disentangled capture of both human head and body, we demonstrate that both hair and clothing can be simultaneously transferred to arbitrary reconstructed human body. See Figure 1 for an illustration.

Distinct from existing work [Li et al. 2022; Pavlakos et al. 2022], at the very heart of DELTA is our novel mesh-integrated volumetric renderer, which not only drives the disentanglement of different parts of the avatar (*i.e.*, face, hair, body, clothing), but also enables the end-to-end differentiable learning directly from monocular videos without any 3D supervision. We expect the idea of hybrid 3D representation to be quite general, and DELTA aims to demonstrate the power of hybrid 3D representation by bringing together meshes and NeRFs in modeling human avatars.

Why is disentanglement so important for learning avatars? We answer this question by listing some key desiderata for photorealistic avatar creation. First, the pose-dependent factors should be disentangled from the appearance such that the captured avatar can be easily reusable in new environments. Second, disentangling the human body, hair, and clothing is crucial to accurately model their respective dynamics, since the motion dynamics of the human body, hair, and clothing are completely distinct from each other. Moreover, modeling the interaction between body and hair/clothing also requires an accurate disentanglement. Such a disentanglement becomes even more important when performing physical simulation on the reconstructed avatar. Third, human body, hair and clothing have totally different material and physical properties, which results in different lighting phenomena. In order to construct realistic and generalizable avatars, human body and hair/clothing have to be disentangled and modeled separately. Towards the goal of learning disentangled avatars, our contributions are listed below:

- By substantially extending our previous work [Feng et al. 2022], we propose the disentangled avatar that models face/body and hair/clothing with a hybrid 3D representation. Such an hybrid representation marries the statistical prior from mesh surfaces and the representation flexibility from implicit functions. DELTA is one of the first methods that uses a hybrid explicit-implicit representation to reconstruct high-fidelity disentangled avatars.
- We design a novel differentiable volumetric rendering method that incorporates meshes into volumetric rendering.
- The framework of DELTA is fully differentiable and end-to-end trainable. It is trained on a monocular video (*e.g.*, from web cameras) without requiring any 3D supervision.
- For the face and body, DELTA delivers high-fidelity details while being able to effortlessly reposed. For the hair and clothing region, DELTA yields realistic hair and clothing reconstruction owing to the powerful implicit NeRF representation.
- We emphasize that the major contribution of DELTA is to serve as a demonstration to showcase the potentials of hybrid 3D representation in modeling human avatars.

2 RELATED WORK

2.1 Head Avatar Creation

Explicit head avatars. Explicit head avatars are typically based on explicit 3D representations (*e.g.*, triangular meshes). 3D morphable models (3DMM) [Blanz and Vetter 1999], which are obtained from a population of 3D head scans [Egger et al. 2020], are widely used as a stronger statistical prior to represent the geometry of faces. Built upon 3DMM, many improved variants have been proposed, including multi-linear models for shape and expression [Cao et al. 2013; Vlasic et al. 2006], full-head models [Dai et al. 2020; Li et al.

2017; Ploumpis et al. 2020], and deep nonlinear models [Ranjan et al. 2018; Tran and Liu 2018]. Besides, morphable models also provide a linear model for textures [Aldrian and Smith 2010; Blanz and Vetter 1999, 2003; Paysan et al. 2009]. 3DMM and its variants can be used to reconstruct faces through an optimization procedure [Gecer et al. 2019; Romdhani and Vetter 2005; Schönborn et al. 2017; Thies et al. 2016] or learning-based estimation [Deng et al. 2019; Dib et al. 2021; Feng et al. 2021b; Khakhulin et al. 2022; Lattas et al. 2020; Li et al. 2018; Sanyal et al. 2019; Shang et al. 2020; Tewari et al. 2019, 2018, 2017; Wen et al. 2021]. Besides 3DMM template priors, other priors (*e.g.*, symmetry [Liu et al. 2022b; Wu et al. 2020], causality [Liu et al. 2022b; Wen et al. 2021], identity [Cole et al. 2017; Feng et al. 2021b]) are also considered in 3D face reconstruction. Despite producing good coarse facial geometry, these methods are usually unable to reconstruct fine-grained facial details and the entire head (*e.g.*, hair). Some methods [Alldieck et al. 2018a; Cao et al. 2015; Feng et al. 2021b] use mesh displacements to reconstruct fine details such as wrinkles, producing fine-grained geometry. Following a similar spirit, Grassal et al. [2022] use a geometry refinement network that learns a pose-dependent offset function for geometry corrections, and produces photorealistic outputs under novel views. PointAvatar [Zheng et al. 2023b] uses a deformable point-based representation to reconstruct human heads from videos. Unlike previous work, DELTA captures the head avatar with disentangled face and hair components. DELTA adopts the explicit mesh-based representation to model the face region, making it easily animatable. For the hair, we utilize an implicit NeRF-based representation, capable of accommodating various hair types. With this approach, we can utilize models tailored for faces and hair, and it also unlocks potential applications like hairstyle transfer.

Implicit head avatars. Implicit models normally encode the 3D head avatar with NeRF-based representation [Mildenhall et al. 2020; Müller et al. 2022] or implicit surface functions [Chen and Zhang 2019; Kellnhofer et al. 2021; Mescheder et al. 2019; Park et al. 2019; Yariv et al. 2020]. NeRF-based methods have been explored for 3D face modeling from images or videos [Chan et al. 2021; Gafni et al. 2021; Park et al. 2021; Wang et al. 2021]. Gafni et al. [2021] reconstruct an animatable NeRF from a single monocular video, which is conditioned on the expression code from a 3DMM. Gao et al. [2022] propose a NeRF-based linear blending representation where expression is encoded by multi-level voxel fields. AvatarMAV [Xu et al. 2023a] uses neural voxel fields to represent motion and appearance to achieve fast head reconstruction. LatentAvatar [Xu et al. 2023b] reconstructs a NeRF-based head avatar that is driven by latent expression codes, and these expression codes are learned in an end-to-end and self-supervised manner without the tracking of templates. However, NeRF-based head representations generally suffer from poor 3D geometry and struggles to generalize to unseen poses/expressions. Approaches utilizing implicit surface functions generally provide better geometry for faces. Yenamandra et al. [2021] proposes an implicit morphable face model that disentangles texture and geometry. Zheng et al. [2022] parameterize the head with implicit surface functions in the canonical space, and represents the expression- and pose-dependent deformations via learned blendshapes and skinning fields. Ramon et al. [2021] use an optimization-based approach to estimate the signed distance function (SDF) of a full head from a few images, and this optimization is constrained by a pre-trained 3D head SDF model. In contrast to both explicit and implicit head avatars that use a holistic 3D representation, DELTA is the first method that adopts a hybrid explicit-implicit 3D representation to separately model face and hair. DELTA marries the strong controllability of the mesh-based face and the high-fidelity rendering of the NeRF-based hair.

2.2 Full Body Avatar Creation

Explicit Body Avatars. The 3D surface of a human body is typically represented by a learned statistical 3D model using an explicit mesh representation [Anguelov et al. 2005; Joo et al. 2018; Loper et al. 2015; Osman et al. 2020; Pavlakos et al. 2019]. The parametric models [Loper et al. 2015; Pavlakos et al. 2019] can produce a minimal clothed body when the shape parameters are provided. Numerous optimization and regression methods have been proposed to compute 3D shape and pose parameters from images, videos, and scans. See [Liu et al. 2022a; Tian et al. 2022] for recent surveys. We focus on methods that capture full-body pose and shape, including the hands and facial expressions [Choutas et al. 2020; Feng et al. 2021a; Pavlakos et al. 2019; Rong et al. 2021; Xiang et al. 2019; Xu et al. 2020; Zhou et al. 2021]. Such methods, however, do not capture hair, clothing, or anything that deviates the body. Also, they rarely recover texture information, due to the large geometric discrepancy between the clothed human in the image and captured minimal clothed body mesh. Some methods choose to model body along with clothing. However, clothing is more complex than the body in terms of geometry, non-rigid deformation, and appearance, making the capture of clothing from images challenging. Explicit ways to capture clothing often use additional vertex offsets relative to the body mesh [Alldieck et al. 2019a, 2018a,b, 2019b; Jin et al. 2020; Lazova et al. 2019; Ma et al. 2020; Xiu et al. 2023]. While such an approach generally works well for tight clothing, it still struggles to capture loose clothing like skirts and dresses.

Implicit Body Avatars. Recently, implicit representations have gained traction in modeling the human body [Alldieck et al. 2021; Xu et al. 2020]. Correspondingly, methods have been developed to estimate implicit body shape from images [Xu et al. 2020]. However, similar to explicit body model [Pavlakos et al. 2019], they only model minimal clothed body. When it comes to clothed avatars, recent methods are leveraging implicit representations to handle more complex variations in clothing styles, aiding in the recovery of clothing structures. For instance, [He et al. 2021; Huang et al. 2020; Saito et al. 2019, 2020; Xiu et al. 2022; Zheng et al. 2021] extract pixel-aligned spatial features from images and map them to an implicit shape representation. To animate the captured non-parametric clothed humans, Yang et al. [2021] predict skeleton and skinning weights from images to drive the representation. Corona et al. [2021] represent clothing layers with deep unsigned distance functions [Chibane et al. 2020], and learn the clothing style and clothing cut space with an auto-decoder. Once trained, the clothing latent code can be optimized to match image observations, but it produces oversmooth results without detailed wrinkles. PoseVocab [Li et al. 2023b] models NeRF-based human avatars by learning pose encoding. Although such implicit models can capture various clothing styles

much better than explicit mesh-based approaches, faces and hands are usually poorly recovered due to the lack of a strong prior on the human body. In addition, such approaches typically require a large set of manually cleaned 3D scans as training data. Recently, various methods recover 3D clothed humans directly from multi-view or monocular RGB videos [Chen et al. 2021b; Jiang et al. 2022; Liu et al. 2021; Peng et al. 2021a, 2022, 2021b; Qiu et al. 2023; Su et al. 2021; Weng et al. 2022]. They optimize avatars from image information using implicit shape rendering [Liu et al. 2020; Niemeyer et al. 2020; Yariv et al. 2021, 2020] or volume rendering [Mildenhall et al. 2020], no 3D scans are needed. Although these approaches demonstrate impressive performance, hand gestures and facial expressions are difficult to capture and animate due to the lack of model expressiveness and controllability. AvatarReX [Zheng et al. 2023c] learns a NeRF-based full-body avatar with disentangled modeling of face, body and hands, but the clothing is still entangled with body.

Unlike prior methods, we view clothing as a separate layer above the body and combine explicit body models and implicit clothing to leverage the advantages of both. The mesh-based body model allows us to create human shapes with detailed components (*e.g.*, hands) and to control the body (*e.g.*, expressions and hand articulations). With implicit representation, we can capture a variety of clothing using images, without the need for 3D scans. Moreover, the disentangled modeling of explicit body and implicit clothing facilitates seamless clothing transfer, enabling applications like virtual try-ons.

2.3 Other Related Work

Hybrid 3D representation. The potentials of hybrid 3D representation have also been demonstrated in other 3D reconstruction tasks. Pavlakos et al. [2022] represent the background static scene as a NeRF and the people inside as SMPL models. Li et al. [2022] model the eye-ball surface with an explicit parametric surface model and represents the periocular region and the interior of the eye with deformable volumetric representations. Hybrid explicit-implicit representation has also been explored in transparent object reconstruction [Xu et al. 2022] and haptic rendering [Kim et al. 2004].

Hair modeling. How to represent hair is a long-standing problem in human modeling [Ward et al. 2007]. Strand-based modeling is widely adopted to model human hair [Beeler et al. 2012; Chai et al. 2013, 2012; Herrera et al. 2012; Hu et al. 2014; Luo et al. 2012, 2013; Nam et al. 2019; Rosu et al. 2022; Sun et al. 2021; Yang et al. 2019; Zhang et al. 2017; Zhang and Zheng 2019; Zhou et al. 2018]. Zheng et al. [2023a] recover the strand-based 3D hair from an intermediate representation that consists of a strand map and a depth map. Neural Haircut [Sklyarova et al. 2023] uses a two-stage coarse-to-fine optimization to reconstruct the strand-level hair. More recently, volumetric representation is also applied to perform hair modeling [Saito et al. 2018; Wang et al. 2022]. Their primary focus is on hair reconstruction, and they typically utilize head-tracked meshes from multi-view images [Rosu et al. 2022; Wang et al. 2021, 2022] or reconstruct faces from videos with stationary heads [Sklyarova et al. 2023]. None of these methods, however, are designed to learn faces from monocular videos with dynamic facial expressions. In contrast, our approach distinguishes itself by learning both facial features and hair from monocular videos, even when the head is moving. Since the primary objective of DELTA is to disentangle the representation of faces and hair rather than accurately capturing hair geometry, we employ a NeRF representation for hair modeling. The disentangled capture of face, upper body and hair is a necessary step before one can perform high-fidelity hair modeling, so DELTA also serves as a stepping stone for future work that combines better hair modeling in creating disentangled head avatars.

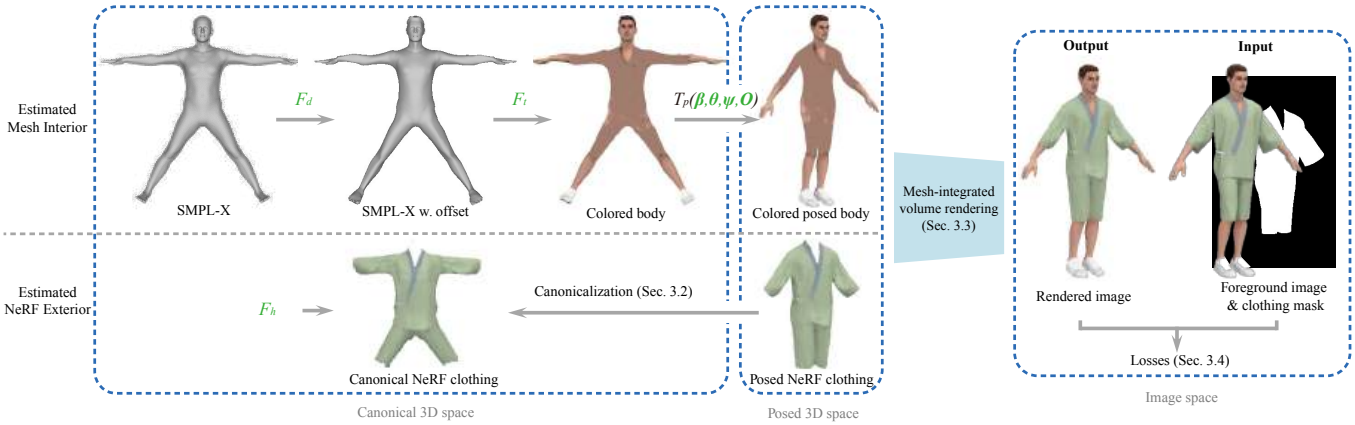
Garment reconstruction. The task of reconstructing 3D garments from images or videos has proven to be a complex challenge [Daněřek et al. 2017; Hong et al. 2021; Li et al. 2021; Qiu et al. 2023; Su et al. 2022; Zhao et al. 2021; Zhu et al. 2020]. This complexity arises from the wide diversity in clothing topologies. To tackle this, existing methods often rely on either clothing template meshes or implicit surface functions. Typically, these approaches demand access to 3D data. Many approaches employ training data produced by physics-based simulations [Bertiche et al. 2020; Patel et al. 2020; Santesteban et al. 2019; Vidaurre et al. 2020] or require template meshes fit to 3D scans [Chen et al. 2021a; Halimi et al. 2022; Pons-Moll et al. 2017; Tiwari et al. 2020; Xiang et al. 2021]. Jiang et al. [2020] train a mesh-based multi-clothing model on 3D datasets with various clothing styles. Zhu et al. [2020] introduce a adaptable template that allows for encoding clothing with diverse topologies within a single mesh template. Then during inference, a trained network produces the 3D clothing as a separate mesh-based layer by recognizing and predicting the clothing style from an image. Zhu et al. [2022] fit template meshes to non-parametric 3D reconstructions. While these methods recover garments from images, they are limited in visual fidelity, as they do not capture clothing appearance. Additionally, methods with such predefined clothing style templates can not easily handle the real clothing variations, limiting their applications. In contrast, Corona et al. [2021] represent clothing layers with deep unsigned distance functions [Chibane et al. 2020], and learn the clothing style and clothing cut space with an auto-decoder. Once trained, the clothing latent code can be optimized to match image observations, but it produces over-smooth results without detailed wrinkles. Instead, DELTA models the clothing layer with a neural radiance field, and optimizes the body and clothing layer from scratch instead of the latent space of a learned clothing model. Therefore, DELTA produces avatars with higher visual fidelity (see Section 5).

3 **DELTA**: LEARNING DISENTANGLED AVATARS

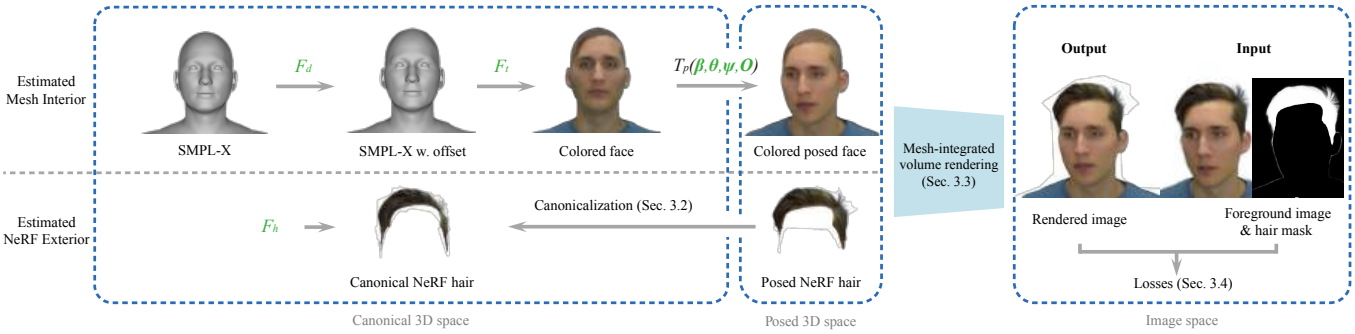
Given a monocular video, DELTA reconstructs a head (or body) avatar where head/body and hair/clothing are fully disentangled. Once the avatar is built, we can animate it with novel poses and change the hairstyle and clothing effortlessly. Because the way that DELTA reconstructs head and body shares many similarities, we simplify the description by referring the face or body as *avatar interior* and the hair or clothing as *avatar exterior*.

3.1 Hybrid Explicit-Implicit 3D Representations

Previous work on face and body modeling [Bi et al. 2021; Grassal et al. 2022; Li et al. 2017; Lombardi et al. 2018; Loper et al. 2015; Pavlakos et al. 2019] has demonstrated that both human faces and bodies can be accurately modeled by mesh-based representations. In the light of these encouraging results, we choose mesh as the representation



(a) DELTA for disentangled body and clothing reconstrcution



(b) DELTA for disentangled head and hair reconstrcution

Fig. 2. DELTA takes a monocular RGB video and clothing/hair segmentation masks as input, and outputs a human avatar with separate body and clothing/hair layers. Green letters indicate optimizable modules or parameters.

for the face and body. Specifically, we use SMPL-X [Pavlakos et al. 2019] to make full use of the human geometry priors. When it comes to representing hair and clothing, it remains an open problem which representation works the best. Because of the complex geometry of hair and clothing, we propose to model both hair and clothing with NeRF [Mildenhall et al. 2020] – a more flexible and expressive implicit representation. Distinct from meshes, NeRF is agnostic to the style, geometry and topology of hair and clothing.

Explicit avatar interior by SMPL-X. SMPL-X is an expressive body model with detailed face shape and expressions. A subject's face and body with neutral expression in the rest pose is defined as

$$T_P(\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{\psi}) = \bar{\boldsymbol{T}} + B_S(\boldsymbol{\beta};\mathcal{S}) + B_P(\boldsymbol{\theta};\mathcal{P}) + B_E(\boldsymbol{\psi};\mathcal{E}), \tag{1}$$

where $\bar{\boldsymbol{T}} \in \mathbb{R}^{n_v\times 3}$ is a template of body shape in the rest pose, $\boldsymbol{\beta} \in \mathbb{R}^{|\boldsymbol{\beta}|}$ is the body identity parameters, and $B_S(\boldsymbol{\beta};\mathcal{S}) : \mathbb{R}^{|\boldsymbol{\beta}|} \rightarrow \mathbb{R}^{n_v\times 3}$ are the identity blend shapes. More specifically, $B_S(\boldsymbol{\beta};\mathcal{S}) = \sum_{i=1}^{|\boldsymbol{\beta}|} \boldsymbol{\beta}_i\mathcal{S}_i$ where $\boldsymbol{\beta}_i$ is the i-th linear coefficient and $\mathcal{S}_i$ is the i-th orthonormal principle component. $\boldsymbol{\theta} \in \mathbb{R}^{3n_k+3}$ denotes the pose parameters, and $\boldsymbol{\psi} \in \mathbb{R}^{|\boldsymbol{\psi}|}$ denotes the facial expression parameters. Similar to the shape space $\mathcal{S}$, $B_P(\boldsymbol{\theta};\mathcal{P}) : \mathbb{R}^{|\boldsymbol{\theta}|} \rightarrow \mathbb{R}^{n_v\times 3}$ denotes the pose blend shapes ($\mathcal{P}$ is the pose space), and $B_E(\boldsymbol{\psi};\mathcal{E}) : \mathbb{R}^{|\boldsymbol{\psi}|} \rightarrow \mathbb{R}^{n_v\times 3}$ denotes the expression blend shapes from the SMPL-X model ($\mathcal{E}$ is the expression space). To increase the flexibility of SMPL-X, we add additional vertex offsets $\boldsymbol{O} := \{F_d(\boldsymbol{t}_1), F_d(\boldsymbol{t}_2), \cdots, F_d(\boldsymbol{t}_{n_v})\}^\top \in \mathbb{R}^{n_v\times 3}$ in the canonical space. The offset is modeled by a vertex-wise implicit function $F_d : \boldsymbol{t} \rightarrow \boldsymbol{o}$, which predicts an offset $\boldsymbol{o} \in \mathbb{R}^3$ for the vertex $\boldsymbol{t} \in \mathbb{R}^3$ in the rest template. Therefore, we augment the body shape with the following set of offsets:

$$\tilde{T}_P(\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{\psi},\boldsymbol{O}) = T_P(\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{\psi}) + \boldsymbol{O}. \tag{2}$$

The albedo is represented by an implicit function $F_t : \boldsymbol{t} \rightarrow \boldsymbol{c}^{\text{mesh}}$ which predicts the RGB color $\boldsymbol{c}^{\text{mesh}}$ of each given vertex $\boldsymbol{t}$ on the surface. Specifically, we sample vertex $\boldsymbol{t}$ from the template mesh $\bar{\boldsymbol{T}}$ if the video is under uniform lighting. For more complex lighting conditions, in order to better model the texture, we sample $\boldsymbol{t}$ from the surface after the pose deformation. More details can be found in Section 5.2. To capture more geometric details, we use an upsampled version of SMPL-X with $n_v = 38,703$ vertices and $n_t = 77,336$ faces [Feng et al. 2022]. Similar to [Grassal et al. 2022], we also add additional faces inside the mouth region for head avatar modeling.

Implicit avatar exterior by NeRF. Based on NeRF [Mildenhall et al. 2020], we define the avatar exterior (hair or clothing) in the

canonical 3D space as an implicit function $F_h : \boldsymbol{x}^c \rightarrow (\boldsymbol{c}^{\text{nerf}}, \sigma)$ which can be parameterized by a multi-layer perceptron (MLP). $\boldsymbol{c}^{\text{nerf}}$ represents the RGB color. Given a query point $\boldsymbol{x}^c \in \mathbb{R}^3$ in the canonical space, the implicit NeRF-based function F_h outputs an emitted RGB color $\boldsymbol{c}^{\text{nerf}}$ and a volume density σ.

3.2 Pose-dependent Deformation

Explicit avatar interior deformation. Given the monocular video, we need to model the movement of this subject. Since our avatar interior model is based on SMPL-X, it provides a good way to capture the pose deformation and facial expressions. For each frame of given video, we estimate the parameters of shape $\boldsymbol{\theta} \in \mathbb{R}^{|\boldsymbol{\theta}|}$ and expression $\boldsymbol{\psi} \in \mathbb{R}^{|\boldsymbol{\psi}|}$. Then we can deform the head/body to the observation pose using the linear blend skinning function (*i.e.*, LBS). The deformation for the explicit SMPL-X mesh model is modeled by a differential function $M(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\psi}, \boldsymbol{O})$ that outputs a 3D human body mesh $(\boldsymbol{V}, \boldsymbol{F})$ where $\boldsymbol{V} \in \mathbb{R}^{n_v \times 3}$ is a set of n_v vertices and $\boldsymbol{F} \in \mathbb{R}^{n_t \times 3}$ is a set of n_t faces with a fixed topology:

$$M(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\psi}, \boldsymbol{O}) = \text{LBS}(\tilde{T}_P(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\psi}, \boldsymbol{O}), J(\boldsymbol{\beta}), \boldsymbol{\theta}, \boldsymbol{W}), \tag{3}$$

in which $\boldsymbol{W} \in \mathbb{R}^{n_k \times n_v}$ is the blend skinning weights used in the LBS function. $J(\boldsymbol{\beta}) \in \mathbb{R}^{n_k \times 3}$ is a function of body shape [Pavlakos et al. 2019], representing the shape-dependent joints. Given a template vertex $\boldsymbol{t}_i$, the vertex $\boldsymbol{v}_i$ can be computed with simple linear transformation. Specifically, the forward vertex-wise deformation can be written as the following equation in the homogeneous coordinates:

$$\underbrace{\boldsymbol{v}_i}_{\text{Posed vertex}} = \underbrace{\sum_{k=1}^{n_k} \boldsymbol{W}_{k,i} G_k(\boldsymbol{\theta}, J(\boldsymbol{\beta})) \cdot \begin{bmatrix} \boldsymbol{I} & \boldsymbol{o}_i + \boldsymbol{b}_i \\ \boldsymbol{0} & 1 \end{bmatrix}}_{M_i(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\psi}, \boldsymbol{O})\text{: Deformation to the posed space}} \cdot \underbrace{\boldsymbol{t}_i}_{\text{Template vertex}},$$

where $M_i(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\psi}, \boldsymbol{O}) \in \mathbb{R}^{4\times4}$ is the deformation function of template vertex $\boldsymbol{t}_i$. $\boldsymbol{W}_{k,i}$ is the (k, i)-th element of the blend weight matrix $\boldsymbol{W}$, $G_k(\boldsymbol{\theta}, J(\boldsymbol{\beta})) \in \mathbb{R}^{4\times4}$ is the world transformation of the k-th joint and $\boldsymbol{b}_i$ is the i-th vertex of the sum of all blend shapes $\boldsymbol{B} := B_S(\boldsymbol{\beta}) + B_P(\boldsymbol{\theta}) + B_E(\boldsymbol{\psi})$. We denote $\boldsymbol{V}$ as the vertex set of the posed avatar ($\boldsymbol{v}_i \in \boldsymbol{V}$). Both $\boldsymbol{v}_i$ and $\boldsymbol{t}_i$ are the homogeneous coordinates when applying this deformation function.

Implicit avatar exterior deformation. Aiming to learn the NeRF-based clothing/hair representation in the canonical space, we need to deform from the posed space to the canonical space. Therefore, we perform backward deformation on the top of the explicit body skinning. Given a query point $\boldsymbol{x}^p$ in the posed space (from the observed video frame), we first find the nearest k points on the body surface M. Then we use the weighted backward skinning function to transform the posed point $\boldsymbol{x}^p$ to the canonical space (*i.e.*, x^c). To model more accurate clothing/hair movement and deformation, we further learn a pose-dependent deformation function $F_e : (\boldsymbol{x}^c, \boldsymbol{v}^p_{n(\boldsymbol{x}^p)}) \in \mathbb{R}^6 \rightarrow \Delta\boldsymbol{x}^c \in \mathbb{R}^3$, where $\boldsymbol{x}^p$ denotes a point in observation space and $n(\boldsymbol{x}^p)$ is the set of indices of the nearest points to $\boldsymbol{x}^p$ in $\boldsymbol{V}^p$ which denotes the posed body meshes in $M(\boldsymbol{0}, \boldsymbol{\theta}, \boldsymbol{0}, \boldsymbol{0})$. F_e aims to predict the detailed non-rigid deformation for the query point in the canonical space. Then the residual $\Delta\boldsymbol{x}^c$ is added back to $\boldsymbol{x}^c$, and the displaced point $\tilde{\boldsymbol{x}}^c = \boldsymbol{x}^c + \Delta\boldsymbol{x}^c$ is fed to the canonical NeRF model F_h in order to compensate the exterior clothing/hair deformation in the observation space. Specifically, we have the inverse blend skinning mapping from the observation space to the posed space as the following transformation:

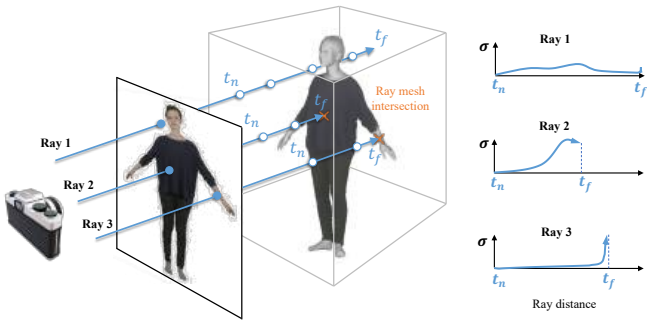


Fig. 3. Illustration of mesh-integrated volume rendering.

$$\underbrace{\boldsymbol{x}^c}_{\substack{\text{Canonical} \\ \text{vertex}}} = \underbrace{\sum_{\boldsymbol{v}_i \in n(\boldsymbol{x}^p)} \alpha_i(\boldsymbol{x}^p) \cdot M_i(\boldsymbol{0}, \boldsymbol{\theta}, \boldsymbol{0}, \boldsymbol{0}) \cdot M_i^{-1}(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\psi}, \boldsymbol{O}) \cdot \underbrace{\boldsymbol{x}^p}_{\substack{\text{Observed} \\ \text{vertex}}}}_{\text{Transformation to the canonical space}},$$

where α_i is the parameter that balances the importance:

$$\alpha_i(\boldsymbol{x}^p) = \frac{1}{Z} \exp\left(-\frac{1}{2\sigma^2} \cdot \|\boldsymbol{x}^p - \boldsymbol{v}_i\| \cdot \|\boldsymbol{w}_{nn(\boldsymbol{x}^p)} - \boldsymbol{w}_i\|\right).$$

Where $Z := \sum_{\boldsymbol{v}_i \in n(\boldsymbol{x}^p)} \alpha_i(\boldsymbol{x}^p)$ is a normalizing coefficient, $\boldsymbol{w}_i \in \mathbb{R}^{n_k}$ is the blend weights of $\boldsymbol{v}_i$, σ is a constant and $nn(\boldsymbol{x}^p)$ denotes the index of the nearest point of $\boldsymbol{x}^p$ in $\boldsymbol{V}^p$.

3.3 Mesh-integrated Volume Rendering

Camera model. We simplify the problem by using a scaled orthographic camera model $\mathbf{p} = \{s, \boldsymbol{t}^\top\}^\top$ where $s \in \mathbb{R}$ is the isotropic scale and $\boldsymbol{t} \in \mathbb{R}^2$ denotes the translation.

Mesh rasterization. With the geometry parameters $(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\psi})$, the vertex offsets $\boldsymbol{O}$, the RGB color $\boldsymbol{c}^{\text{mesh}}$ of vertices in the upsampled SMPL-X template and the camera parameters $\mathbf{p}$, we render the colored mesh into an image with $\mathcal{R}_m(M(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\psi}, F_d), \boldsymbol{c}^{\text{mesh}}, \mathbf{p})$ where $\mathcal{R}_m$ denotes the differentiable rasterizer function.

Mesh-integrated volume rendering. Finally we discuss how to take mesh into consideration while performing volumetric rendering. The basic idea is that the camera ray will stop when it intersects with the mesh in the 3D space. Given a camera ray $\boldsymbol{r}(t) = \boldsymbol{q} + t\boldsymbol{d}$ with center $\boldsymbol{q} \in \mathbb{R}^3$ and direction $\boldsymbol{d} \in \mathbb{R}^3$. The rendering interval is $t \in [t_n, t_f] \subset \mathbb{R}$ (near and far bounds). Unlike previous work, we integrate the body model, $M(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\psi}, O)$, into the volumetric rendering. Specifically, if $\boldsymbol{r}(t)$ intersects M, we set the t_f such that $\boldsymbol{r}(t_f)$ is the intersection point with M. In this case, we use the mesh color instead of the NeRF color $\boldsymbol{c}^{\text{nerf}}(\boldsymbol{r}(t_f))$ (see Figure 3). More formally, the expected color of the camera ray r is defined as

$$\boldsymbol{c}(\boldsymbol{r}) = \int_{t_n}^{t_f} \boldsymbol{c}^{\text{nerf}}(\boldsymbol{r}(t)) \cdot T(t) \cdot \sigma(\boldsymbol{r}(t)) + \mathbb{1}_{\text{s}}(\boldsymbol{r}) \cdot \delta(t - t_f) \cdot \boldsymbol{c}^{\text{mesh}} dt,$$

where $\mathbb{1}_{\text{s}}(\boldsymbol{r})$ is the indicator function for whether the ray intersects the mesh surface (1 if true, 0 otherwise), $\delta(\cdot)$ denotes the Dirac delta function and $T(t) = \exp(-\int_{t_n}^{t} \sigma(\boldsymbol{r}(s))ds)$. When $\mathbb{1}_{\text{s}}(\boldsymbol{r})$ is true, we

set the t_f such that $\boldsymbol{t}(t_f)$ is the intersection point with the SMPL-X mesh M. $\boldsymbol{c}^{\text{mesh}}$ is the vertex color of the intersected mesh. We approximate the integral with evenly split n_b bins in practice:

$$\boldsymbol{c}(\boldsymbol{r}) = \Big(1 - \sum_{k=1}^{n_b-1} T_k\big(1-\exp(-\sigma_k\Delta_k)\big)\Big)\cdot\Big((1-\mathbb{1}_{\text{s}}(\boldsymbol{r}))\boldsymbol{c}^{\text{nerf}}(\boldsymbol{r}^c_{n_b}) + \mathbb{1}_{\text{s}}(\boldsymbol{r})\cdot\boldsymbol{c}^{\text{mesh}}(\boldsymbol{r}_{n_b})\Big) + \sum_{j=1}^{n_b-1} T_j\big(1-\exp(-\sigma_j\Delta_j)\big)\boldsymbol{c}^{\text{nerf}}(\boldsymbol{r}^c_j),$$

where we define $T_j = \exp(-\sum_{q=1}^{j-1}\sigma_j\Delta_j)$. $\boldsymbol{r}_j$ is sampled from the j-th bin along the camera ray $\boldsymbol{r}$. $\boldsymbol{r}^c_i$ is the corresponding canonical point for the observed point $\boldsymbol{r}_i$.

3.4 Objective Function

Overall objective function. Given a sequence of n_f images, I_f ($1 \le f \le n_f$), we optimize $\boldsymbol{\beta}$ and the weights of the MLPs F_d, F_h, F_t, F_e jointly across the entire sequence, and $\boldsymbol{\theta}_f$ and $\mathbf{p}_f$ per frame. We use the following overall objective function:

$$\mathcal{L} = \mathcal{L}_{\text{recon}} + \mathcal{L}_{\text{ext}} + \mathcal{L}_{\text{int}} + \mathcal{L}_{\text{reg}}, \tag{4}$$

with reconstruction loss $\mathcal{L}_{\text{recon}}$, avatar exterior loss $\mathcal{L}_{\text{ext}}$, avatar interior loss $\mathcal{L}_{\text{int}}$ ($\mathcal{L}^{\text{body}}_{\text{int}}$ or $\mathcal{L}^{\text{face}}_{\text{int}}$) and regularization $\mathcal{L}_{\text{reg}}$. For simplicity, we omit the frame index f and the optimization arguments whenever there is no ambiguity. For videos, the final objective function is the average over all frames.

Reconstruction loss. We minimize the difference between the rendered image and the input image with the following objective:

$$\mathcal{L}_{\text{recon}} = \lambda_{\text{pixel}}\cdot\mathcal{L}_\delta(\mathcal{R}_v - I) + \lambda_{\text{semantic}}\cdot\mathcal{L}_{\text{semantic}}(\mathcal{R}_v, I), \tag{5}$$

where $\mathcal{L}_\delta$ is the Huber loss [Huber 1964] that penalizes the pixel-level difference. $\mathcal{L}_{\text{semantic}}$ is used to regularize the semantic difference. More specifically, we use an ID-MRF loss [Wang et al. 2018] $\mathcal{L}_{\text{mrf}}$ as $\mathcal{L}_{\text{semantic}}$ for reconstructing the body avatar, and an perceptual loss [Johnson et al. 2016] $\mathcal{L}_{\text{per}}$ as $\mathcal{L}_{\text{semantic}}$ for reconstructing the head avatar. While the Huber loss focuses on the overall reconstruction, the semantic loss allows us to reconstruct more details as previously shown by Feng et al. [2021b].

Avatar exterior loss Only minimizing the reconstruction error $\mathcal{L}_{\text{recon}}$ results in a NeRF that models the entire avatar including the body/face regions. Our goal is to only capture exterior components such as clothing or hair using F_h. To achieve this, we employ a segmentation mask to explicitly limit the space within which the NeRF density can be. Given a segmentation mask S_e, which is represented by **1** for every exterior pixel (clothing or hair) and **0** elsewhere, we minimize the following exterior loss:

$$L_{\text{ext}} = \lambda_{\text{ext}}\left\|S_v - S_e\right\|_{1,1}, \tag{6}$$

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

$$\boldsymbol{s_v}(\boldsymbol{r}) = \sum_{k=1}^{n_b-1} T_k\big(1-\exp(-\sigma_k\Delta_k)\big). \tag{7}$$

Minimizing L_{ext} ensures that the aggregated density across rays (excluding the far bound) outside of clothing or hair is 0. Therefore, only the intended exterior region is captured by the NeRF model.

Avatar interior loss. To further disentangle the avatar interior and exterior, we need to ensure that the interior mesh model does not capture any exterior variation. To this end, we define a few additional loss functions based on prior knowledge.

First, the interior mesh should match the masked image. Given a binary mask S of the entire avatar (1 for inside, 0 elsewhere), we minimize the difference between the silhouette of the rendered body (denoted by $\mathcal{R}^s_m(M,\mathbf{p})$) and the given mask as

$$\mathcal{L}_{\text{silhouette}} = \lambda_{\text{silhouette}}\mathcal{L}_\delta(\mathcal{R}^s_m(M,\mathbf{p}) - S). \tag{8}$$

Second, the interior mesh should match visible avatar interior (*e.g.*, for reconstructing the body, the body mesh should match the visible body region). Only optimizing $\mathcal{L}_{\text{silhouette}}$ results in meshes that also fit the avatar exterior (*e.g.*, clothing or hair). This is undesired especially for loose clothing or long hair, and also leads to visible artifacts when transferring clothing between subjects. Instead, given a binary mask S_b of the visible body parts (1 for body parts, 0 elsewhere), we minimize the following part-based silhouette loss

$$\mathcal{L}_{\text{int-mask}} = \lambda_{\text{int-mask}}\mathcal{L}_\delta(S_b \odot \mathcal{R}^s_m(M,\mathbf{p}) - S_b), \tag{9}$$

and a part-based photometric loss

$$\mathcal{L}_{\text{skin}} = \lambda_{\text{skin}}\mathcal{L}_\delta(S_b \odot (\mathcal{R}_m(M,\boldsymbol{c},\mathbf{p}) - I)), \tag{10}$$

to put special emphasis on fitting visible interior parts.

Third, the interior mesh should stay within the exterior region. Specifically, the body or face should be generally covered by the clothing or hair, yielding to the following loss function:

$$\mathcal{L}_{\text{inside}} = \lambda_{\text{inside}}\mathcal{L}_\delta(ReLU(\mathcal{R}^s_m(M,\mathbf{p}) - S_c)). \tag{11}$$

Fourth, the skin color of occluded body vertices should be similar to visible skin regions. For this, we minimize the difference between the body colors in occluded regions and the average skin color as

$$\mathcal{L}_{\text{skin-inside}} = \lambda_{\text{skin-inside}}\mathcal{L}_\delta(S_c \odot (\mathcal{R}_m(M,\boldsymbol{c},\mathbf{p}) - \mathbf{C}_{\text{skin}})), \tag{12}$$

where $\mathbf{C}_{\text{skin}}$ is the average color of the visible skin regions. In practice, we encountered challenges with skin detection not performing effectively. Therefore, for body video sequences, we assume that the hands are visible and utilize these hand regions to compute the average skin color. Moreover, for face videos, we determine the skin color by computing the mean color of the cheek region.

Combining the loss functions above, we use the following $\mathcal{L}_{\text{int}}$ for reconstructing the interior avatar:

$$\mathcal{L}_{\text{int}} = \mathcal{L}_{\text{silhouette}} + \mathcal{L}_{\text{int-mask}} + \mathcal{L}_{\text{skin}} + \mathcal{L}_{\text{inside}} + \mathcal{L}_{\text{skin-inside}}. \tag{13}$$

Regularization. We regularize the reconstructed mesh surface with

$$\mathcal{L}_{\text{reg}} = \lambda_{\text{edge}}\mathcal{L}_{\text{edge}}(M) + \lambda_{\text{offset}}\left\|\boldsymbol{O}\right\|_{2,2}, \tag{14}$$

where $\mathcal{L}_{\text{edge}}$ denotes the relative edge loss [Hirshberg et al. 2012] between the optimized interior mesh with and without the applied offsets. For the offset loss, we apply different weights to the body, hand and face region. Details are given in the experiment section.

4 INTRIGUING INSIGHTS

Hybrid representation for general 3D modeling. While the proposed DELTA demonstrates the effectiveness of hybrid 3D representation for human avatar modeling, the idea of hybrid representation can be broadly useful for modeling general 3D objects and scenes, especially for objects whose components have quite different physical properties. For example, a burning candle can be represented with a mesh-based candle and a NeRF-based flame, and a hourglass can be represented with mesh-based glass and point-based sand. DELTA shows the power of hybrid 3D representation through the lens of human avatar modeling, and we expect more future efforts can be put in exploring hybrid 3D representation.

Hybrid vs. holistic 3D representation. It has been a long-standing debate regarding the optimal holistic 3D representation for shape modeling. In the existing graphics pipeline, meshes are still a *de facto* choice for holistic 3D representation due to its efficiency in storage and rendering. However, meshes can be quite limited in representing certain geometric structures, such as hair strand, fluid, smoke and complex clothing. Implicit 3D representations [Chen and Zhang 2019; Mescheder et al. 2019; Mildenhall et al. 2020; Park et al. 2019] demonstrate strong flexibility in complex shape representation, and in particular, NeRF further shows great novel view synthesis quality. However, it is difficult for NeRF to capture thin shell geometry like human body. While there is no single perfect 3D representation for all objects, why not combine the advantages of different representations and use them together? However, hybrid representation also inevitably introduces some shortcomings. First, the rendering process for hybrid representation becomes highly nontrivial and case-dependent. For example, our mesh-integrated volume rendering only works for the hybrid mesh and NeRF representation. Second, the representational heterogeneity makes subsequent learning and processing more difficult. For example, learning a generative model on hybrid representation is far more complicated than holistic representation. Moreover, editing hybrid representation will also become more challenging for designers. Third, how to choose the right 3D representations to combine is task-dependent. While DELTA uses meshes for human head and NeRFs for hair, it could be better to use a strand-based representation for hair.

5 EXPERIMENTS AND RESULTS

5.1 Datasets

DELTA offers a solution for capturing dynamic objects from monocular video. We demonstrate the effectiveness of our approach by applying it to the challenging tasks of capturing clothing and hair from videos. To evaluate our approach, we introduce two types of datasets, one for full-body and one for head capture.

Full-body datasets. To compare with other state-of-the-art methods of realistic human capturing. We evaluate DELTA on sequences from public sources: People Snapshot [Alldieck et al. 2018b], iPER [Liu et al. 2019], SelfRecon [Jiang et al. 2022]. However, none of them provide complicated clothes such as long dresses. Thus, we capture our own data MPIIS-SCARF, where we record videos of each subject wearing short and long dresses. For People Snapshot, we use the provided SMPL pose as initialization instead of running PIXIE [Feng et al. 2021a]. To be specific, we use 4 subjects ("male-3-casual", "female-3-casual", "male-4-casual", "female-4-casual") from People Snapshot [Alldieck et al. 2018b] for qualitative and quantitative evaluation. The quantitative evaluation follows the settings of Anim-NeRF [Chen et al. 2021b]. We further use 4 subjects ("subject003", "subject016", "subject022", "subject023") with outfit 1 and motion 1 from iPER [Liu et al. 2019] and 4 synthetic video data ("female outfit1", "female outfit2", "female outfit3", "male outfit1") and 1 self-captured video ("CHH female") from SelfRecon [Jiang et al. 2022] for qualitative evaluation. For MPIIS-SCARF, we use A-pose videos of subject "Yao" with six types of clothing for qualitative evaluation, those videos include loose dressing and short skirts. For each subject, we use around 100-150 images for optimization. For each frame, we run PIXIE [Feng et al. 2021a] to initialize $(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\psi})$, and camera $\mathbf{p}$. For datasets without providing silhouette masks, we compute S with [Lin et al. 2022], and [Dabhi 2022] for S_c.

Head datasets. We also evaluate DELTA on head videos from public sources. To be specific, we use video "MVI_1810" from IMAvatar [Zheng et al. 2022], "person_0000" and "person_0004" from neural head avatar [Grassal et al. 2022]. As subjects with long hair are missing, we further collected one video with long hair from the Internet, named video "b0_0" [Xiao 2022] (2:30). For each image from the video, we detect the upper body region and resize it to an image with 512x512 size. We then estimate 68 landmarks [Bulat and Tzimiropoulos 2017] and iris [Lugaresi et al. 2019], portrait matting with MODNet [Ke et al. 2022], and segment face and hair with face parsing [zllrunning 2019]. Given the estimated labels and SMPL-X model, we roughly estimate the shape and texture parameters for the subject, and camera, pose, expression and lighting (Spherical harmonic) for each frame. Subsequently, for enhanced SMPL-X shape fitting, we perform parameter optimization across all frames, where shape and texture parameters are shared across frames. These optimized parameters serve as the initialization for our model training. Nonetheless, these videos often lack backviews of the head as they predominantly focus on face-related areas. To demonstrate our method's capacity for capturing complete hairs, we also incorporate synthetic data from the AGORA dataset [Patel et al. 2021]. We select three subjects from Agora, each containing the mesh, texture, and corresponding SMPL fits. 200 images are rendered from the textured mesh for training DELTA.

5.2 Implementation Details

We choose $\sigma = 0.1$ and $|\mathcal{N}(\mathbf{x})| = 6$. For full-body video, we set $t_n = -0.6$, and $t_f = 0.6$ and weight the individual losses with $\lambda_{\text{pixel}} = 1.0$, $\lambda_{\text{semantic}} = 0.0005$, $\lambda_{\text{ext}} = 0.5$, $\lambda_{\text{silhouette}} = 0.001$, $\lambda_{\text{int-mask}} = 30$, $\lambda_{\text{skin}} = 1.0$, $\lambda_{\text{inside}} = 40$, $\lambda_{\text{skin-inside}} = 0.01$, $\lambda_{\text{edge}} = 500$, $\lambda_{\text{offset}} = 400$. For λ_{offset}, the weight ratio of body, face and hands region is $2:3:12$. Note that it is important to perform the first stage NeRF training without optimizing the non-rigid deformation model. In this stage, we also set $\lambda_{\text{semantic}} = 0$. In the second stage, the non-rigid deformation model then explains clothing deformations that cannot be explained by the body transformation. And $L_{semantic}$ helps capture more details that can not be modelled by the non-rigid deformation. The overall optimization time is around 40 hours with NVIDIA V100. In head video settings, we conducted SMPL-X fitting

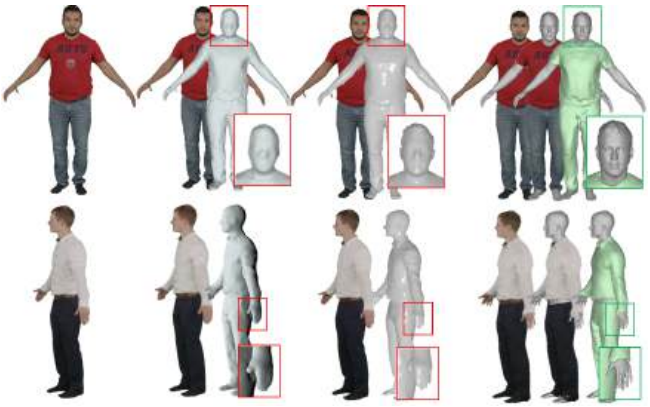


Fig. 4. Qualitative comparison with SelfRecon [Jiang et al. 2022] and Anim-NeRF [Chen et al. 2021b] for reconstruction. While all methods capture the clothing with comparable quality, our approach has much more detailed face and hands due to the disentangled representation of clothing and body.

Fig. 5. Qualitative comparison with neural head avatar (NHA) [Grassal et al. 2022] and IMavatar [Zheng et al. 2022] for reconstruction. Our method exhibits superior performance in capturing the geometry of the face and shoulders. Moreover, it achieves exceptional rendering quality for the hair. This can be attributed to the effective utilization of a disentangled representation for separating the hair and face components in DELTA.

for all frames during data processing, that ensures accurate face fitting. By employing this as our initialization for DELTA training, we can directly train both mesh-based face and NeRF-based hair components. The chosen hyperparameters include $t_n = -1.5$, and $t_f = 1.5$. We assign weights to individual losses as follows: $\lambda_{\text{pixel}} = 1.0$, $\lambda_{\text{semantic}} = 0.015$, $\lambda_{\text{ext}} = 0.5$, $\lambda_{\text{silhouette}} = 0.001$, $\lambda_{\text{int-mask}} = 30$, $\lambda_{\text{skin}} = 1.0$, $\lambda_{\text{inside}} = 40$, $\lambda_{\text{skin-inside}} = 0.001$, $\lambda_{\text{edge}} = 500$, $\lambda_{\text{offset}} = 400$. To enhance training efficiency, we adopt Instant-NGP [Li et al. 2023a; Müller et al. 2022] for parameterizing the hair component. Unlike the MLP layers in the original NeRF model, Instant-NGP leverages a hash table to store feature grids at various coarseness scales, resulting in fast training and inference speeds. We then require around 40 minutes of optimization time with NVIDIA A100.

5.3 Comparison to Existing Methods

Our approach enables the creation of hybrid explicit-implicit avatars from monocular videos. We note that this has not been achieved by previous methods, which typically model clothed bodies or heads holistically using either implicit or explicit representations. To evaluate the effectiveness of our approach, we compare it to existing state-of-the-art methods on the challenging tasks of clothed-body and head modeling. The explicit-implicit modeling of DELTA also naturally disentangles objects such as the body and clothing, thereby enabling garment reconstruction. Unlike previous methods that reconstruct cloth geometry from a single image with the help of extensive 3D scan data, our approach can reconstruct garments from images alone. We evaluate the effectiveness of DELTA for garment reconstruction by comparing it to existing methods.

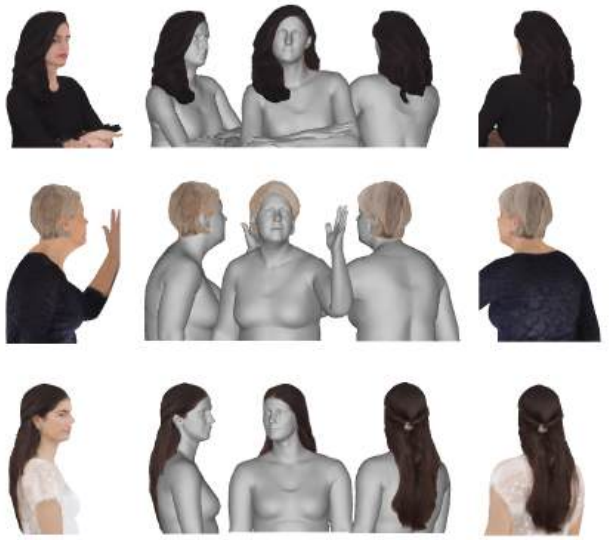


Fig. 6. Qualitative result on synthetic upper-body videos. The leftmost and rightmost images show the colored rendering of the learned avatars. The middle images show the hybrid rendering of the estimated upper body and hair. The results validate DELTA's ability to accurately represent complete hair views, including both short and long hair types.

Body and clothing modeling. We quantitatively compare NB [Omran et al. 2018], SMPLpix [Prokudin et al. 2021], Neural Body [Peng et al. 2021b] and Anim-NeRF [Chen et al. 2021b], following the evaluation protocol of [Chen et al. 2021b]. To be specfic, we use 4 subjects ("subject003", "subject016", "subject022", "subject023") with outfit 1 and motion 1 from iPER [Liu et al. 2019] for qualitative evaluation. For all subjects, we uniformly select frames 1-490 with a step-size 4 for optimization. We use 4 synthetic video data ("female outfit1", "female outfit2", "female outfit3", "male outfit1") and 1 self-captured video ("CHH female") from SelfRecon [Jiang et al. 2022]. For each subject, we use 100 frames for optimization. For self-captured data, we use A-pose videos of subject "Yao" with six types of clothing for qualitative evaluation, those videos include loose dressing and short skirts. For each video, we uniformly select frames 0-400 with a step-size 2 for optimization. Table 1 shows that DELTA is more accurate than the other methods under most metrics. The qualitative comparison in Figure 4 demonstrates that DELTA can better reconstruct the hand and face geometry compared to SelfRecon [Jiang et al. 2022] and Anim-NeRF [Chen et al. 2021b].

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

Table 1. Quantitative comparison of novel view synthesis on People-Snapshot [Alldieck et al. 2018b].

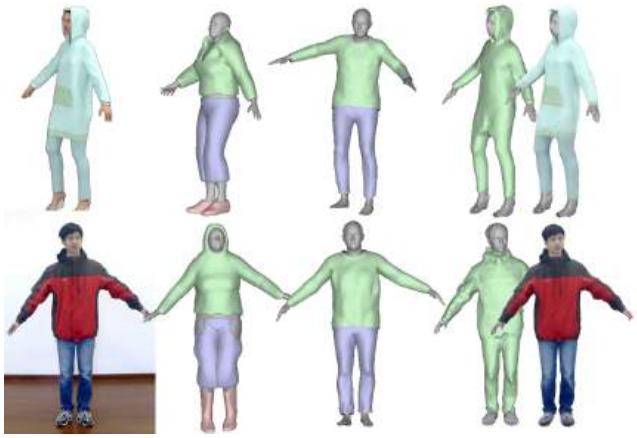


Fig. 7. Qualitative comparison of garment reconstruction. DELTA reconstructs different clothing types more faithfully than SMPLicit [Corona et al. 2021] and BCNet [Jiang et al. 2020].

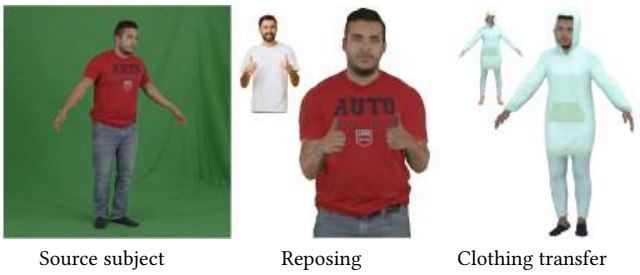


Fig. 8. Applications of DELTA. The hybrid representation enables (middle) reposing with detailed control over the body pose and (right) dressing up the source subject with target clothing. The target pose and clothing are shown in the inset images.

Fig. 9. Applications of DELTA. The hybrid representation enables transferring NeRF-based hairs into another face. Picture in the left indicates the source of the original hair. The avatar can also be animated with different poses and expressions.

Face and hair modeling. We conduct an evaluation of our proposed method using four real-world videos. To assess the effectiveness of our approach, we compare it with two state-of-the-art methods, neural head avatar (NHA) [Grassal et al. 2022] and IMavatar [Zheng et al. 2022]. To ensure a fair comparison, we adopt the same experimental protocol, where we train NHA and IMavatar using exactly the same set of video frames and reserve the remaining frames for evaluation. To be specific, for subjects "person_0000", "person_0004" and "MVI_1810", we sample every 50 frames for evaluation, and for the subject "b0_0", we sample every 5 frames. Following neural head avatar [Grassal et al. 2022], for each image, we keep the trained model and optimize per-frame parameters such as camera, pose, and expression. Consistent with prior research [Gafni et al. 2021; Grassal et al. 2022; Zheng et al. 2022], we employ four image-based metrics to evaluate our approach. These metrics include pixel-wise L1 loss, peak signal-to-noise ratio (PSNR), structural similarity metric (SSIM), and the learned perceptual image patch similarity (LPIPS). We find that NHA only focuses on the face, neck, and hair regions for training and evaluation. For a fair comparison, we compute the metrics on both the whole human region and only face, neck and hair regions.

The quantitative comparison presented in Table 2 demonstrates that our method attains the highest level of quality when considering the entire human region. However, when specifically focusing on the face, hair, and neck regions, it is worth noting that NHA achieves superior results for subjects with short hair, such as "person_0000". Nevertheless, when it comes to subjects with longer hair, NHA struggles to capture both hair and face details, as exemplified in instances such as "MVI_1810" and "b0_0". In contrast, our method performs effectively across various hair types and successfully captures the entirety of the avatar, including changes in the shoulders. This capability can be attributed to the utilization of hybrid representations within our approach.

We additionally provide qualitative comparisons for novel view images and shapes in Figure 5, along with supplementary qualitative results of DELTA applied to synthetic upper-body videos from the AGORA [Patel et al. 2021] dataset in Figure 6. Our method showcases superior performance in capturing accurate face and shoulder geometry, while also delivering high-quality renderings of the hair.

5.4 Applications

Body and garment reconstruction. We show comparisons on Garment reconstruction with SMPLicit [Corona et al. 2021] and

Video	Model	Whole				Face, Hair and Neck			
		L1 ↓	PSNR ↑	SSIM ↑	LIPIS ↓	L1 ↓	PSNR ↑	SSIM ↑	LIPIS ↓
person_0000	NHA [Grassal et al. 2022]	0.094	12.15	0.843	0.198	**0.012**	**24.92**	**0.920**	**0.046**
	IMavatar [Zheng et al. 2022]	0.024	22.55	0.882	0.177	0.015	23.70	0.917	0.089
	DELTA	**0.021**	**24.04**	**0.892**	**0.122**	0.017	23.37	0.914	0.086
MVI_1810	NHA [Grassal et al. 2022]	0.054	16.01	0.817	0.195	0.038	18.94	0.842	0.149
	IMavatar [Zheng et al. 2022]	**0.039**	20.33	0.829	0.171	**0.031**	21.44	0.851	0.137
	DELTA	**0.039**	**21.33**	**0.835**	**0.156**	0.034	**22.12**	**0.852**	**0.132**
b0_0	NHA [Grassal et al. 2022]	0.062	15.60	0.874	0.203	0.042	16.12	0.896	0.137
	IMavatar [Zheng et al. 2022]	0.043	19.61	0.871	0.188	0.030	20.13	0.905	**0.097**
	DELTA	**0.025**	**23.28**	**0.909**	**0.096**	**0.022**	**21.47**	**0.917**	0.103

Table 2. Quantitative comparison of novel pose and expression synthesis on public real videos.

BCNet [Jiang et al. 2020] in Fig 7. DELTA gives better visual quality than SMPLicit and BCNet. Note that the training/optimization settings are different, they reconstruct the body and garment from a single image, while our results are learned from video. However, they require a large set of 3D scans and manually designed cloth templates for training, while we do not need any 3D supervision, and capture the garment appearance as well. Figure 7 shows that DELTA reconstructs different clothing types more faithfully.

Reposing. For clothed body modeling, unlike previous methods that represent clothed bodies holistically, DELTA offers more fine-grained control over body pose especially hand pose. Figure 8 shows reposing into novel poses. Similar to the face and hair, utilizing an explicit shape model to present face region facilitates generalization across a wide range of facial expression animations. As Figure 9 shows different expressions of the reconstructed avatar.

Clothing and hair transfer. Figures 1, 8 and 9 qualitatively demonstrate the capability of our hybrid 3D representation in enabling clothing and hair transfer between avatars. We note that the clothing and hair is able to seamlessly adapt to accommodate various body shapes. Furthermore, the trained hair and clothing models can be both seamlessly transferred to different subjects. One potential application involves utilizing an existing body estimation method like PIXIE [Feng et al. 2021a] to estimate the body shape from a single image. Subsequently, our captured hair and clothing models can be applied to this subject, offering a streamlined approach for virtual try-on applications, as shown in Figure 10.

Altering human Shape. Figure 11 highlights an additional facet of DELTA's capabilities. We show the capacity to alter human body or face shape through adjustments in SMPL-X shape parameters. Subsequently, the NeRF-based clothing or hair seamlessly adjusts to align with the modified shape.

5.5 Ablation Study

We run different ablation experiments to show the impact of different components of our hybrid representation, and to show the impact of the pose refinements.

Effect of representations. DELTA consists of a NeRF to represent clothing, and a mesh with vertex displacements. Figure 12 compares NeRF to holistically represent body and clothing (i.e., DELTA w/o body-clothing segmentation) and mesh-only based representation

Fig. 10. Virtual try-on Application of DELTA. Given a single image, we can estimate the body shape using PIXIE [Feng et al. 2021a]. The body texture is from PIXIE template. Both the trained hair and clothing can be subsequently applied to this subject, resulting in smooth virtual try-on applications. In this instance, the captured hair is derived from the second example in Figure 6, and the clothing is from the second example of Figure 7.

(i.e., DELTA w/o NeRF). Our hybrid representation is better able to estimate the face, hands, and complex clothing. Note that, unlike our hybrid representation, none of the existing body NeRF methods can transfer clothing between avatars.

Effect of pose refinement. Since the pose estimation for each frame is not accurate, the pose refinement is important to gain details. We try learning our method without pose refinement. Figure 14 shows that pose refinement improves the image quality a lot.

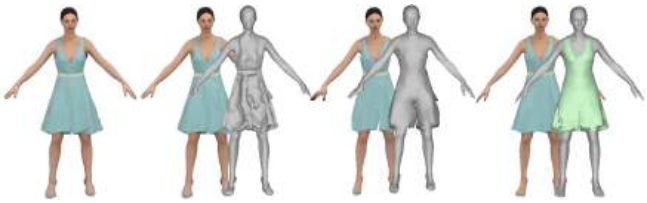
6 DISCUSSION AND LIMITATION

Segmentation. DELTA requires body and clothing/hair segmentation for training. Segmentation errors of the subject and background negatively impact the visual quality of the extracted avatar, and erroneous clothing or hair segmentation results in poor separation of mesh-based body and NeRF-based clothing or hair part. Figure 13 shows the wrong reconstruction due to consistent clothing segmentation errors, e.g. the belt is not recognized as part of clothing in segmentation, this results in wrong disentanglement between human body and clothing. Enforcing temporal consistency by exploiting optical flow could improve the segmentation quality.

Geometric quality. The strength of NeRF is its visual quality and the ability to synthesize realistic images, even when the geometry is not perfect. Figure 15 and Figure 16 show examples of noisy

Source subject Reconstruction Changing shape

Fig. 11. DELTA can change underlying body/face shapes by modifying SMPL-X shape parameters, and the NeRF-based clothing/hair will adapt to the new body/face shape accordingly.



Reference image NeRF SMPL-X +D Ours

Fig. 12. Rendered images and extracted meshes from different components of DELTA. Our hybrid representation gives a better estimated face, hand, and clothing geometry than vanilla NeRF or a mesh-based representation.

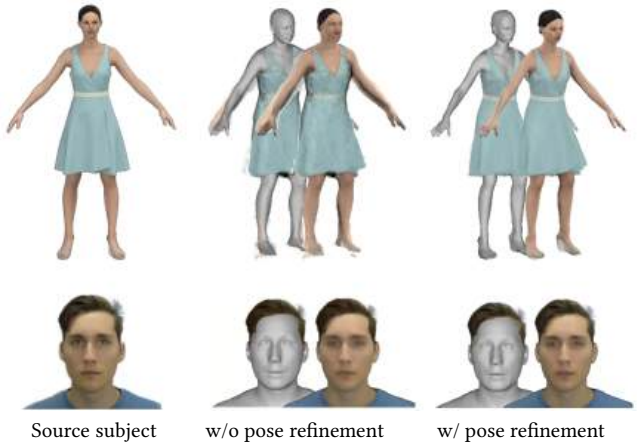
Input image and clothing segmentations Reconstruction

Fig. 13. The wrong clothing segmentation masks result in a visible gap within the reconstructed clothing.

geometry despite good visual quality. In contrast, recent SDF-based methods have demonstrated good geometric reconstruction (e.g., [Jiang et al. 2022]). It may be possible to leverage their results to better represent the underlying clothed shape or to regularize NeRF.

Novel poses and views. Although DELTA demonstrates generalization to unseen poses, artifacts may occur in extreme poses. As depicted in Figure 17, the animation results for new poses exhibit satisfactory performance for the body and face regions. However, artifacts are prevalent in the non-rigid fusion (clothing or hair) component. Notably, for regions that have not been encountered in the training data, our model will fail to capture the desired details. For instance, in the example featuring short hair, the hair in the head top is always missing in all poses and views in the video. To address these limitations, potential solutions include incorporating regularization techniques during NeRF optimization or training a generative model using a diverse set of training examples encompassing different individuals and poses. These approaches have the potential to enhance the robustness and accuracy of the model when dealing with unseen regions and extreme poses.

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



Source subject w/o pose refinement w/ pose refinement

Fig. 14. Rendering results of clothed body (up) and head (bottom) w/o and w/ pose refinement. The pose refinement improves the visual quality of the reconstruction, as more texture details are reconstructed. For the face subject, please zoom in to check the difference.

Reference image Captured clothing appearance & geometry

Fig. 15. While DELTA gives good visual quality for clothing renderings, the underlying geometry of the NeRF clothing is sometimes noisy.

Fig. 16. Two examples of captured hair appearance and geometry. While DELTA gives good visual quality for hair renderings, the underlying geometry of the NeRF hair is noisy.

Fig. 17. Some failure cases for avatar animation. Artifacts may occur in extreme poses. In the example shown at the bottom-left, the top part of the hair is absent since it was never observed in the training data.

Dynamics. DELTA handles non-rigid cloth deformation with the pose-conditioned deformation model. While the global pose can account for some deformation, how to accurately model the clothing and hair dynamics as a function of body movement remains an open problem and is an important future work.

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

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

7 CONCLUDING REMARKS

DELTA is able to automatically extract human body, clothing or hair from a monocular video. Our key novelty is a hybrid representation that combines a mesh-based body model with a neural radiance field to separately model the body and clothing/hair. This factored representation enables DELTA to transfer clothing/hair between avatars, animate the body pose of the avatars including finger articulation, alter their body shape and facial expression, and visualize them from unseen viewing directions. This property makes DELTA well suited to VR and virtual try-on applications. Finally, DELTA outperforms existing avatar extraction methods from videos in terms of visual quality and generality.

ACKNOWLEDGMENTS

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

Disclosure. MJB has received research gift funds from Adobe, Intel, Nvidia, Meta/Facebook, and Amazon. MJB has financial interests in Amazon, Datagen Technologies, and Meshcapade GmbH. While MJB is a consultant for Meshcapade, his research in this project was performed solely at, and funded solely by, the Max Planck Society.

REFERENCES

Oswald Aldrian and WA Smith. 2010. A linear approach of 3d face shape and texture recovery using a 3d morphable model. In *BMVC*.

Thiemo Alldieck, Marcus Magnor, Bharat Lal Bhatnagar, Christian Theobalt, and Gerard Pons-Moll. 2019a. Learning to reconstruct people in clothing from a single RGB camera. In *CVPR*.

Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. 2018a. Detailed human avatars from monocular video. In *3DV*.

Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. 2018b. Video based reconstruction of 3d people models. In *CVPR*.

Thiemo Alldieck, Gerard Pons-Moll, Christian Theobalt, and Marcus Magnor. 2019b. Tex2shape: Detailed full human body geometry from a single image. In *ICCV*.

Thiemo Alldieck, Hongyi Xu, and Cristian Sminchisescu. 2021. imGHUM: Implicit Generative Models of 3D Human Shape and Articulated Pose. In *ICCV*.

Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. 2005. SCAPE: Shape completion and animation of people. *ACM Transactions on Graphics* 24, 3 (2005), 408–416.

Ronen Basri and David W Jacobs. 2003. Lambertian reflectance and linear subspaces. *IEEE transactions on pattern analysis and machine intelligence* 25, 2 (2003), 218–233.

Thabo Beeler, Bernd Bickel, Gioacchino Noris, Paul Beardsley, Steve Marschner, Robert W Sumner, and Markus Gross. 2012. Coupled 3D reconstruction of sparse facial hair and skin. *ACM Transactions on Graphics* 31, 4 (2012), 1–10.

Hugo Bertiche, Meysam Madadi, and Sergio Escalera. 2020. CLOTH3D: clothed 3d humans. In *ECCV*.

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* 40, 4 (2021), 1–15.

Volker Blanz and Thomas Vetter. 1999. A morphable model for the synthesis of 3D faces. In *SIGGRAPH*.

Volker Blanz and Thomas Vetter. 2003. Face recognition based on fitting a 3D morphable model. *IEEE Transactions on pattern analysis and machine intelligence* 25, 9 (2003), 1063–1074.

Adrian Bulat and Georgios Tzimiropoulos. 2017. How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). In *ICCV*.

Chen Cao, Derek Bradley, Kun Zhou, and Thabo Beeler. 2015. Real-time high-fidelity facial performance capture. *ACM Transactions on Graphics* 34, 4 (2015), 1–9.

Chen Cao, Yanlin Weng, Shun Zhou, Yiying Tong, and Kun Zhou. 2013. Facewarehouse: A 3d facial expression database for visual computing. *IEEE Transactions on Visualization and Computer Graphics* 20, 3 (2013), 413–425.

Menglei Chai, Lvdi Wang, Yanlin Weng, Xiaogang Jin, and Kun Zhou. 2013. Dynamic hair manipulation in images and videos. *ACM Transactions on Graphics* 32, 4 (2013), 1–8.

Menglei Chai, Lvdi Wang, Yanlin Weng, Yizhou Yu, Baining Guo, and Kun Zhou. 2012. Single-view hair modeling for portrait manipulation. *ACM Transactions on Graphics* 31, 4 (2012), 1–8.

Eric R Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. 2021. pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In *CVPR*.

Jianchuan Chen, Ying Zhang, Di Kang, Xuefei Zhe, Linchao Bao, Xu Jia, and Huchuan Lu. 2021b. Animatable neural radiance fields from monocular rgb videos. *arXiv preprint arXiv:2106.13629* (2021).

Xin Chen, Anqi Pang, Wei Yang, Peihao Wang, Lan Xu, and Jingyi Yu. 2021a. TightCap: 3D Human Shape Capture with Clothing Tightness Field. *ACM Transactions on Graphics* 41, 1 (2021), 1–17.

Zhiqin Chen and Hao Zhang. 2019. Learning implicit fields for generative shape modeling. In *CVPR*.

Julian Chibane, Aymen Mir, and Gerard Pons-Moll. 2020. Neural Unsigned Distance Fields for Implicit Function Learning. In *NeurIPS*.

Vasileios Choutas, Georgios Pavlakos, Timo Bolkart, Dimitrios Tzionas, and Michael J. Black. 2020. Monocular Expressive Body Regression through Body-Driven Attention. In *ECCV*.

Forrester Cole, David Belanger, Dilip Krishnan, Aaron Sarna, Inbar Mosseri, and William T Freeman. 2017. Synthesizing normalized faces from facial identity features. In *CVPR*.

Enric Corona, Albert Pumarola, Guillem Alenya, Gerard Pons-Moll, and Francesc Moreno-Noguer. 2021. Smplicit: Topology-aware generative model for clothed people. In *CVPR*.

Levin Dabhi. 2022. Clothes Segmentation using U2NET. https://github.com/levindabhi/cloth-segmentation

Hang Dai, Nick Pears, William Smith, and Christian Duncan. 2020. Statistical modeling of craniofacial shape and texture. *International Journal of Computer Vision* 128, 2

(2020), 547–571.
Radek Daněček, Michael J Black, and Timo Bolkart. 2022. EMOCA: Emotion driven monocular face capture and animation. In *CVPR*.
R Daněřek, Endri Dibra, Cengiz Öztireli, Remo Ziegler, and Markus Gross. 2017. Deepgarment: 3d garment shape estimation from a single image. In *Computer Graphics Forum*, Vol. 36. Wiley Online Library, 269–280.
Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. 2019. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *CVPR Workshops*.
Abdallah Dib, Cédric Thébault, Junghyun Ahn, Philippe-Henri Gosselin, Christian Theobalt, and Louis Chevallier. 2021. Towards high fidelity monocular face reconstruction with rich reflectance using self-supervised learning and ray tracing. In *ICCV*.
Bernhard Egger, William AP Smith, Ayush Tewari, Stefanie Wuhrer, Michael Zollhoefer, Thabo Beeler, Florian Bernard, Timo Bolkart, Adam Kortylewski, Sami Romdhani, et al. 2020. 3d morphable face models—past, present, and future. *ACM Transactions on Graphics* 39, 5 (2020), 1–38.
Yao Feng, Vasileios Choutas, Timo Bolkart, Dimitrios Tzionas, and Michael J Black. 2021a. Collaborative regression of expressive bodies using moderation. In *3DV*.
Yao Feng, Haiwen Feng, Michael J Black, and Timo Bolkart. 2021b. Learning an animatable detailed 3D face model from in-the-wild images. *ACM Transactions on Graphics* 40, 4 (2021), 1–13.
Yao Feng, Jinlong Yang, Marc Pollefeys, Michael J Black, and Timo Bolkart. 2022. Capturing and Animation of Body and Clothing from Monocular Video. In *SIGGRAPH Asia*.
Guy Gafni, Justus Thies, Michael Zollhofer, and Matthias Nießner. 2021. Dynamic neural radiance fields for monocular 4d facial avatar reconstruction. In *CVPR*.
Xuan Gao, Chenglai Zhong, Jun Xiang, Yang Hong, Yudong Guo, and Juyong Zhang. 2022. Reconstructing Personalized Semantic Facial NeRF Models From Monocular Video. In *SIGGRAPH Asia*.
Baris Gecer, Stylianos Ploumpis, Irene Kotsia, and Stefanos Zafeiriou. 2019. Ganfit: Generative adversarial network fitting for high fidelity 3d face reconstruction. In *CVPR*.
Philip-William Grassal, Malte Prinzler, Titus Leistner, Carsten Rother, Matthias Nießner, and Justus Thies. 2022. Neural head avatars from monocular RGB videos. In *CVPR*.
Oshri Halimi, Fabian Prada, Tuur Stuyck, Donglai Xiang, Timur Bagautdinov, He Wen, Ron Kimmel, Takaaki Shiratori, Chenglei Wu, and Yaser Sheikh. 2022. Garment avatars: Realistic cloth driving using pattern registration. *arXiv preprint arXiv:2206.03373* (2022).
Tong He, Yuanlu Xu, Shunsuke Saito, Stefano Soatto, and Tony Tung. 2021. ARCH++: Animation-ready clothed human reconstruction revisited. In *ICCV*.
Tomas Lay Herrera, Arno Zinke, and Andreas Weber. 2012. Lighting hair from the inside: A thermal approach to hair reconstruction. *ACM Transactions on Graphics* 31, 6 (2012), 1–9.
David A. Hirshberg, Matthew Loper, Eric Rachlin, and Michael J. Black. 2012. Coregistration: Simultaneous Alignment and Modeling of Articulated 3D Shape. In *ECCV*.
Fangzhou Hong, Liang Pan, Zhongang Cai, and Ziwei Liu. 2021. Garment4d: Garment reconstruction from point cloud sequences. In *NIPS*.
Liwen Hu, Chongyang Ma, Linjie Luo, and Hao Li. 2014. Robust hair capture using simulated examples. *ACM Transactions on Graphics* 33, 4 (2014), 1–10.
Zeng Huang, Yuanlu Xu, Christoph Lassner, Hao Li, and Tony Tung. 2020. ARCH: Animatable reconstruction of clothed humans. In *CVPR*.
Peter J. Huber. 1964. Robust Estimation of a Location Parameter. *The Annals of Mathematical Statistics* 35, 1 (1964), 73 – 101.
Boyi Jiang, Yang Hong, Hujun Bao, and Juyong Zhang. 2022. SelfRecon: Self Reconstruction Your Digital Avatar from Monocular Video. In *CVPR*.
Boyi Jiang, Juyong Zhang, Yang Hong, Jinhao Luo, Ligang Liu, and Hujun Bao. 2020. Bcnet: Learning body and cloth shape from a single image. In *ECCV*.
Ning Jin, Yilin Zhu, Zhenglin Geng, and Ronald Fedkiw. 2020. A Pixel-Based Framework for Data-Driven Clothing. In *Computer Graphics Forum*, Vol. 39. Wiley Online Library, 135–144.
Justin Johnson, Alexandre Alahi, and Li Fei-Fei. 2016. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*.
Hanbyul Joo, Tomas Simon, and Yaser Sheikh. 2018. Total capture: A 3d deformation model for tracking faces, hands, and bodies. In *CVPR*.
Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. 2018. End-to-end recovery of human shape and pose. In *CVPR*.
Zhanghan Ke, Jiayu Sun, Kaican Li, Qiong Yan, and Rynson W.H. Lau. 2022. MODNet: Real-Time Trimap-Free Portrait Matting via Objective Decomposition. In *AAAI*.
Petr Kellnhofer, Lars C Jebe, Andrew Jones, Ryan Spicer, Kari Pulli, and Gordon Wetzstein. 2021. Neural lumigraph rendering. In *CVPR*.
Taras Khakhulin, Vanessa Sklyarova, Victor Lempitsky, and Egor Zakharov. 2022. Realistic One-shot Mesh-based Head Avatars. In *ECCV*.
Laehyun Kim, Gaurav S Sukhatme, and Mathieu Desbrun. 2004. A haptic-rendering technique based on hybrid surface representation. *IEEE computer graphics and applications* 24, 2 (2004), 66–75.
Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. 2019. Learning to reconstruct 3D human pose and shape via model-fitting in the loop. In *ICCV*.
Alexandros Lattas, Stylianos Moschoglou, Baris Gecer, Stylianos Ploumpis, Vasileios Triantafyllou, Abhijeet Ghosh, and Stefanos Zafeiriou. 2020. AvatarMe: Realistically Renderable 3D Facial Reconstruction" in-the-wild". In *CVPR*.
Verica Lazova, Eldar Insafutdinov, and Gerard Pons-Moll. 2019. 360-Degree Textures of People in Clothing from a Single Image. In *3DV*.
Gengyan Li, Abhimitra Meka, Franziska Mueller, Marcel C Buehler, Otmar Hilliges, and Thabo Beeler. 2022. EyeNeRF: a hybrid representation for photorealistic synthesis, animation and relighting of human eyes. *ACM Transactions on Graphics* 41, 4 (2022), 1–16.
Ruilong Li, Hang Gao, Matthew Tancik, and Angjoo Kanazawa. 2023a. NerfAcc: Efficient Sampling Accelerates NeRFs. *arXiv preprint arXiv:2305.04966* (2023).
Tianye Li, Timo Bolkart, Michael J Black, Hao Li, and Javier Romero. 2017. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics* 36, 6 (2017), 194–1.
Xiaoxing Li, Tao Jia, and Hao Zhang. 2009. Expression-insensitive 3D face recognition using sparse representation. In *CVPR*.
Yue Li, Marc Habermann, Bernhard Thomaszewski, Stelian Coros, Thabo Beeler, and Christian Theobalt. 2021. Deep physics-aware inference of cloth deformation for monocular human performance capture. In *3DV*.
Yue Li, Liqian Ma, Haoqiang Fan, and Kenny Mitchell. 2018. Feature-preserving detailed 3d face reconstruction from a single image. In *ACM SIGGRAPH European Conference on Visual Media Production*.
Zhe Li, Zerong Zheng, Yuxiao Liu, Boyao Zhou, and Yebin Liu. 2023b. PoseVocab: Learning Joint-structured Pose Embeddings for Human Avatar Modeling. In *SIGGRAPH*.
Shanchuan Lin, Linjie Yang, Imran Saleemi, and Soumyadip Sengupta. 2022. Robust Video Matting (RVM). https://github.com/PeterL1n/RobustVideoMatting
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* 40, 6 (2021), 1–16.
Shaohui Liu, Yinda Zhang, Songyou Peng, Boxin Shi, Marc Pollefeys, and Zhaopeng Cui. 2020. Dist: Rendering deep implicit signed distance function with differentiable sphere tracing. In *CVPR*.
Wu Liu, Qian Bao, Yu Sun, and Tao Mei. 2022a. Recent advances of monocular 2d and 3d human pose estimation: A deep learning perspective. *Comput. Surveys* 55, 4 (2022), 1–41.
Weiyang Liu, Zhen Liu, Liam Paull, Adrian Weller, and Bernhard Schölkopf. 2022b. Structural causal 3d reconstruction. In *ECCV*.
Wen Liu, Zhixin Piao, Min Jie, Wenhan Luo, Lin Ma, and Shenghua Gao. 2019. Liquid Warping GAN: A Unified Framework for Human Motion Imitation, Appearance Transfer and Novel View Synthesis. In *ICCV*.
Stephen Lombardi, Jason Saragih, Tomas Simon, and Yaser Sheikh. 2018. Deep appearance models for face rendering. *ACM Transactions on Graphics* 37, 4 (2018), 1–13.
Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. 2015. SMPL: A Skinned Multi-Person Linear Model. *ACM Transactions on Graphics* 34, 6 (2015), 248:1–248:16.
Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al. 2019. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*.
Linjie Luo, Hao Li, Sylvain Paris, Thibaut Weise, Mark Pauly, and Szymon Rusinkiewicz. 2012. Multi-view hair capture using orientation fields. In *CVPR*.
Linjie Luo, Hao Li, and Szymon Rusinkiewicz. 2013. Structure-aware hair capture. *ACM Transactions on Graphics* 32, 4 (2013), 1–12.
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 *CVPR*.
Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. 2019. Occupancy networks: Learning 3d reconstruction in function space. In *CVPR*.
Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. 2020. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*.
Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. 2022. Instant Neural Graphics Primitives with a Multiresolution Hash Encoding. *ACM Transactions on Graphics* 41, 4, Article 102 (july 2022), 15 pages. https://doi.org/10.1145/3528223.3530127
Giljoo Nam, Chenglei Wu, Min H Kim, and Yaser Sheikh. 2019. Strand-accurate multi-view hair capture. In *CVPR*.
Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. 2020. Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In *CVPR*.

Mohamed Omran, Christoph Lassner, Gerard Pons-Moll, Peter V. Gehler, and Bernt Schiele. 2018. Neural Body Fitting: Unifying Deep Learning and Model Based Human Pose and Shape Estimation. In *3DV*.

Ahmed AA Osman, Timo Bolkart, and Michael J Black. 2020. Star: Sparse trained articulated human body regressor. In *ECCV*.

Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. 2019. Deepsdf: Learning continuous signed distance functions for shape representation. In *CVPR*.

Keunhong Park, Utkarsh Sinha, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Steven M Seitz, and Ricardo Martin-Brualla. 2021. Nerfies: Deformable neural radiance fields. In *ICCV*.

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 *CVPR*.

Priyanka Patel, Chun-Hao Paul Huang, Joachim Tesch, David Hoffmann, Shashank Tripathi, and Michael J. Black. 2021. AGORA: Avatars in Geography Optimized for Regression Analysis. In *CVPR*.

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. 2019. Expressive Body Capture: 3D Hands, Face, and Body From a Single Image. In *CVPR*.

Georgios Pavlakos, Ethan Weber, Matthew Tancik, and Angjoo Kanazawa. 2022. The One Where They Reconstructed 3D Humans and Environments in TV Shows. In *ECCV*.

Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas Vetter. 2009. A 3D face model for pose and illumination invariant face recognition. In *IEEE international conference on advanced video and signal based surveillance*.

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 *ICCV*.

Sida Peng, Shangzhan Zhang, Zhen Xu, Chen Geng, Boyi Jiang, Hujun Bao, and Xiaowei Zhou. 2022. Animatable Neural Implicit Surfaces for Creating Avatars from Videos. *arXiv preprint arXiv:2203.08133* (2022).

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 *CVPR*.

Stylianos Ploumpis, Evangelos Ververas, Eimear O'Sullivan, Stylianos Moschoglou, Haoyang Wang, Nick Pears, William AP Smith, Baris Gecer, and Stefanos Zafeiriou. 2020. Towards a complete 3D morphable model of the human head. *IEEE transactions on pattern analysis and machine intelligence* 43, 11 (2020), 4142–4160.

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

Sergey Prokudin, Michael J. Black, and Javier Romero. 2021. SMPLpix: Neural Avatars from 3D Human Models. In *WACV*.

Lingteng Qiu, Guanying Chen, Jiapeng Zhou, Mutian Xu, Junle Wang, and Xiaoguang Han. 2023. REC-MV: REconstructing 3D Dynamic Cloth from Monocular Videos. In *CVPR*.

Eduard Ramon, Gil Triginer, Janna Escur, Albert Pumarola, Jaime Garcia, Xavier Giro-i Nieto, and Francesc Moreno-Noguer. 2021. H3d-net: Few-shot high-fidelity 3d head reconstruction. In *ICCV*.

Anurag Ranjan, Timo Bolkart, Soubhik Sanyal, and Michael J Black. 2018. Generating 3D faces using convolutional mesh autoencoders. In *ECCV*.

Sami Romdhani and Thomas Vetter. 2005. Estimating 3D shape and texture using pixel intensity, edges, specular highlights, texture constraints and a prior. In *CVPR*.

Yu Rong, Takaaki Shiratori, and Hanbyul Joo. 2021. FrankMocap: A Monocular 3D Whole-Body Pose Estimation System via Regression and Integration. In *ICCV Workshops*.

Radu Alexandru Rosu, Shunsuke Saito, Ziyan Wang, Chenglei Wu, Sven Behnke, and Giljoo Nam. 2022. Neural Strands: Learning Hair Geometry and Appearance from Multi-View Images. In *ECCV*.

Shunsuke Saito, Liwen Hu, Chongyang Ma, Hikaru Ibayashi, Linjie Luo, and Hao Li. 2018. 3D hair synthesis using volumetric variational autoencoders. *ACM Transactions on Graphics* 37, 6 (2018), 1–12.

Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Hao Li, and Angjoo Kanazawa. 2019. PIFu: Pixel-Aligned Implicit Function for High-Resolution Clothed Human Digitization. In *ICCV*.

Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. 2020. PIFuHD: Multi-Level Pixel-Aligned Implicit Function for High-Resolution 3D Human Digitization. In *CVPR*.

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

Soubhik Sanyal, Timo Bolkart, Haiwen Feng, and Michael J Black. 2019. Learning to regress 3D face shape and expression from an image without 3D supervision. In *CVPR*.

Sandro Schönborn, Bernhard Egger, Andreas Morel-Forster, and Thomas Vetter. 2017. Markov chain monte carlo for automated face image analysis. *International Journal of Computer Vision* 123, 2 (2017), 160–183.

Jiaxiang Shang, Tianwei Shen, Shiwei Li, Lei Zhou, Mingmin Zhen, Tian Fang, and Long Quan. 2020. Self-supervised monocular 3d face reconstruction by occlusion-aware multi-view geometry consistency. In *ECCV*.

Vanessa Sklyarova, Jenya Chelishev, Andreea Dogaru, Igor Medvedev, Victor Lempitsky, and Egor Zakharov. 2023. Neural Haircut: Prior-Guided Strand-Based Hair Reconstruction. In *ICCV*.

Shih-Yang Su, Frank Yu, Michael Zollhöfer, and Helge Rhodin. 2021. A-NeRF: Articulated neural radiance fields for learning human shape, appearance, and pose. In *NeurIPS*.

Zhaoqi Su, Tao Yu, Yangang Wang, and Yebin Liu. 2022. Deepcloth: Neural garment representation for shape and style editing. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 2 (2022), 1581–1593.

Tiancheng Sun, Giljoo Nam, Carlos Aliaga, Christophe Hery, and Ravi Ramamoorthi. 2021. Human Hair Inverse Rendering using Multi-View Photometric data. In *Eurographics Symposium on Rendering*.

Ayush Tewari, Florian Bernard, Pablo Garrido, Gaurav Bharaj, Mohamed Elgharib, Hans-Peter Seidel, Patrick Pérez, Michael Zollhofer, and Christian Theobalt. 2019. Fml: Face model learning from videos. In *CVPR*.

Ayush Tewari, Michael Zollhöfer, Pablo Garrido, Florian Bernard, Hyeongwoo Kim, Patrick Pérez, and Christian Theobalt. 2018. Self-supervised multi-level face model learning for monocular reconstruction at over 250 hz. In *CVPR*.

Ayush Tewari, Michael Zollhofer, Hyeongwoo Kim, Pablo Garrido, Florian Bernard, Patrick Perez, and Christian Theobalt. 2017. Mofa: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction. In *ICCV*.

Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. 2016. Face2face: Real-time face capture and reenactment of rgb videos. In *CVPR*.

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

Garvita Tiwari, Bharat Lal Bhatnagar, Tony Tung, and Gerard Pons-Moll. 2020. SIZER: A dataset and model for parsing 3d clothing and learning size sensitive 3d clothing. In *ECCV*.

Luan Tran and Xiaoming Liu. 2018. Nonlinear 3d face morphable model. In *CVPR*.

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

Daniel Vlasic, Matthew Brand, Hanspeter Pfister, and Jovan Popovic. 2006. Face transfer with multilinear models. In *SIGGRAPH*.

Yi Wang, Xin Tao, Xiaojuan Qi, Xiaoyong Shen, and Jiaya Jia. 2018. Image inpainting via generative multi-column convolutional neural networks. In *NeurIPS*.

Ziyan Wang, Timur Bagautdinov, Stephen Lombardi, Tomas Simon, Jason Saragih, Jessica Hodgins, and Michael Zollhofer. 2021. Learning compositional radiance fields of dynamic human heads. In *CVPR*.

Ziyan Wang, Giljoo Nam, Tuur Stuyck, Stephen Lombardi, Michael Zollhöfer, Jessica Hodgins, and Christoph Lassner. 2022. HVH: Learning a Hybrid Neural Volumetric Representation for Dynamic Hair Performance Capture. In *CVPR*.

Kelly Ward, Florence Bertails, Tae-Yong Kim, Stephen R Marschner, Marie-Paule Cani, and Ming C Lin. 2007. A survey on hair modeling: Styling, simulation, and rendering. *IEEE transactions on visualization and computer graphics* 13, 2 (2007), 213–234.

Yandong Wen, Weiyang Liu, Bhiksha Raj, and Rita Singh. 2021. Self-supervised 3d face reconstruction via conditional estimation. In *ICCV*.

Chung-Yi Weng, Brian Curless, Pratul P. Srinivasan, Jonathan T. Barron, and Ira Kemelmacher-Shlizerman. 2022. HumanNeRF: Free-Viewpoint Rendering of Moving People From Monocular Video. In *CVPR*.

Shangzhe Wu, Christian Rupprecht, and Andrea Vedaldi. 2020. Unsupervised learning of probably symmetric deformable 3d objects from images in the wild. In *CVPR*.

Donglai Xiang, Hanbyul Joo, and Yaser Sheikh. 2019. Monocular Total Capture: Posing Face, Body, and Hands in the Wild. In *CVPR*.

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* 40, 6 (2021), 1–15.

Yiqing Xiao. 2022. head video. https://www.bilibili.com/video/BV1b84y1q7Vm/?spm_id_from=333.999.0.0

Yuliang Xiu, Jinlong Yang, Xu Cao, Dimitrios Tzionas, and Michael J. Black. 2023. ECON: Explicit Clothed humans Optimized via Normal integration. In *CVPR*.

Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J. Black. 2022. ICON: Implicit Clothed humans Obtained from Normals. In *CVPR*.

Hongyi Xu, Eduard Gabriel Bazavan, Andrei Zanfir, William T. Freeman, Rahul Sukthankar, and Cristian Sminchisescu. 2020. GHUM & GHUML: Generative 3D Human Shape and Articulated Pose Models. In *CVPR*.

Jiamin Xu, Zihan Zhu, Hujun Bao, and Wewei Xu. 2022. A Hybrid Mesh-neural Representation for 3D Transparent Object Reconstruction. *arXiv preprint arXiv:2203.12613* (2022).

Yuelang Xu, Lizhen Wang, Xiaochen Zhao, Hongwen Zhang, and Yebin Liu. 2023a. AvatarMAV: Fast 3D Head Avatar Reconstruction Using Motion-Aware Neural Voxels. In *SIGGRAPH*.

Yuelang Xu, Hongwen Zhang, Lizhen Wang, Xiaochen Zhao, Huang Han, Qi Guojun, and Yebin Liu. 2023b. LatentAvatar: Learning Latent Expression Code for Expressive Neural Head Avatar. In *SIGGRAPH*.

Lingchen Yang, Zefeng Shi, Youyi Zheng, and Kun Zhou. 2019. Dynamic hair modeling from monocular videos using deep neural networks. *ACM Transactions on Graphics* 38, 6 (2019), 1–12.

Ze Yang, Shenlong Wang, Sivabalan Manivasagam, Zeng Huang, Wei-Chiu Ma, Xinchen Yan, Ersin Yumer, and Raquel Urtasun. 2021. S3: Neural shape, skeleton, and skinning fields for 3D human modeling. In *CVPR*.

Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. 2021. Volume rendering of neural implicit surfaces. In *NeurIPS*.

Lior Yariv, Yoni Kasten, Dror Moran, Meirav Galun, Matan Atzmon, Basri Ronen, and Yaron Lipman. 2020. Multiview neural surface reconstruction by disentangling geometry and appearance. In *NeurIPS*.

Tarun Yenamandra, Ayush Tewari, Florian Bernard, Hans-Peter Seidel, Mohamed Elgharib, Daniel Cremers, and Christian Theobalt. 2021. i3dmm: Deep implicit 3d morphable model of human heads. In *CVPR*.

Andrei Zanfir, Eduard Gabriel Bazavan, Mihai Zanfir, William T Freeman, Rahul Sukthankar, and Cristian Sminchisescu. 2021. Neural descent for visual 3d human pose and shape. In *CVPR*.

Meng Zhang, Menglei Chai, Hongzhi Wu, Hao Yang, and Kun Zhou. 2017. A data-driven approach to four-view image-based hair modeling. *ACM Transactions on Graphics* 36, 4 (2017), 1–11.

Meng Zhang and Youyi Zheng. 2019. Hair-GAN: Recovering 3D hair structure from a single image using generative adversarial networks. *Visual Informatics* 3, 2 (2019), 102–112.

Fang Zhao, Wenhao Wang, Shengcai Liao, and Ling Shao. 2021. Learning anchored unsigned distance functions with gradient direction alignment for single-view garment reconstruction. In *ICCV*.

Yufeng Zheng, Victoria Fernández Abrevaya, Marcel C Bühler, Xu Chen, Michael J Black, and Otmar Hilliges. 2022. Im avatar: Implicit morphable head avatars from videos. In *CVPR*.

Yujian Zheng, Zirong Jin, Moran Li, Haibin Huang, Chongyang Ma, Shuguang Cui, and Xiaoguang Han. 2023a. Hairstep: Transfer synthetic to real using strand and depth maps for single-view 3d hair modeling. In *CVPR*.

Yufeng Zheng, Wang Yifan, Gordon Wetzstein, Michael J. Black, and Otmar Hilliges. 2023b. PointAvatar: Deformable Point-based Head Avatars from Videos. In *CVPR*.

Zerong Zheng, Tao Yu, Yebin Liu, and Qionghai Dai. 2021. Pamir: Parametric model-conditioned implicit representation for image-based human reconstruction. *IEEE transactions on pattern analysis and machine intelligence* 44, 6 (2021), 3170–3184.

Zerong Zheng, Xiaochen Zhao, Hongwen Zhang, Boning Liu, and Yebin Liu. 2023c. AvatarRex: Real-time Expressive Full-body Avatars. *ACM Transactions on Graphics* 42, 4 (2023).

Yuxiao Zhou, Marc Habermann, Ikhsanul Habibie, Ayush Tewari, Christian Theobalt, and Feng Xu. 2021. Monocular Real-Time Full Body Capture With Inter-Part Correlations. In *CVPR*.

Yi Zhou, Liwen Hu, Jun Xing, Weikai Chen, Han-Wei Kung, Xin Tong, and Hao Li. 2018. Hairnet: Single-view hair reconstruction using convolutional neural networks. In *ECCV*.

Heming Zhu, Yu Cao, Hang Jin, Weikai Chen, Dong Du, Zhangye Wang, Shuguang Cui, and Xiaoguang Han. 2020. Deep Fashion3D: A dataset and benchmark for 3D garment reconstruction from single images. In *ECCV*.

Heming Zhu, Lingteng Qiu, Yuda Qiu, and Xiaoguang Han. 2022. Registering Explicit to Implicit: Towards High-Fidelity Garment mesh Reconstruction from Single Images. In *CVPR*.

zllrunning. 2019. face-parsing.PyTorch. https://github.com/zllrunning/face-parsing.PyTorch