LayerNet: High-Resolution Semantic 3D Reconstruction of Clothed People

Enric Corona Guillem Alenyà Gerard Pons-Moll Francesc Moreno-Noguer

Abstract—In this paper we introduce SMPLicit, a novel generative model to jointly represent body pose, shape and clothing geometry; and LayerNet, a deep network that given a single image of a person simultaneously performs detailed 3D reconstruction of body and clothes. In contrast to existing learning-based approaches that require training specific models for each type of garment, SMPLicit can represent in a unified manner different garment topologies (*e.g.* from sleeveless tops to hoodies and open jackets), while controlling other properties like garment size or tightness/looseness.

LayerNet follows a coarse-to-fine multi-stage strategy by first predicting smooth cloth geometries from SMPLicit, which are then refined by an image-guided displacement network that gracefully fits the body recovering high-frequency details and wrinkles. LayerNet achieves competitive accuracy in the task of 3D reconstruction against current 'garment-agnostic' state of the art for images of people in up-right positions and controlled environments, and consistently surpasses these methods on challenging body poses and uncontrolled settings. Furthermore, the semantically rich outcome of our approach is suitable for performing Virtual Try-on tasks directly on 3D, a task which, so far, has only been addressed in the 2D domain.

Index Terms—3D human reconstruction, 3D virtual try-on

1 Introduction

Building a differentiable and low dimensional generative model capable to control garments style and deformations under different body shapes and poses would open the door to many exciting applications in *e.g.* digital animation of clothed humans, 3D content creation and virtual try-on. However, while such representations have been shown effective for the case of the undressed human body [1], [2], where body shape variation can be encoded by a few parameters of a linear model, there exist so far, no similar approach for doing so on clothes.

The standard practice to represent the geometry of dressed people has been to treat clothing as an additive displacement over canonical body shapes, typically obtained with SMPL [3], [4], [5], [6]. Nevertheless, these types of approaches cannot tackle the main challenge in garment modeling, which is the large variability of types, styles, cut, and deformations they can have. For instance, upper body clothing can be either a sleeveless top, a long-sleeve hoodie or an open jacket. In order to handle such variability, existing approaches need to train specific models for each type of garment, hampering thus their practical utilization.

In this paper, we introduce SMPLicit, a topologically-aware generative model for clothed bodies that can be controlled by a low-dimensional and interpretable vector of parameters. SMPLicit builds upon an implicit network architecture conditioned on the body pose and shape. With these two factors, we can predict clothing deformation in 3D as a function of the body geometry, while controlling the garment style (cloth category) and cut (*e.g.* sleeve length, tight or loose-fitting). We independently train this model for two distinct cloth clusters, namely *upper body* (including sleeveless tops, T-shirts, hoodies and jackets) and *lower body* (including pants, shorts and skirts). Within each cluster, the same model is able to represent garments with very different geometric properties and topology while allowing to smoothly and consistently interpolate between their geometries. *Shoes* and *hair* categories are also modeled as independent categories. Interestingly, SMPLicit is fully differentiable and can be easily deployed and integrated into larger end-to-end deep learning systems.

Concretely, we demonstrate that SMPLicit can be readily applied to two different problems. First, for fitting 3D scans of dressed people. In this problem, our multi-garment "generic" model is on a par with other approaches that were specifically trained for each garment [5], [6]. We also apply SMPLicit for the challenging problem of 3D reconstruction from images, where we compare favorably to state-of-the-art, being able to retrieve complex garment geometries under different body poses, and can tackle situations with multiple clothing layers. Fig. 1 shows one such example, where besides reconstructing the geometry of the full outfit, SMPLicit provides semantic knowledge of the shape, allowing then for garment editing and body re-posing, key ingredients of virtual try-on systems.

A preliminary version of this work was presented in [7], in which we showed our generative model was able to represent multiple cloth topologies, interpolate between them and fit clothes from 3D scans and images. In this paper we also present LayerNet, which is based on SMPLicit to predict body and garments as layers, while still preserving high fidelity detail and texture, given a single in-the-wild image of a person in an arbitrary pose. For this purpose, we formulate a two-stage pipeline that combines the best of

- *Enric Corona, Guillem Alenyà and Francesc Moreno-Noguer were with the Institut de Robòtica i Informàtica Industrial, CSIC-UPC, Barcelona, Spain.*
- *Gerard Pons-Moll is with the University of Tubingen and the Max Planck Institute for Informatics, Germany.*

E-mail: ecorona@iri.upc.edu

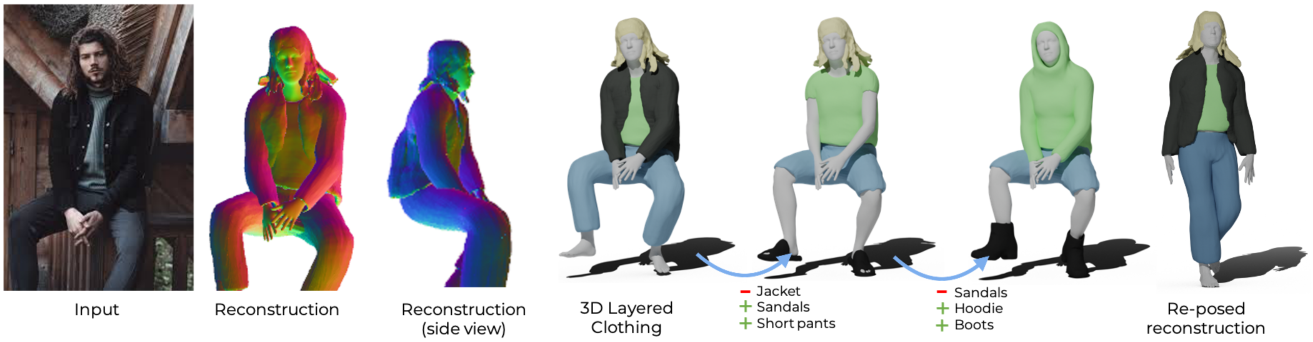


Fig. 1. We introduce SMPLicit, a fully differentiable generative model for clothed bodies, capable of representing garments with different topology. The four figures on the left show the application of the model to the problem of 3D body and cloth reconstruction from an input image. We are able to predict different models per cloth, even for multi-layer cases. Three right-most images: The model can also be used for editing the outfits, removing/adding new garments and re-posing the body.

model-based and model-free paradigms. Concretely, in the first stage, we predict body shape and pose parameters with SMPL [1] and clothing with SMPLicit. In the second stage, we process the image with pixel-aligned feature maps: we estimate features, normal maps and front and back RGB completion. We then introduce a deformation network that predicts point-wise displacements on the smooth shape to add detail based on pixel-aligned features, obtained by projecting the SMPLicit model vertices onto the feature maps. This makes the deformation network independent of the number of vertices of the layers. In addition, the deformation network reasons about the different cloth layers, enabling the full reconstruction of garments that are mostly occluded. Similar to the deformation network, a texture network predicts front and back per-vertex colors, yielding a high-resolution, textured, and complete geometric reconstruction for all garments.

A thorough evaluation demonstrates that LayerNet achieves reconstructions in a forward pass which are competitive with PIFuHD [8] (arguably the SOTA of single-layer models) for people in upright positions, and consistently overcomes this approach on images with complex body poses. Most importantly, the detailed reconstructions we provide are accompanied by a 3D semantic segmentation into the meshes of the body and each one of the garments, even for multi-layer outfits. This rich output opens the door to a number of novel applications that involve 3D mesh editing, like 3D virtual try-on. Fig. 1 shows one such example, in which after applying SMPLicit to the images of two people we can easily swap their clothes.

2 Related Work

Reconstruction and modelling of clothes is a long-standing goal lying at the intersection of computer vision and computer graphics. We next discuss related works, grouping them in *Generative cloth models*, *3D reconstruction of clothed humans* and *Cloth editing*, the three main topics in which we contribute.

2.1 Generative cloth models

Drawing inspiration on the success of the data driven methods for modeling the human body [1], [2], [10], [11], [12], [13], [14], a number of approaches aim to learn clothing models from real data, obtained using multiple images [3], [15], [16], [17], [18], 3D scans [19], [20], [21] or RGBD sensors [22], [23]. Nevertheless, capturing a sufficiently large volume of data to represent the complexity of clothes is still an open challenge, and methods built using real data [24], [25], [26] have problems to generalize beyond the deformation patterns of the training data. [5] addresses this limitation by means of a probabilistic formulation that predicts clothing displacements on the graph defined by the SMPL mesh. While this strategy improves the generalization capabilities, the clothes it is able to generate can not largely depart from the shape of a "naked" body defined by SMPL.

An alternative to the use of real data is to learn clothing models using data from physics simulation engines [6], [27], [28], [29], [30]. The accuracy of these models, however, is again constrained by the quality of the simulations. Additionally, their underlying methodologies still rely on displacement maps from a template, and can not produce different topologies.

Very recently, [4], [31], [32] have proposed strategies to model garments with topologies departing from the SMPL body mesh, like skirts or dresses. [4] does so by predicting generic skinning weights for the garment, independent from those of the body mesh. In [32], the garment is characterized by means of 2D sewing patterns, with a set of parameters that control its 3D shape. A limiting factor of these approaches is that they require training specific models for each type of garment, penalizing thus their practical use. [31] uses also sewing patterns to build a unified representation encoding different clothes. This representation, however, is too complex to allow controlling the generation process with just a few parameters. SMPLicit, in contrast, is able to represent using a single low-dimensional parametric model a large variety of clothes, which largely differ in their geometric properties, topology and cut.

Table 1 summarizes the main properties of the most recent generative cloth models we have discussed.

2.2 Reconstructing clothed humans from images

Most approaches for reconstructing 3D humans from images return the SMPL parameters, and thus only retrieve

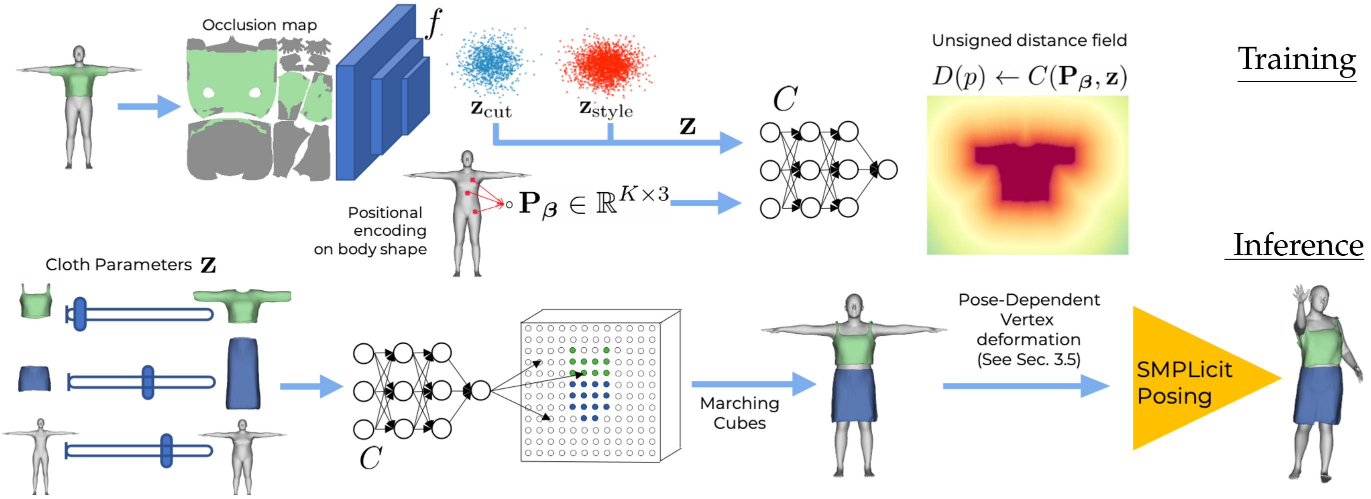

Occlusion map
f
$\mathbf{z}_{\text{cut}}$
$\mathbf{z}_{\text{style}}$
$\mathbf{z}$
C
Unsigned distance field
$D(p) \leftarrow C(\mathbf{P}_{\beta}, \mathbf{z})$
Training
Positional encoding on body shape
$\mathbf{P}_{\beta} \in \mathbb{R}^{K\times3}$
Cloth Parameters $\mathbf{Z}$
Inference
C
Marching Cubes
Pose-Dependent Vertex deformation (See Sec. 3.5)
SMPLicit Posing

Fig. 2. **Architecture of SMPLicit during training (top row) and inference (bottom row)**. At the core of SMPLicit lies an implicit-function network C that predicts unsigned distance from the query point $\mathbf{p}$ to the cloth iso-surface. The input $\mathbf{P}_{\beta}$ is encoded from $\mathbf{p}$ given a body shape. During training, we jointly train the network C as the latent space representation is created. We include an image encoder f that takes SMPL occlusion maps from ground truth garments and maps them to shape representations $\mathbf{z}_{\text{cut}}$, and a second component $\mathbf{z}_{\text{style}}$ trained as an auto-decoder [9]. At inference, we run the network $C(\cdot)$ for a densely sampled 3D space and use Marching Cubes to generate the 3D garment mesh. We finally pose each cloth vertex using the learnt skinning parameters [1] of the closest SMPL vertex.

TABLE 1
Comparison of our method with other works.

Method	Body Pose Variations	Body Shape Variations	Topology	Low-Dimension Latent Vector	Model is public
Santesteban [29]	✓	✓			
DRAPE [27]	✓	✓		✓	
Wang [30]		✓		✓	✓
GarNet [28]	✓	✓			✓
CAPE [5]	✓	✓		✓	✓
TailorNet [6]	✓	✓		✓	✓
BCNet [4]	✓	✓		✓	✓
Vidaurre [32]	✓	✓		✓	
Shen [31]	✓	✓	✓		✓
SMPLicit	✓	✓	✓	✓	✓

3D body meshes, but not clothing [2], [33], [34], [35], [36], [37], [38], [39], [40], [41], [42]. To reconstruct clothed people, a standard practice is to represent clothing geometry as an offset over the SMPL body mesh [3], [15], [16], [17], [43], [44], [45], [46], [47]. However, these approaches are prone to fail for loose garments that exhibit large displacements over the body. Non-parametric representations have also been explored for reconstructing arbitrary clothing topologies. These include approaches based on volumetric voxelizations [48], geometry images [49], bi-planar depth maps [50] or visual hulls [51]. Certainly, the most powerful model-free representations are those based on implicit functions [8], [52], [53], [54]. Recent approaches have also combined parametric and model-free representations, like SMPL plus voxels [55] and SMPL plus implicit functions [21], [54], [56].

While these approaches retrieve rich geometric detail, the resulting surfaces can not be controlled in both pose and clothing. SMPLicit is also built upon implicit functions, but our output contains multiple layers for the body and garments, and allows control over pose and clothing. Moreover, in this work we design a pipeline that builds on parametric and parametric-free models, extending SMPLicit to achieve high-definition and expressive meshes on in-the-wild images.

2.3 Cloth editing.

Several recent cloth editing approaches are focused on 2D, mostly tailored to virtual-try-on applications [57], [58], [59], [60], [61], [62], [63], [64], [65]. In this work, we are interested in doing such editing tasks, like swapping clothes between people, by extending initial approaches [66] to work in the wild. For this to be possible, it is necessary to segment the meshes of clothed humans into body and garment components. However, as mentioned above, most existing approaches on reconstructing clothed humans estimate cloth and body geometry as a single surface (mesh or voxels).

There exist a few works that provide rich cloth representations potentially applicable to editing tasks [16], [23], [43], [67], [68], although so far, none of them is ready to be used on single in-the-wild images. [6], [31], [32] propose generative models to control garment style (*e.g.* sleeve length or size), but do not demonstrate reconstructions from images. BCNet [4] and Multi-Garment Net [17] demonstrate reconstruction from images and show convincing cloth editing results, although they are evaluated on synthetic data [4] and require up to 8 video frames [17].

While SMPLicit has the capacity to perform 3D cloth editing on in-the-wild images, the editing results are also undermined by the lack of high-frequency details and color of the estimated 3D meshes. As we will show in the experimental results, our extension has the capacity to represent this detail when performing cloth editing tasks.

3 SMPLicit

We next describe SMPLicit's formulation, training scheme and how it can be used to interpolate between clothes. Fig. 2 shows the whole train and inference process.

3.1 Vertex Based SMPL vs SMPLicit

We build on the parametric human model SMPL [1] to generate clothes that adjust to a particular human body $M(\boldsymbol{\beta}, \boldsymbol{\theta})$, given its shape $\boldsymbol{\beta}$ and pose $\boldsymbol{\theta}$. SMPL is a function

$$M(\boldsymbol{\beta}, \boldsymbol{\theta}) : \boldsymbol{\theta} \times \boldsymbol{\beta} \mapsto \mathbf{V} \in \mathbb{R}^{3N}, \tag{1}$$

which predicts the N vertices $\mathbf{V}$ of the body mesh as a function of pose and shape. Our goal is to add a layer of clothing on top of SMPL. Prior work adds displacements [15], [16] on top of the body, or learns garment category-specific vertex-based models [6], [28]. The problem with predicting a fixed number of vertices is that different topologies (T-shirt vs open jacket) and extreme geometry changes (sleeve-less vs long-sleeve) can not be represented in a single model.

Our main contribution is *SMPLicit-core* (Sec.3.2-3.4), which departs from vertex models, and predicts clothing on T-pose with a learned implicit function

$$C(\mathbf{p}, \boldsymbol{\beta}, \mathbf{z}_{\text{cut}}, \mathbf{z}_{\text{style}}) \mapsto \mathbb{R}^{+}. \tag{2}$$

Specifically, we predict the *unsigned distance* to the clothing surface for a given point $\mathbf{p} \in \mathbb{R}^3$. By sampling enough points, we can reconstruct the desired mesh by thresholding the distance field and running Marching Cubes [69]. In addition to shape, we want to control the model with intuitive parameters $(\mathbf{z}_{\text{cut}}, \mathbf{z}_{\text{style}})$ representing the *cut* (*e.g.* , long vs short) and *style* (*e.g.* , hoodie vs not hoodie) of the clothing. Moreover, although it is not the focus of this paper, we also learn a point-based displacement field (Sec.3.5) to model pose-dependent deformations, and use SMPL skinning to pose the garments. The full model is called SMPLicit and outputs posed meshes $\mathcal{G}$ on top of the body:

$$C'(\boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{z}_{\text{cut}}, \mathbf{z}_{\text{style}}) \mapsto \mathcal{G}. \tag{3}$$

3.2 SMPLicit-Core Formulation

We explain here how we learn the *input representation*: two latent spaces to control clothing cut and style, and body shape to control fit; and the *output representation*. Together, these representations allow to generate and control garments of varied topology in a single model.

Clothing cut: We aim to control the output clothing cut, which we define as the body area occluded by clothing. To learn a latent space of cut, for each garment-body pair in the training set, we compute a UV body *occlusion image* denoted as $\mathbf{U}$. That is, we set every pixel in the SMPL body UV map to 1 if the corresponding body vertex is occluded by the garment, and 0 otherwise, see Fig. 2. Then we train an image encoder $f : \mathbf{U} \mapsto \mathbf{z}_{\text{cut}} \in \mathbb{R}^D$ to map the occlusion image to a latent vector $\mathbf{z}_{\text{cut}}$.

Clothing style: Different clothes might have the same body occlusion image $\mathbf{U}$, but their geometry can differ in tightness, low-frequency wrinkles or collar details. Thus we add another subset of parameters $\mathbf{z}_c$ which are initialized as a zero-vector and trained following the auto-decoder procedure from [9].

The set of parameters $\mathbf{z} = [\mathbf{z}_{\text{cut}}, \mathbf{z}_{\text{style}}] \in \mathbb{R}^N$ fully describes a garment cut and style.

Body shape: Since we want the model to vary with body shape, instead of learning a mapping from points to occupancy [9], [70], [71], we first encode points relative to the body. For each garment, we identify SMPL vertices that are close to ground truth models (*e.g.* torso vertices for upper-body clothes), and obtain K vertex clusters $\mathbf{v}_k \in \mathbb{R}^3$ that are distributed uniformly on the body in a T-pose. Then we map a 3D point in space $\mathbf{p} \in \mathbb{R}^3$ to a body relative encoding $\mathbf{P}_{\boldsymbol{\beta}} \in \mathbb{R}^{K \times 3}$ matrix, with rows storing the displacements to the clusters $\mathbf{P}_{\boldsymbol{\beta},k} = (\mathbf{p} - \mathbf{v}_k)$. This over-parameterized representation allows the network to reason about body boundaries, and we empirically observed superior performance compared to Euclidean or Barycentric distances. However, too many clusters might increase overfitting.

Output representation: One of the main challenges in learning a 3D generative clothing model is registering training garments [17], [20] (known to be a hard problem due to the lack of semantics and correspondences, and the diversity of body poses in raw data), which is necessary for vertex-based models [5], [6]. Implicit surface representations do not require registration, but necessitate closed surfaces for learning occupancies [71], [72] or signed distances [9], [73]. Since garments are open surfaces, we follow recent work [74] by predicting unsigned distance fields.

Given a query point $\mathbf{p}$, its positional encoding $\mathbf{P}_{\boldsymbol{\beta}}$ and cloth parameters $\mathbf{z}$, we train a decoder network $C(\mathbf{P}_{\boldsymbol{\beta}}, \mathbf{z}) \mapsto \mathbb{R}^{+}$ to predict the unsigned distance $D(\mathbf{p})$ to the ground truth cloth surface.

3.3 SMPLicit-core Training

Training entails learning the network parameters $\mathbf{w}_1$ of the clothing cut image encoder $\mathbf{z}_{\text{cut}} = f(\mathbf{U}; \mathbf{w_1})$, the style latent parameters $\mathbf{z}_{\text{style}}$ for each training example, and the parameters of the decoder network $C(\cdot; \mathbf{w}_2)$. For one training example, and one sampled point $\mathbf{p}$, we have the following loss:

$$\mathcal{L}_d = |C(\mathbf{P}_{\boldsymbol{\beta}}, f(\mathbf{U}; \mathbf{w}_1), \mathbf{z}_{\text{style}}; \mathbf{w}_2) - D(\mathbf{p})|. \tag{4}$$

During training, we sample points uniformly on a body bounding box, and also near the ground-truth surface, and learn a model of *all* garment categories jointly (we train separate models for upper-body, pants, skirts, shoes and hair though, because interpolation among them is not meaningful). At inference, we discard the encoder $f : \mathbf{U} \mapsto \mathbf{z}_{\text{cut}}$ network, and control SMPLicit directly with $\mathbf{z}_{\text{cut}}$.

To smoothly interpolate and generate new clothing, we constrain the latent space $\mathbf{z} = [\mathbf{z}_{\text{cut}}, \mathbf{z}_{\text{style}}]$ to be distributed normally with a second loss component $\mathcal{L}_z = |\mathbf{z}|$.

We also add zero mean identity covariance Gaussian noise $\mathbf{z}_\sigma \sim \mathcal{N}(\mathbf{0}, \sigma_n \mathbf{I})$ in the cloth representations before the forward pass during training, taking as input $C(\mathbf{P}_{\boldsymbol{\beta}}, \mathbf{z}+\mathbf{z}_\sigma)$, which proves specially helpful for garment types where we have a very small amount of data. The network C and the cloth latent spaces are jointly learned by minimizing a linear combination of the previously defined losses $\mathcal{L}_d + \lambda_z \mathcal{L}_z$, where λ_z is a hyper-parameter.

3.4 SMPLicit-core Inference

To generate a 3D garment mesh, we evaluate our network $C(\cdot)$ at densely sampled points around the body in a T-pose, and extract the iso-surface of the distance field at threshold t_d using Marching Cubes [69]. We set the hyperparameter $t_d = 0.1\,mm$ such that reconstructed garments do not have artifacts and are smooth. Since $C(\cdot)$ predicts unsigned distance and $t_d > 0$, the reconstructed meshes have a slightly larger volume than ground truth data; this is still better than closing the garments for training which requires voxelization. Thinner surfaces could be obtained with Neural Distance Fields [74], but we leave this for future work.

In summary, we can generate clothes that fit a body shape $\boldsymbol{\beta}$ by: (1) sampling $\mathbf{z} \sim \mathcal{N}(\mu * \mathbf{1}, \sigma * \mathbf{I})$, with a single mean and variance ($\mu, \sigma \in \mathbb{R}$) for all latent components obtained from the training latent spaces; (2) estimating the positional encoding $\mathbf{P}_{\boldsymbol{\beta}}$ for points around the T-pose and evaluating $C(\mathbf{P}_{\boldsymbol{\beta}}, \mathbf{z})$; (3) thresholding the distance field, and (4) running marching cubes to get a mesh.

3.5 Pose Dependent Deformation

SMPLicit-core can drape garments on a T-posed SMPL, but does not predict pose dependent deformations. Although *pose deformation is not the focus* of this work, we train a pose-dependent model to make SMPLicit readily available for animation applications. Similar to prior work [6], we learn the pose-deformation model on a canonical T-pose, and use SMPL learned skinning to pose the deformed mesh. Here, we leverage the publicly available TailorNet [6] dataset of simulated garments. Specifically, we learn a second network which takes body pose $\boldsymbol{\theta}$, a learnable latent variable $\mathbf{z}_{\boldsymbol{\theta}}$ and maps them to a per-point displacement $P : \mathbf{p} \times \boldsymbol{\theta} \times \mathbf{z}_{\boldsymbol{\theta}} \mapsto \mathbf{d} \in \mathbb{R}^3$. The latent space of $\mathbf{z}_{\boldsymbol{\theta}}$ is learned in an auto-decoding fashion like $\mathbf{z}_{\text{style}}$.

During training, since we are only interested in the displacement field on the surface, we only evaluate the model on points sampled along the cloth surface template on a T-Pose. We also encode the position of the input points $\mathbf{p} \mapsto \mathbf{P}_{\boldsymbol{\beta}}$ as a function of the body surface and train the model to minimize the difference between ground truth displacement and prediction.

During inference, we only evaluate P on the vertices of the recovered SMPLicit-core mesh, and displace them accordingly $\mathbf{p} \mapsto \mathbf{p} + \mathbf{d}$ to obtain a deformed mesh (still in the T-pose). Then we apply SMPL [1] to both body and deformed garment to pose them with $\boldsymbol{\theta}$. In particular, we deform each garment vertex using the skinning deformation of the closest SMPL body vertex. This process determines the SMPLicit function $C'(\cdot)$ defined in Eq. (3).

4 Fitting SMPLicit

In this section, we show the potential of SMPLicit for several computer vision and graphics applications. We show how SMPLicit can be fitted to 3D scans of dressed humans, or directly to in-the-wild images for perception tasks, taking advantage of the full differentiability of the predicted unsigned distance field with respect to cloth parameters.

4.1 Fitting SMPLicit to 3D scans of dressed people

Here we show how to fit SMPLicit to 3D scans of the Sizer dataset [75] which includes cloth segmentation. Intuitively, the main objective for fitting is to impose that SMPLicit-core evaluates to zero at the *unposed* scan points. We sample 3D points uniformly on the segmented scan upper-body and lower-body clothes, and also the 3D empty space around it. Let $\mathbf{q} \in \mathbb{R}^3$ be a point in the posed scan space, and let $\mathbf{d} = \text{dist}(\mathbf{q}, \mathcal{S})$ be the distance to the scan. Since SMPLicit-core is defined on the T-pose, we unpose $\mathbf{q}$ using the differentiable SMPL parameters (we associate to the closest SMPL vertex), and obtain the body relative encoding $\mathbf{P}_{\boldsymbol{\beta}}(\boldsymbol{\theta}, \boldsymbol{\beta})$, now as a function of shape *and* pose. Then we impose that our model C evaluates to the same distance at the encoding of the unposed point:

$$\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{z}) = |C(\mathbf{P}_{\boldsymbol{\beta}}(\boldsymbol{\theta}, \boldsymbol{\beta}), \mathbf{z}) - \mathbf{d}|. \tag{5}$$

We run the optimization for a number of iterations and for the cloth parameters of all garments the person is wearing. We also minimize the Chamfer distance between scan points and SMPL vertices, the MSE between SMPL joints and predicted scan joints, an SMPL prior loss [33], and a regularization term for $\boldsymbol{z}$. We use scheduling and first optimize the pose and shape, and finally all parameters jointly.

4.2 Fitting SMPLicit to images

Similar to SMPL for undressed bodies, SMPLicit provides the robustness and semantic knowledge to represent clothed people in images, especially in presence of severe occlusions, difficult poses, low-resolution images and noise. We first detect people and obtain an estimate of each person's pose and shape [40], as well as a 2D cloth semantic segmentation [76]. We then fit SMPLicit to every detection to obtain layered 3D clothing.

For every detected garment, we uniformly sample the space around the T-Posed SMPL, deform those points to the target SMPL pose ($\mathbf{p} \mapsto \bar{\mathbf{p}}$), and remove those that are occluded by the own body shape. Each *posed* point $\bar{\mathbf{p}}$ is then projected, falling into a semantic segmentation pixel (u, v) that matches its garment class $s_{\mathbf{p}} = 1$ or another class/background $s_{\mathbf{p}} = 0$. We have the following loss for a single point $\mathbf{p}$:

$$\mathcal{L}_I(\mathbf{z}) = \begin{cases} |C(\mathbf{P}_{\boldsymbol{\beta}}, \mathbf{z}) - \mathbf{d}_{\max}|, & \text{if } s_{\mathbf{p}} = 0 \\ \min_i |C(\mathbf{P}^i_{\boldsymbol{\beta}}, \mathbf{z})|, & \text{if } s_{\mathbf{p}} = 1 \end{cases} \tag{6}$$

When $s_{\mathbf{p}} = 0$ we force our model to predict the maximum cut-off distance $\mathbf{d}_{\max}$ of our distance fields (we force the point to be off-surface). When $s_{\mathbf{p}} = 1$ we force prediction to be zero distance (point in surface). Since many points $\bar{\mathbf{p}}^i$ (along the camera ray) might project to the same pixel (u, v), we take the $\min_i(\cdot)$ to consider only the point with minimum distance (closest point to the current garment surface estimate). Experimentally, this prevents thickening of clothes, which helps when we represent more than one cloth layer. We also add a regularization loss $\mathcal{L}_z = |\mathbf{z}|$ and optimize it jointly with $\mathcal{L}_I$.

5 LayerNet

We next describe our model LayerNet which takes advantage of SMPLicit's robustness and predicts deformations on its clothing. In this section, we first give an overview of the pipeline, and then describe each of its modules and the data pre-processing required to train it.

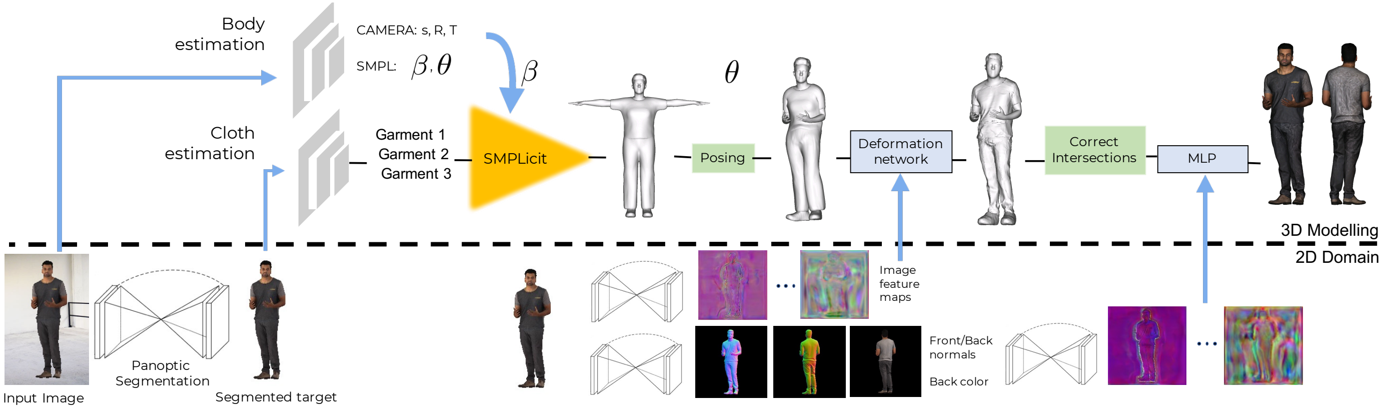


Fig. 3. **Overview of LayerNet.** Given a single input image, we first perform 2D person segmentation, predict SMPL [1] parameters and cloth categories from SMPLicit to obtain a coarse approximation of clothes. A novel Deformation Network that leverages image information allows to progressively refine the geometry of the clothes and recover fine details and wrinkles while adjusting to the body. A final branch estimates per-vertex normals and color, resulting in a high-definition person avatar with controllable garment meshes.

5.1 Pipeline Overview

After having described the SMPL and SMPLicit models, we can now formally summarize our pipeline (see Fig. 3). Given an input image $\mathbf{I}$ of a clothed and posed person, we initially use Frankmocap [40] to estimate SMPL/SMPL-X parameters $\{\boldsymbol{\beta}, \boldsymbol{\theta}\}$ encoding body shape and pose, respectively. The input image is also fed into a pre-trained 2D cloth segmentation network [76] that predicts cloth types m_g and the matrix of latent vectors $\mathbf{z}$. We then apply the SMPLicit mapping of Eq. 3 to recover the cloth meshes $\mathcal{G}$. However, recall that these meshes are still very coarse and lack of detail. Let us define $\mathbf{g}$ as a single vertex of $\mathcal{G}$. To recover the details, we train a deformation network $D(\cdot)$ that learns the mapping $D : \mathbf{g} \rightarrow \mathbf{g}'$, where $\mathbf{g}' \in \mathcal{G}'$ are high-definition cloth meshes. We then propose an optimization approach to reason about cloth layers and remove interpenetrations, and we finally predict per-vertex color and normals for both cloth and body.

Since LayerNet is based on SMPL predictions, we require an elaborated process of fitting SMPL, SMPLicit and re-posing scans for generating the data to train the pipeline to estimate body shape and garments from one single in-the-wild image. Data generation and LayerNet training procedure are described in Sections 5.2 and 5.3, respectively.

5.2 Data pre-processing

In order to learn to predict body shape and garment geometry directly from images, we will use both off-the-shelf segmentation networks pretrained on 2D real images, and other modules trained on 3D scans. The 3D scans will need to be pre-processed to extract ground truth parameters for the SMPL and SMPLicit models. For fitting SMPL, we follow a similar strategy as in [17], [20], [77] to minimize a loss combining three terms: L2 distance between SMPL joints and predicted scan joints , Chamfer distance between SMPL and scan vertices and an SMPL prior [33]. The 3D joints are up-lifted from 2D predicted joints in multiple views. We next describe the rest of pre-processing operations onto the training 3D scans.

Segmentation. The goal of this task is to assign to each vertex of the scan a garment/body label according to the 19 categories of the CIHP dataset [78], including hair, face, t-shirt, jacket, dress, pants, skirt or shoe. For this purpose, we render 360 views of the scan and perform semantic segmentation of each view using RP-R-CNN [25]. From the vertices visibility, we can back-project the 2D segmentation maps onto the 3D scan, and assign a per-vertex label according to a majority voting strategy.

SMPLicit fit. Given the 3D segmented scans, we next fit SMPLicit (*i.e*, we estimate the latent representations $\mathbf{z}_{\text{cut}}$ and $\mathbf{z}_{\text{style}}$ from Eq. 3) for each cloth type using the optimization proposed in Section 4.1. Additionally, we compute a Gaussian mixture model (GMM) over a large number of SMPLicit cloth mesh representations. This GMM is incorporated into the optimization of the SMPLicit parameters, which we observed to provide cleaner 3D cloth fits.

Cloth deformation. Given the SMPLicit fit encoded by $\mathbf{z}_{\text{cut}}$ and $\mathbf{z}_{\text{style}}$, we calculate the corresponding meshes $\mathcal{G}$ using Eq. 3. Then the mesh of each garment is non-rigidly deformed towards the corresponding mesh of the original segmented scan. This deformation is encoded by means of a per-vertex displacement $\mathbf{D}$, which is to be added to $\mathcal{G}$ to obtain a refined mesh $\mathcal{G}'$ as close as possible to the 3D scan, *i.e* $\mathcal{G}' = \mathcal{G} + \mathbf{D}$. The deformation $\mathbf{D}$ is obtained based on the following optimization:

$$\min_{\mathbf{D}} = \mathcal{L}_{CD} + \mathcal{L}_E + \mathcal{L}_L + |\mathbf{D}|_2 \,, \tag{7}$$

where $\mathcal{L}_{CD}$ is the Chamfer distance between the scan and $\mathcal{G}'$. $\mathcal{L}_E$ regularizes the edge lengths of $\mathcal{G}'$ and prevents departing from a mean length, that is, $\mathcal{L}_E = \sum_{l \in E_{\mathcal{G}'}} |l - \mu(E_{\mathcal{G}'})|_2$, where $E_{\mathcal{G}'}$ are the edges of $\mathcal{G}'$ and $\mu(\cdot)$ is the mean function. $\mathcal{L}_L$ regularizes over the curvature of neighboring vertices. We estimate the Laplace-Beltrami operator for each vertex and compute the graph Laplacian. $\mathcal{L}_L$ is then defined as the average norm of the product between the Laplacian and the position of each vertex, which represents the local amount of curvature. The last term $|\mathbf{D}|_2$ enforces small deformations.

Reposing scans. In order to extend the range of human poses of our training set, we re-pose each scan using different pose configurations from the AMASS dataset [79]. This reposing, however, needs to be done carefully to prevent self-intersections between arms and torso. For this purpose, we manually validate the fit of SMPL and SMPLicit in each of the scans and discard those where the optimization did not converge or the re-posing generated artifacts.

The output or these dataset processing operations is a large training set of 3D body scans with very diverse poses. Each 3D scan is annotated with the SMPL and SMPLicit parameters, plus a deformation map $\mathbf{D}$ that adds high-frequency details to the coarse SMPLicit meshes.

5.3 From images to 3D semantic reconstructions

We next describe the learning process of LayerNet, which follows a coarse-to-fine strategy and combines branches trained using real images with other branches trained with the registered data described above. Given an input image, the learning pipeline (see Fig. 2) involves the following steps: 1) Estimate coarse garment geometry; 2) Deform garments to retrieve fine details; 3) Reason about multi-layer outfits; 4) Estimate dense normals and texture.

Coarse cloth estimation. Following recent works [8], [52], [56] we initially use DeepLab [80] to perform 2D person segmentation on the input image $\mathbf{I}$. [80] is trained on large datasets of real images [81] which brings robustness to a variety of environmental conditions and body poses. Additionally, FrankMocap [40] is used to predict body shape and pose. This supports both SMPL [1] and SMPL-X [2]. We will use the latter, as it yields a better representation of the hands.

The segmented image is forwarded to MaskR-CNN [82], trained on DeepFashion2 [83], to predict cloth types. These labels, together with the SMPL parameters are then processed by SMPLicit, to obtain a coarse shape estimation $\mathcal{G}$ of the clothes. The garment labels are mapped to clusters computed over the GMM defined in 5.2, which is used to initialize SMPLicit and achieve faster and more robust convergence. SMPLicit allows recovering the following categories and their topological variations: upper clothes (vest/t-shirt/shirt/jacket), pants, skirts and shoes. Additionally, we train it to reconstruct different hair meshes. Long dresses and jumpsuits are modeled using two separate clothes: upper-cloth and skirt or pants.

Cloth deformation. We next refine the coarse cloth geometry, by adjusting it to the body shape and introducing high-frequency details consistent with the wrinkles in the input image $\mathbf{I}$. For this purpose, if we denote by $\mathbf{g}$ a specific vertex of $\mathcal{G}$, we learn a function $F_{disp}(\cdot)$ that predicts the Cartesian displacement $\mathbf{d}$ in camera coordinates:

$$F_{disp} : H(\mathbf{I}(\mathbf{g})), \pi(\mathbf{g}), Z(\mathbf{g}), B(\mathbf{g}), N(\mathbf{g}) \mapsto \mathbf{d}\ , \qquad (8)$$

where $H(\mathbf{I}(\mathbf{g}))$ are image features extracted at different resolutions using a Hourglass architecture [84], as in [8], [52]. $\pi(\cdot)$ and $Z(\cdot)$ represent the projection and normalized depth of the target vertex. We also condition F_{disp} with the undressed body geometry: $B(\cdot)$ is the distance to the SMPL surface and $N(\cdot)$ the normal direction of that vertex. During training, this process is supervised by the ground truth deformation maps $\mathbf{D}$ we described in Sect. 5.2.

Reasoning about 3D cloth layers. So far, we have not introduced constraints that prevent intersections between different body-cloth or cloth-cloth in multi-layer outfits (*e.g.* including T-shirts and jackets). To avoid body-cloth intersections we can exploit that SMPL and SMPLicit clothing are watertight and easily detect cloth vertices that fall inside the body or other clothing. These vertices are iteratively moved towards outside the body, along their normal direction, until the intersection is removed. For the multi-layer case, we first pre-define a specific garment ordering (*e.g.* T-shirts are "inside" jackets). Then, the garment that is in the exterior is slightly deformed, so as to remain outside an slightly inflated version of the body shape. The interior garment is adjusted so as not to intersect the original body shape.

Normals and texture prediction. Recovering cloth normals and color is essential to capture rich details and convey realism. There have been already several works in this direction [8], [52], [55], [56], from which we obtain inspiration in our method. Given the segmented input image, we train pix2pix [85] to predict front and back normal maps $\mathbf{N}_f, \mathbf{N}_b$, and back texture map $\mathbf{I}_b$. This training is performed in a fully supervised manner using renders of the original scans.

Using this information we then train a final function $F_{n,c}$ that for every vertex $\mathbf{g} \in \mathcal{G}$ predicts its normal direction $\mathbf{n}$ and RGB color $\mathbf{c}$:

$$F_{n,c} : H(\mathbf{I}, \mathbf{I}_b, \mathbf{N}_f, \mathbf{N}_b)(\mathbf{g}), \pi(\mathbf{g}), Z(\mathbf{g}), ID(\mathbf{g}) \mapsto \mathbf{n}, \mathbf{c} \qquad (9)$$

where $\mathbf{ID}(\mathbf{g})$ encodes the cloth type.

At inference, for those vertices that are visible, we directly assign them the interpolated normals and colors from the frontal predicted normal map $\mathbf{N}_f$ and input image $\mathbf{I}$, respectively. $F_{n,c}$ is used to assign color and normals to the rest of non-visible vertices.

To improve the ability of this prediction network to infer the color of the unseen parts (*e.g.* skin color when the whole body is covered by garments), during the training procedure of $F_{n,c}$ we swap 10% of the input cloth types $ID(\cdot)$ and assign other types that are present in the person (*e.g.* body). For these data points, we set the ground truth color to be the average color of the new target cloth type (*e.g.* average body color). This will guide the network to learn to predict colors of unseen parts, especially occluded body regions, for which we do not have groundtruth color in the training scans. By doing this we can swap clothes between a reference person wearing, e.g., a short-sleeve T-shirt and a target person wearing a jacket (right-most column in Fig.10).

6 Implementation Details

For the cloth latent space of SMPLicit, we set $|\mathbf{z}|= 18$ for upper-body, pants, skirts, hair and $|\mathbf{z}|= 4$ for shoes; the pose-dependent deformation parameters $|\mathbf{z}_\theta|= 128$, number of positional encoding clusters $K = 500$ and iso-surface threshold $t_d = 0.1$ mm. We clip the unsigned distance field $d_{max} = 10$mm. The implicit network architecture uses three 2-Layered MLPs that separately encode $\mathbf{z}_{\text{cut}}$, $\mathbf{z}_{\text{style}}$ and $\mathbf{P}_{\boldsymbol{\beta}}$ into an intermediate representation before a last 5-Layered MLP predicts the target unsigned distance field. SMPLicit is trained using Adam [88], with an initial learning rate 10^{-3}, $\beta_1 = 0.9$, $\beta_2 = 0.999$ for 1M iterations with linear LR decay after 0.5M iterations. We use $BS = 12$, $\sigma_n = 10^{-2}$ and refine a pre-trained ResNet-18 [89] as image encoder f.

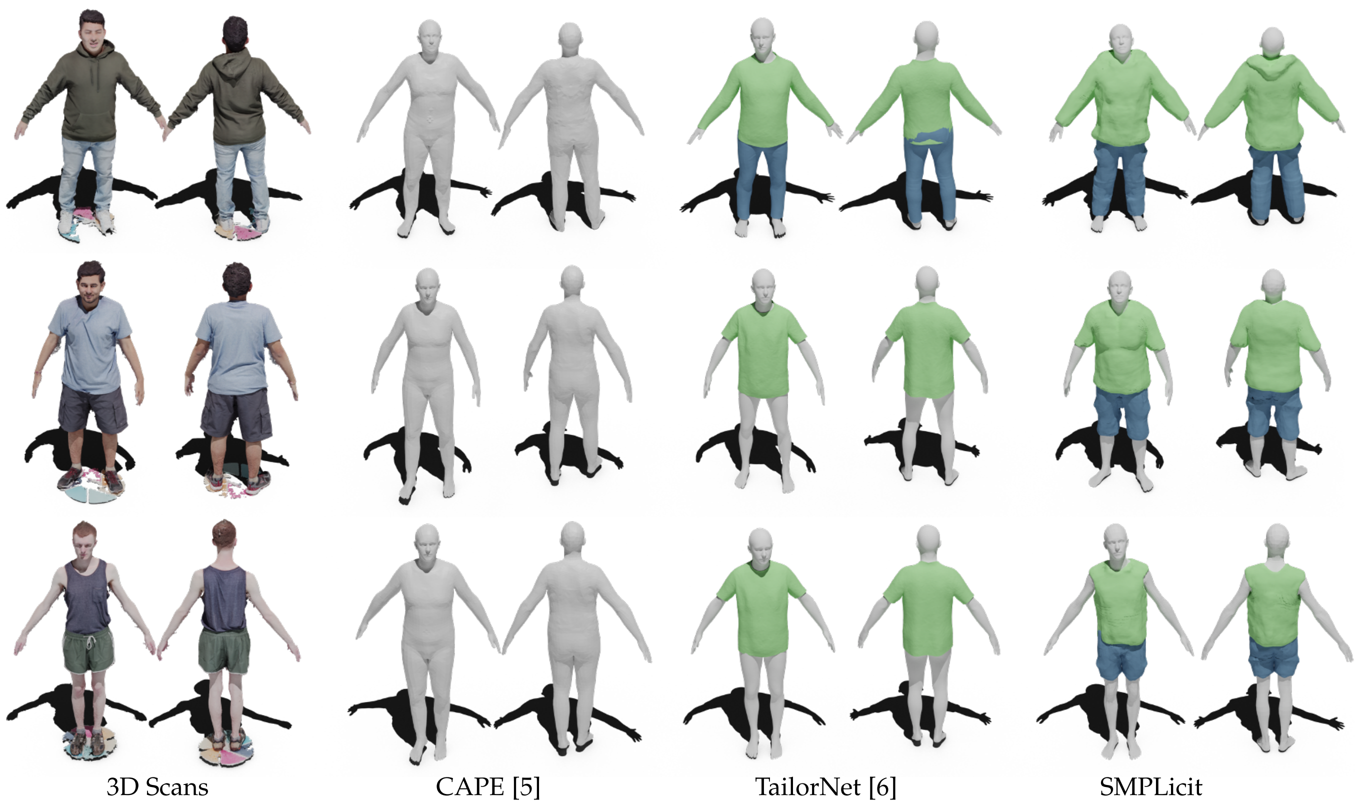


Fig. 4. **Fitting SMPLicit to 3D Scans of the Sizer Dataset [75]**. All three models achieve fitting results of approximately 1 mm of error. However, SMPLicit does this using a single model that can represent varying clothing topologies. For instance, it can model either hoodies (top row) and tank tops (third row) or long and short pants.

TABLE 2
Average Chamfer distance in cm in the AYXZ and BUFF datasets [86]. We use the SMPL-X [2] estimation from [40]. SMPLicit and PIFuHD [8] depict a similar quantitative performance, although recall that SMPLicit also provides a semantic disentanglement of the different garment meshes.

	AXYZ							BUFF Dataset [86]					
Method	Body	Hair	Upperclothes	Pants	Skirts	Shoes	Avg.	Subject 1	Subject 2	Subject 3	Subject 4	Subject 5	Avg.
PIFu [52]	4.7	3.0	2.8	**3.3**	**3.4**	3.5	3.3	2.6	3.7	2.8	2.7	2.7	2.8
PIFuHD [8]	4.4	**2.4**	2.8	3.5	4.8	**3.0**	3.0	**2.5**	**3.6**	2.6	**2.4**	**2.5**	**2.7**
Tex2Shape [3]	3.5	3.1	3.4	4.9	5.6	5.0	3.8	4.3	4.1	5.4	4.6	4.0	4.3
SMPL-X [2]	4.0	3.6	3.3	5.1	5.5	6.0	4.1	3.9	5.0	3.7	3.6	3.8	4.0
SMPLicit	**3.4**	2.5	**2.2**	3.5	3.9	4.4	**2.9**	2.6	3.8	**2.5**	2.6	2.8	2.9

TABLE 3
Capacity of SMPLicit for fitting 3D scans in comparison with TailorNet [6] and CAPE [5]. Note that we fit clothes on either long-sleeves or short-sleeves using a single model, while baselines have particularly trained for such topologies. All models achieve a remarkably accurate fitting within the segmented clothes of the original 3D scans.

	Distance to surface (mm)			
	Short Sleeves		Long Sleeves	
Method	Lower-Body	Upper-Body	Lower-Body	Upper-Body
Cape [5]	1.15	0.87	1.09	1.35
TailorNet [6]	-	0.32	0.48	0.41
SMPLicit	0.78	0.46	0.58	0.52

As [9], we use weight normalization [90] instead of batch normalization [91].

In LayerNet, the cloth estimation, deformation network, color/normal prediction network and normal and texture map prediction networks are trained for 4k, 4k, 1k and 5k epochs respectively, with batch size 6. All image encoders are trained using a pre-trained Resnet-50 [89] while pix2pix [85] and the Hourglass [84] architectures are trained from scratch with Adam [88] and learning rate 10^{-3}, decreasing linearly after reaching half training. The MLPs take as input the pixel representation and process it with 3 FC layers with weight normalization [90].

We render all images and train the networks at a 256×256 resolution. However, to obtain high-resolution reconstructions at test, we train and run the pix2pix at 512×512 which leads to more representative normal maps. The reconstructed meshes are also high-fidelity and contain around 500k.

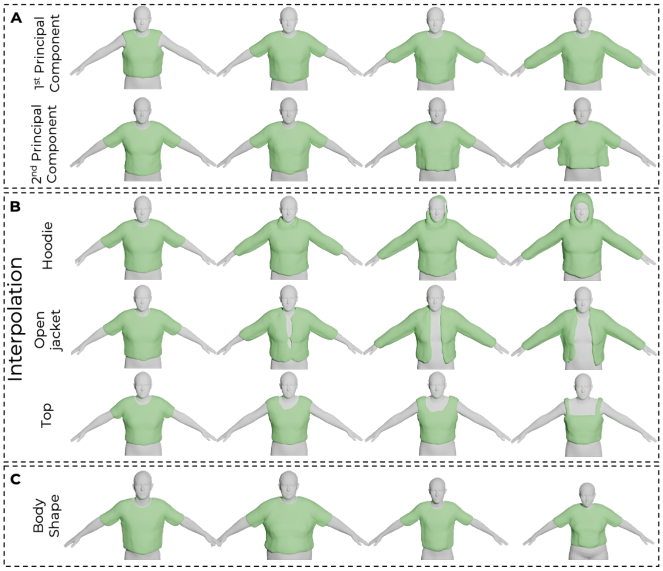


Fig. 5. **Overview of interpolations on latent space.** (A) effect of the two first principal components in the garment geometry. (B) SMPLicit can be used to interpolate from T-shirts to more complex clothes like hoodies, jackets or tops. (C) examples of retargeting an upper-body cloth to different human body shapes.

7 Experiments

This section first describes the datasets used to train SMPLicit and LayerNet. We then show how SMPLicit can interpolate smoothly between different cloth topologies, fit clothes from 3D scans and images with multiple people. We finally evaluate the performance of Layernet in the task of 3D reconstruction of clothed people, which provides a higher level of detail and allows 3D virtual try-on in images in-the-wild.

7.1 Training data

In order to train SMPLicit we resort to several publicly available datasets and augmentations. Concretely, we use the long-sleeved T-shirts (88797), pants (44265) and skirts (44435) from the BCNet Dataset [4]. This data is augmented by manually cutting different sleeve sizes on Blender [92], yielding a total of 800k T-shirts, 973k pants and 933k skirts. We also use 3D cloth models of jackets (23), jumpers (6), suits (2), hoodies (5), tops (12), shoes (28), boots (3) and sandals (3) downloaded from diverse public links of the Internet. We adjust these garments to a canonical body shape $\beta = \mathbf{0}$ and transfer them to randomly sampled body shapes during training, deforming each vertex using the shape-dependent displacement of the closest SMPL body vertex. For hair, we use the USC-HairSalon dataset [93], which contains 343 highly dense hair pointclouds, mostly of long hair. Given the large imbalance on the cloth categories for the upper-body, in each train iteration we sample one of the downloaded models with probability 0.5, otherwise we used one of the BCNet garments.

For training LayerNet, we use the 3D scans of the RenderPeople [94] and AXYZ [95] datasets and a subset of the SIZER Dataset [75]. After discarding not-successful fittings during the pre-processing operations, we keep a total of 346 scans, from which we used 318 for training and 28 for testing. Every scan was reposed to 20 different body poses of AMASS [79]. For rendering, we follow a similar approach as in [8], [52], [56], and consider 360 views per subject with a randomly sampled precomputed radiance transfer [96].

7.2 Generative properties

To provide control over cloth properties, we perform PCA on the latent space to discover directions which vary intuitive cloth characteristics, like sleeve-length, and identify cloth prototypes such as hoodies and tops.

PCA: The latent space $\mathbf{z} = [\mathbf{z}_{\text{cut}}, \mathbf{z}_{\text{style}}]$ of SMPLicit-core is small (4 to 18) in order to better disentangle cloth characteristics. We further perform PCA on the $\mathbf{z}_{\text{cut}}$ latent space and find that, for the upper and lower-body clothes, the first component controls sleeve length, while the second changes overall length (for upper-body garments), or the waist boundary height (for pants and skirts). Fig. 5-(A) shows the effect of the first 2 components for upper-garment. We also notice that perfect disentanglement from cut and style is not possible, as for example the network learns that tops tend to be more loose than t-shirts.

Prototypes: Furthermore, we identify cloth prototypes with interesting characteristics in the train data, such as open jackets, hoodies or tops, and store their average style latent space vectors $\mathbf{z}$. Fig. 5-(B) illustrates interpolation from a T-shirt to each of these prototypes; notice how SMPLicit is able to smoothly transition from short-sleeve to open jacket.

Body Shape: In Fig. 5-(C), we show results of re-targeting a single T-Shirt to significantly different body shapes.

7.3 Fitting SMPLicit to scans of dressed people

We applied SMPLicit-core to the problem of fitting 3D scans of clothed humans from the Sizer dataset [75], comparing against the recent TailorNet [6] and CAPE [5]. Since these methods have been specifically trained for long-sleeved and short-sleeved (for both shirt and pants), we only evaluate the performance of SMPLicit on these garments.

In Table 3 we report the reconstruction error (in mm) of the three methods. Note that in our case, we use a single model for modeling both short- and long-sleeves garments, while the other two approaches train independent models for each case. In any event, we achieve results which are comparable to Tailornet, and significantly better than CAPE. Qualitative results of this experiment are shown in Fig. 4. Note that CAPE does not provide specific meshes for the clothes, and only deforms SMPL mesh vertices. Tailornet yields specific meshes for shirts and long pants. SMPLicit, on the other hand, allows representing different topologies with a single model, from hoodies (first row) to a tank top (third row).

7.4 Fitting SMPLicit to images of clothed humans

Finally, using the optimization pipeline detailed in Sec. 4.2, we demonstrate that SMPLicit can also be fitted to images of clothed people and provide a 3D reconstruction of the body and clothes. Recall that to apply our method, we initially

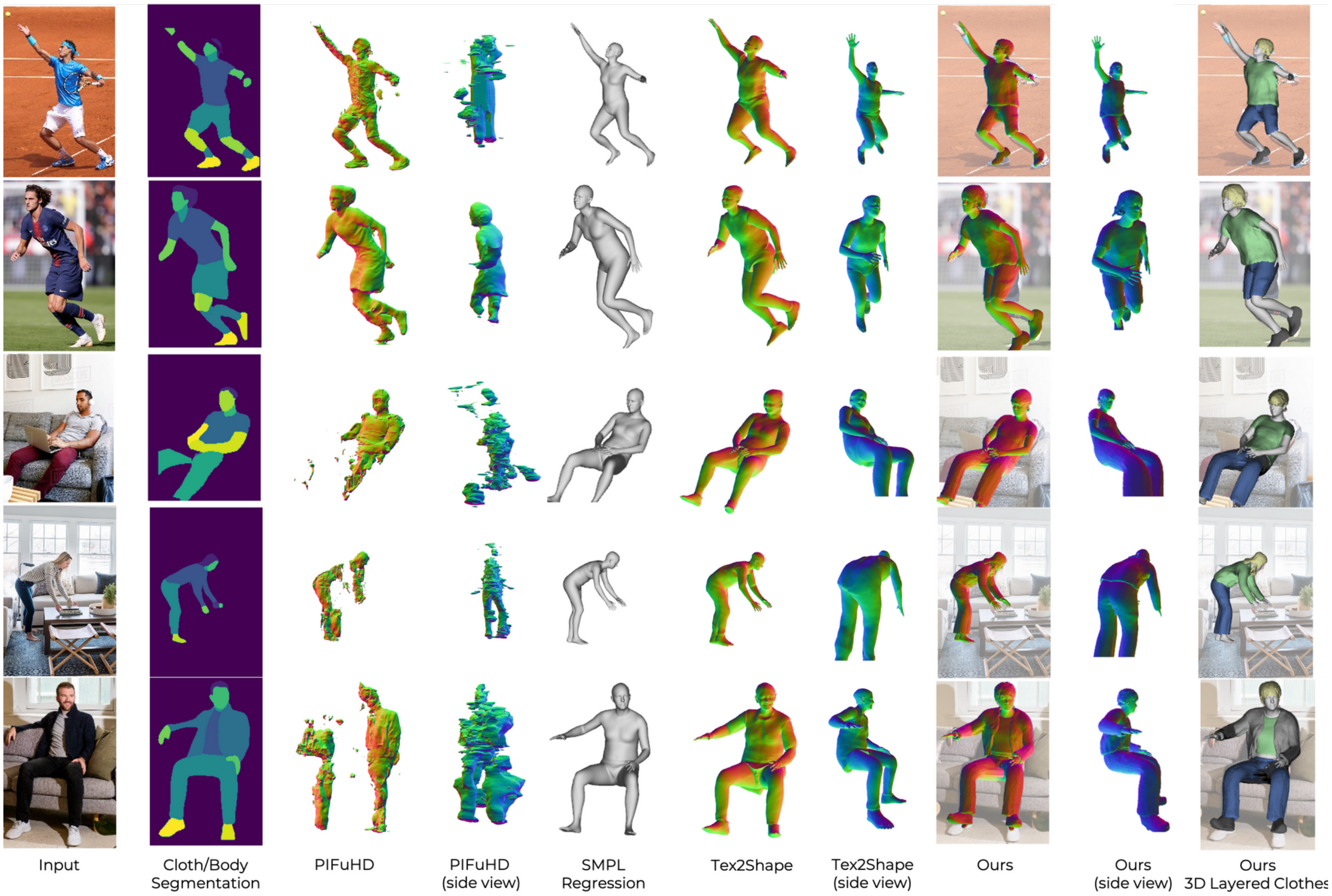


Fig. 6. **3D reconstruction of clothed humans,** in comparison to PIFuHD [8] and Tex2Shape [3]. SMPL regression is from [40].

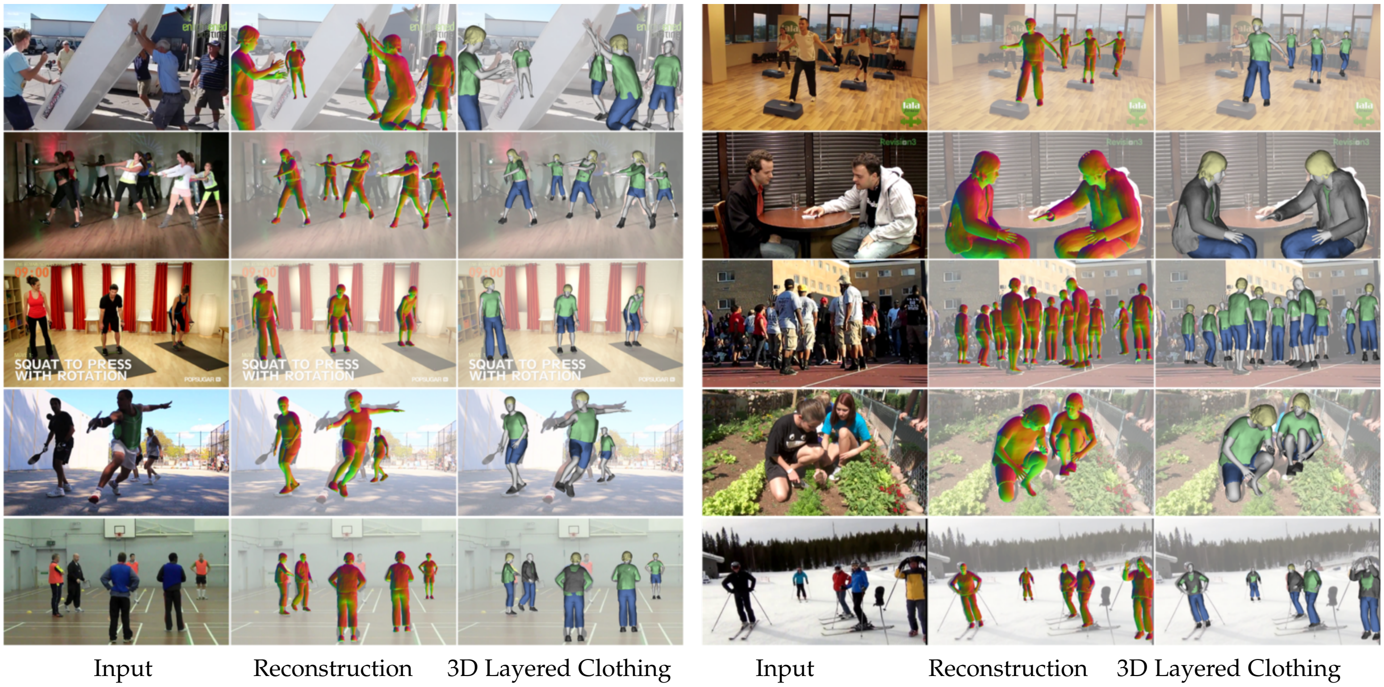


Fig. 7. **Fitting SMPLicit in multi-person images from the MPII [87] dataset.** SMPLicit can dress SMPL with a variety of clothes. Failure case in bottom-right example, where cloth semantic segmentation mixes shirts and jackets in most upper-bodies, and SMPLicit wrongly optimizes two similar intersecting jackets. Best viewed in color with zoom.

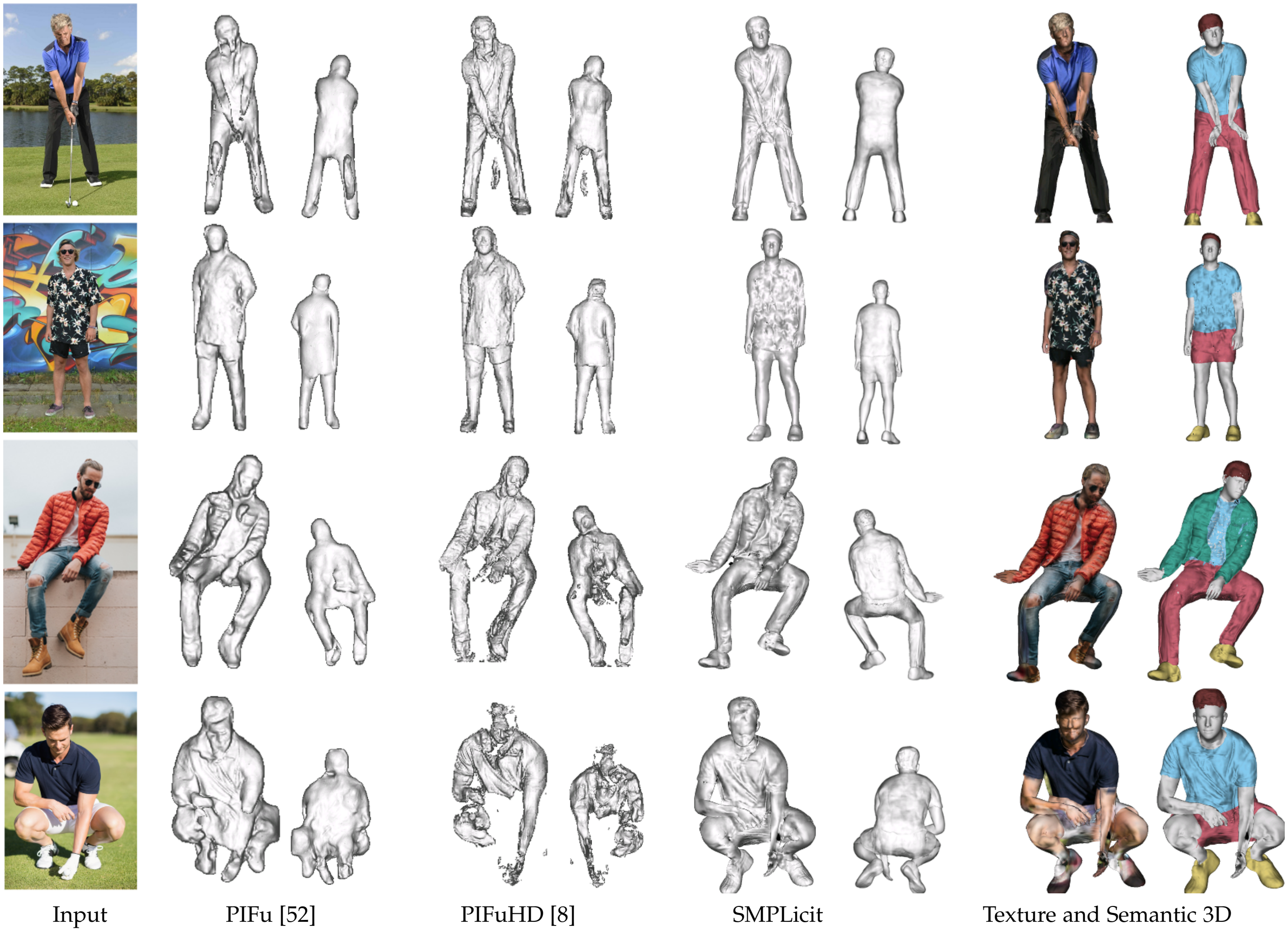


Fig. 8. **Front and back 3D reconstructions from SMPLicit and competing methods.** Both PIFu [52] and PIFuHD [8] do not rely on human body templates such as SMPL [1], which in some cases entails failing to reconstruct entire body limbs (first and second rows), or reconstructions looking unfeasible from the back, while SMPLicit is able to recover high-resolution details from challenging human poses.

use [40] to estimate SMPL parameters and [76] to obtain a pixel-wise segmentation of gross clothing labels (*i.e* upper-clothes, coat, hair, pants, skirts and shoes).

In Fig. 6 we show the results of this fitting on several images in-the-wild with a single person under arbitrary poses. We compare against PIFuHD [8] and Tex2Shape [3]. Before applying PIFuHD, we automatically remove the background using [97], as PIFuHD was trained with no- or simple backgrounds. Tex2Shape requires DensePose [98] segmentations, that map input pixels to the SMPL model. As shown in the Figure, the results of SMPLicit consistently improve other approaches, especially PiFuHD, which fails for poses departing from an upright position. Tex2Shape yields remarkably realistic results, but is not able to correctly retrieve the geometry of all the garments. Observe for instance, the example in the last row, where SMPLicit is capable of reconstructing clothing at different layers (T-shirt and jacket), even if layers are implicitly modelled and learnt from data. Interestingly, once the reconstruction is done, our approach can be used as a virtual try-on, changing garments' style and reposing the person's position. In Fig. 1 we show one such example. In Sec. 5 we also propose LayerNet to add more detail and color that enables realistic virtual try-on.

In Fig. 7 we go a step further in the task of 3D reconstruction, and show that SMPLicit can also be applied on challenging scenarios with multi-persons, taken from the MPII Dataset [87]. For this purpose we iterate over all SMPL detections [40], project the body model onto the image and mask out other people's segmentation. Note that in these examples, the model has to tackle extreme occlusions, but the combination of SMPLicit with powerful body pose detectors, like [40], and cloth segmentation algorithms, like [97], makes this task feasible. Of course, the overall success depends on each individual algorithm. For instance, in the bottom-right example of Fig. 7, errors in the segmentation labels are propagated to our reconstruction algorithm which incorrectly predicts two upper-body garments for certain individuals.

7.5 3D Reconstruction with LayerNet

We next evaluate the experiments regarding LayerNet, which involve the task of 3D reconstruction from monocular images where we compare our approach both quantitatively and qualitatively.

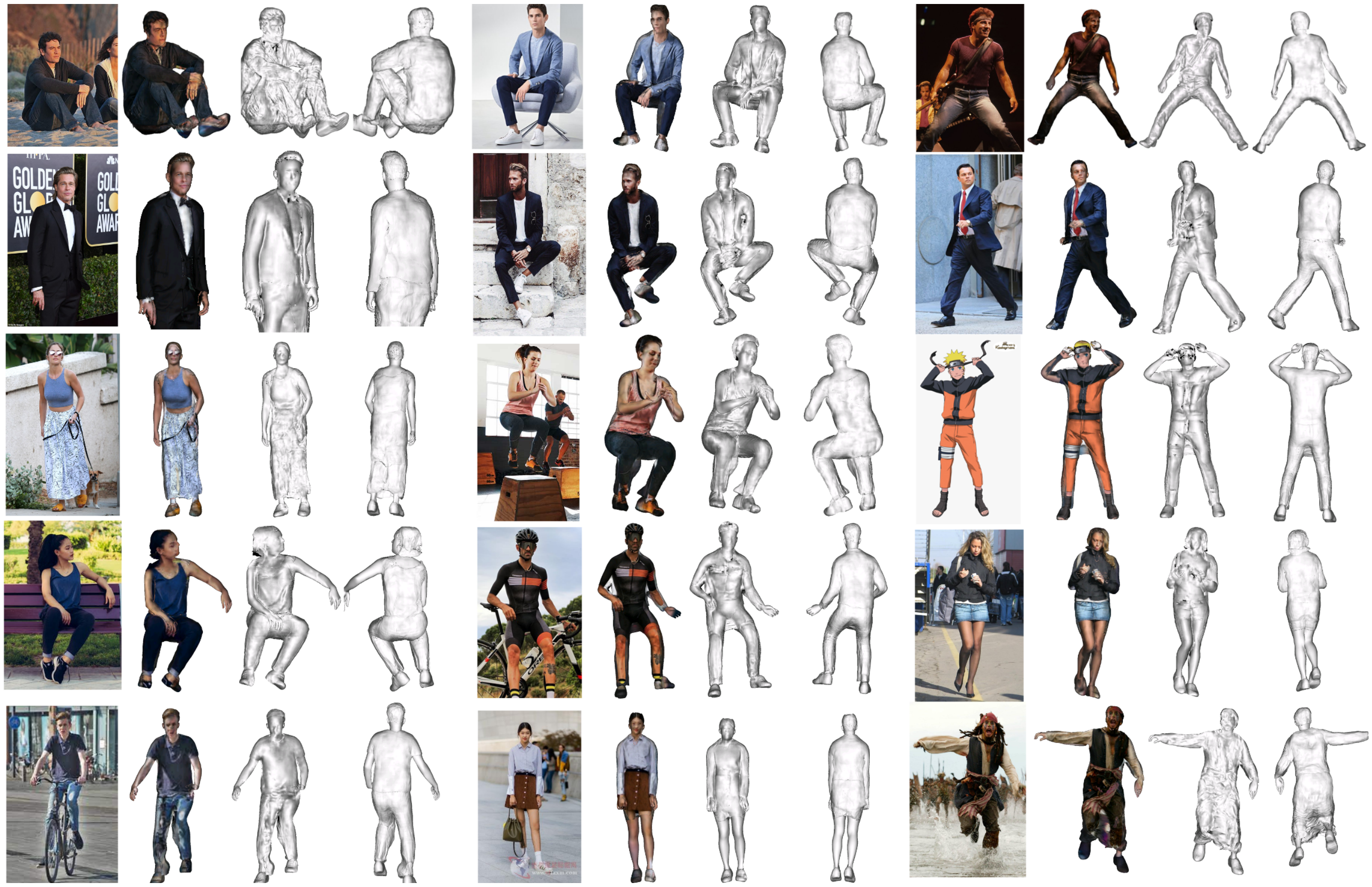


Fig. 9. **Pushing the boundaries of state-of-the-art 3D human reconstruction.** For every input image, we show the colored reconstruction, front and back 3D models of the person. The bottom-right example is a failure case where, due to challenging pose and confusing garment texture, the person is thought to wear a long skirt instead of long pants. Best viewed in zoom.

7.5.1 *Quantitative evaluation*

Even though LayerNet is effective for in-the-wild images, there are no such datasets with 3D ground truth to measure the performance. We therefore provide a quantitative evaluation on our synthetically rendered test set from [95] and renders from the BUFF dataset [86], consisting of 3D scans of people in upright body poses, and report the average Chamfer distance from the ground truth mesh to the reconstructions, in centimeters. We compare against several SOTA, including model-free methods: PIFu [52] and PIFuHD [8], and model-based methods: Tex2Shape [3] and SMPL-X [2]. Extending SMPLicit, LayerNet can use SMPL or SPML-X, and the reported results and figures are based on SMPL-X. Table 2 summarizes the results and demonstrates that LayerNet achieves very competitive reconstructions, close to the best current method, PIFuHD. Recall that on top of these results, LayerNet also provides meshes that are semantically interpretable as different garments. Additionally, thanks to the use of the underlying SMPL model, we can tackle more complex body poses than those of model-free approaches like PIFu or PIFuHD. In the following, we qualitatively demonstrate this.

7.5.2 *Qualitative evaluation*

Fig. 8 compares the performance of PIFu, PIFuHD and LayerNet on images in-the-wild. For a fair comparison, since PIFu and PIFuHD were trained with images without background, we apply the same segmentation mask for all methods. Note that LayerNet, in contrast to the other baselines, achieves consistent reconstructions for different poses and for the back of the body not seen in the image. Unlike LayerNet, note that previous methods do not generate layered reconstructions that include body and clothes.

In Fig. 9 we push the limits of our model and show that it is still robust to diverse outfits, complex body poses and even occluded body parts. This robustness is inherited from the prior of the SMPL model and the fact that we have trained our model with a large variety of re-posed 3D scans.

7.6 3D cloth transfer

We show qualitative results on the problem of cloth retargeting between two people. While SMPLicit does not capture details and cloth colors, we train LayerNet specifically for this purpose, obtaining promising results on the task of 3D virtual try-on. In this section, we apply LayerNet to a reference and a target image. We then unpose the reference cloth meshes to a T-pose configuration, and repose them to the target body shape and pose. Fig. 10 shows several examples that demonstrate we are able to retarget one, multiple or all clothes, even when reference and target images have different poses and body shapes. We also include some examples where the reference person is partially occluded,

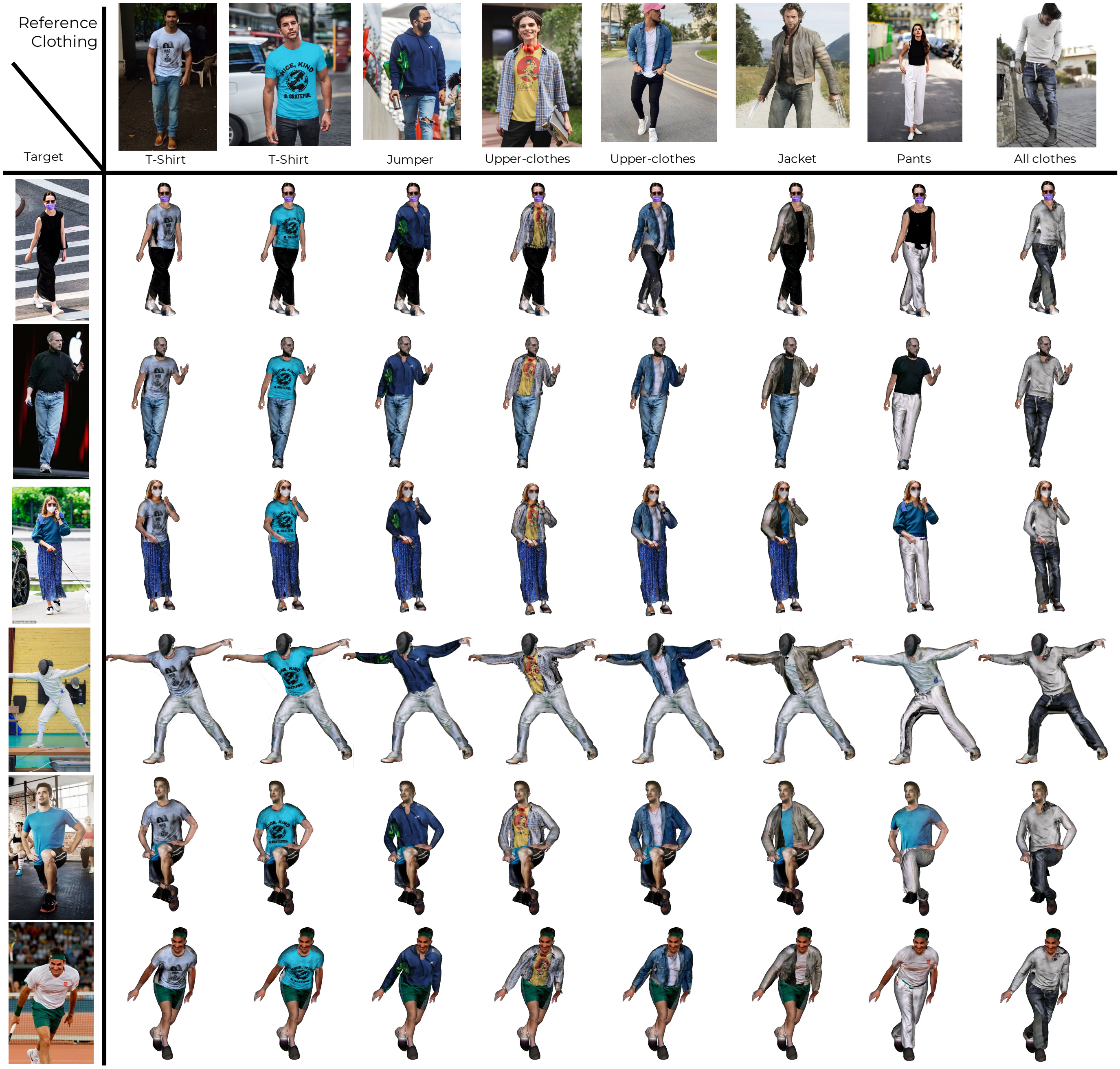


Fig. 10. **3D Cloth transfer.** Given a reference image of a person, we can transfer one, multiple or all their clothes to a new person. Our high-resolution approach enables reconstructing and retargeting clothes while preserving fine-grained details, logos and even text. SMPLicit also infers colors that are not visible in the target image, such as body color in the rightmost column, for those targets that were dressed in long sleeve outfits.

but our results are still consistent. It is also interesting that our pipeline is able to retrieve the textures of unseen parts. See for instance in Fig. 10 the pair fencer-soccer player. The skin tone of the fencer is also reasonably inpainted despite his body being completely occluded. This ability to hallucinate unseen parts is achieved thanks to the ID swapping process we have applied when training the texture prediction network in Sect. 3. Note that we are dealing with far more unconstrained scenarios than previous methods for 3D virtual try-on [68], which consider people in frontal view and mild poses, and a clean image for a single cloth. Our setup with in-the-wild-images, challenging poses and multi-layered garments, makes the problem far more complex.

7.7 Runtime

During inference, SMPLicit is very fast and can generate 3D garment models in less than a second, running marching cubes at a resolution of 128×128×128. In contrast, LayerNet requires representing clothing at a large resolution to recover details like wrinkles and colors accurately, leading to a slower runtime when running the full pipeline. This process takes approximately one minute to obtain a reconstruction given a monocular image on an Nvidia GTX 1080 Ti GPU. This runtime is comparable to works that require marching cubes at large resolutions [8]. The optimization to remove cloth interpenetrations takes an additional minute when multiple cloth layers are detected.

8 Conclusions

We have presented SMPLicit, a generative model for clothing able to represent different garment topologies and controlling their style and cut with just a few interpretable parameters. Our model is fully differentiable, making it possible to be integrated in several computer vision tasks. For instance, we showed that it can be readily used to fit 3D scans, and reconstruct clothed humans in images that pose a number of challenges, like multi-layered garments or strong body occlusions due to the presence of multiple people.

Furthermore, we extended the approach with LayerNet, a 3D reconstruction pipeline that simultaneously reconstructs high-resolution clothed humans and segments multi-layered garments from a single in-the-wild image. This pipeline combines the advantages of model-free and model-based approaches and inherits the flexibility of the former and the robustness of the latter.

8.1 Limitations

Our method still has some limitations which require further study. First, our method is trained on a relatively small dataset of clothing types and might not generalize well to garments that depart from the training distribution or very loose garments, especially considering the enormous diversity in cloth style. We believe that our framework could be trained with larger clothing databases, potentially combining both synthetically generated and real garments. Other open questions reside on how to model pose-dependent deformations in clothing, which would be necessary for animation or more realistic clothing try-on applications. Modeling illumination would enable relighting and realistic scene placement. Finally, recovering and modelling physical properties of clothing is an interesting open avenue for future work that would allow faithful reconstructions and animation.

Acknowledgments

This work is supported in part by the Spanish government with the project MoHuCo PID2020-120049RB-I00, the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - 409792180 (Emmy Noether Programme, project: Real Virtual Humans) and German Federal Ministry of Education and Research (BMBF): Tübingen AI Center, FKZ: 01IS18039A. Gerard Pons-Moll is a member of the Machine Learning Cluster of Excellence, EXC number 2064/1 – Project number 390727645. The project was made possible by funding from the Carl Zeiss Foundation. We thank NVIDIA for donating GPUs.

References

[1] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "Smpl: A skinned multi-person linear model," *ACM Transactions on Graphics (ToG)*, 2015.

[2] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. Osman, D. Tzionas, and M. J. Black, "Expressive body capture: 3d hands, face, and body from a single image," in *Computer Vision and Pattern Recognition Conference (CVPR)*, 2019.

[3] T. Alldieck, G. Pons-Moll, C. Theobalt, and M. Magnor, "Tex2shape: Detailed full human body geometry from a single image," in *International Conference on Computer Vision (ICCV)*, 2019.

[4] B. Jiang, J. Zhang, Y. Hong, J. Luo, L. Liu, and H. Bao, "Bcnet: Learning body and cloth shape from a single image," in *European Conference on Computer Vision (ECCV)*, 2020.

[5] Q. Ma, J. Yang, A. Ranjan, S. Pujades, G. Pons-Moll, S. Tang, and M. J. Black, "Learning to dress 3d people in generative clothing," in *Computer Vision and Pattern Recognition Conference (CVPR)*, 2020.

[6] C. Patel, Z. Liao, and G. Pons-Moll, "Tailornet: Predicting clothing in 3d as a function of human pose, shape and garment style," in *Computer Vision and Pattern Recognition Conference (CVPR)*, 2020.

[7] E. Corona, A. Pumarola, G. Alenyà, G. Pons-Moll, and F. Moreno-Noguer, "Smplicit: Topology-aware generative model for clothed people," in *Computer Vision and Pattern Recognition Conference (CVPR)*, 2021.

[8] S. Saito, T. Simon, J. Saragih, and H. Joo, "Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization," in *Computer Vision and Pattern Recognition Conference (CVPR)*, 2020.

[9] J. J. Park, P. Florence, J. Straub, R. Newcombe, and S. Lovegrove, "Deepsdf: Learning continuous signed distance functions for shape representation," in *Computer Vision and Pattern Recognition Conference (CVPR)*, 2019.

[10] D. Anguelov, P. Srinivasan, D. Koller, S. Thrun, J. Rodgers, and J. Davis, "Scape: shape completion and animation of people," *SIGGRAPH*, 2005.

[11] G. Pons-Moll, J. Romero, N. Mahmood, and M. J. Black, "Dyna: A model of dynamic human shape in motion," *SIGGRAPH*, 2015.

[12] E. Corona, A. Pumarola, G. Alenya, F. Moreno-Noguer, and G. Rogez, "Ganhand: Predicting human grasp affordances in multi-object scenes," in *Computer Vision and Pattern Recognition Conference (CVPR)*, 2020.

[13] H. Joo, T. Simon, and Y. Sheikh, "Total capture: A 3d deformation model for tracking faces, hands, and bodies," in *Computer Vision and Pattern Recognition Conference (CVPR)*, 2018.

[14] J. Romero, D. Tzionas, and M. J. Black, "Embodied hands: Modeling and capturing hands and bodies together," *ACM Transactions on Graphics (ToG)*, 2017.

[15] T. Alldieck, M. Magnor, B. L. Bhatnagar, C. Theobalt, and G. Pons-Moll, "Learning to reconstruct people in clothing from a single rgb camera," in *Computer Vision and Pattern Recognition Conference (CVPR)*, 2019.

[16] T. Alldieck, M. Magnor, W. Xu, C. Theobalt, and G. Pons-Moll, "Video based reconstruction of 3d people models," in *Computer Vision and Pattern Recognition Conference (CVPR)*, 2018.

[17] B. L. Bhatnagar, G. Tiwari, C. Theobalt, and G. Pons-Moll, "Multi-garment net: Learning to dress 3d people from images," in *International Conference on Computer Vision (ICCV)*, 2019.

[18] M. Habermann, W. Xu, M. Zollhoefer, G. Pons-Moll, and C. Theobalt, "Livecap: Real-time human performance capture from monocular video," *ACM Transactions on Graphics (TOG)*, 2019.

[19] A. Neophytou and A. Hilton, "A layered model of human body and garment deformation," in *International Conference on 3D Vision (3DV)*. IEEE, 2014.

[20] G. Pons-Moll, S. Pujades, S. Hu, and M. Black, "ClothCap: Seamless 4D clothing capture and retargeting," *SIGGRAPH*, 2017.

[21] B. L. Bhatnagar, C. Sminchisescu, C. Theobalt, and G. Pons-Moll, "Combining implicit function learning and parametric models for 3d human reconstruction," in *European Conference on Computer Vision (ECCV)*. Springer, August 2020.

[22] T. Yu, Z. Zheng, K. Guo, J. Zhao, Q. Dai, H. Li, G. Pons-Moll, and Y. Liu, "Doublefusion: Real-time capture of human performances with inner body shapes from a single depth sensor," in *Computer Vision and Pattern Recognition Conference (CVPR)*, 2018.

[23] T. Yu, Z. Zheng, Y. Zhong, J. Zhao, Q. Dai, G. Pons-Moll, and Y. Liu, "Simulcap: Single-view human performance capture with cloth simulation," in *Computer Vision and Pattern Recognition Conference (CVPR)*, 2019.

[24] E. Corona, G. Alenya, A. Gabas, and C. Torras, "Active garment recognition and target grasping point detection using deep learning," *Pattern Recognition*, vol. 74, 2018.

[25] J. Yang, J.-S. Franco, F. Hétroy-Wheeler, and S. Wuhrer, "Analyzing clothing layer deformation statistics of 3d human motions," in *European Conference on Computer Vision (ECCV)*, 2018.

[26] Z. Lahner, D. Cremers, and T. Tung, "Deepwrinkles: Accurate and realistic clothing modeling," in *European Conference on Computer Vision (ECCV)*, 2018.

[27] P. Guan, L. Reiss, D. A. Hirshberg, A. Weiss, and M. J. Black, "Drape: Dressing any person," *ACM Transactions on Graphics (ToG)*, 2012.

[28] E. Gundogdu, V. Constantin, A. Seifoddini, M. Dang, M. Salzmann, and P. Fua, "Garnet: A two-stream network for fast and accurate 3d cloth draping," in *International Conference on Computer Vision (ICCV)*, 2019.

[29] I. Santesteban, M. A. Otaduy, and D. Casas, "Learning-based animation of clothing for virtual try-on," in *Computer Graphics Forum*, vol. 38. Wiley Online Library, 2019.

[30] T. Y. Wang, D. Ceylan, J. Popovic, and N. J. Mitra, "Learning a shared shape space for multimodal garment design," *arXiv preprint arXiv:1806.11335*, 2018.

[31] Y. Shen, J. Liang, and M. C. Lin, "Gan-based garment generation using sewing pattern images," in *European Conference on Computer Vision (ECCV)*, 2020.

[32] R. Vidaurre, I. Santesteban, E. Garces, and D. Casas, "Fully Convolutional Graph Neural Networks for Parametric Virtual Try-On," *Computer Graphics Forum*, 2020.

[33] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black, "Keep it smpl: Automatic estimation of 3d human pose and shape from a single image," in *European Conference on Computer Vision (ECCV)*. Springer, 2016.

[34] P. Guan, A. Weiss, A. O. Balan, and M. J. Black, "Estimating human shape and pose from a single image," in *International Conference on Computer Vision (ICCV)*. IEEE, 2009.

[35] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik, "End-to-end recovery of human shape and pose," in *Computer Vision and Pattern Recognition Conference (CVPR)*, 2018.

[36] N. Kolotouros, G. Pavlakos, M. J. Black, and K. Daniilidis, "Learning to reconstruct 3d human pose and shape via model-fitting in the loop," in *International Conference on Computer Vision (ICCV)*, 2019.

[37] N. Kolotouros, G. Pavlakos, and K. Daniilidis, "Convolutional mesh regression for single-image human shape reconstruction," in *Computer Vision and Pattern Recognition Conference (CVPR)*, 2019.

[38] C. Lassner, J. Romero, M. Kiefel, F. Bogo, M. J. Black, and P. V. Gehler, "Unite the people: Closing the loop between 3d and 2d human representations," in *Computer Vision and Pattern Recognition Conference (CVPR)*, 2017.

[39] M. Omran, C. Lassner, G. Pons-Moll, P. Gehler, and B. Schiele, "Neural body fitting: Unifying deep learning and model based human pose and shape estimation," in *International Conference on 3D Vision (3DV)*. IEEE, 2018.

[40] Y. Rong, T. Shiratori, and H. Joo, "Frankmocap: Fast monocular 3d hand and body motion capture by regression and integration," *arXiv preprint arXiv:2008.08324*, 2020.

[41] D. Smith, M. Loper, X. Hu, P. Mavroidis, and J. Romero, "Facsimile: Fast and accurate scans from an image in less than a second," in *International Conference on Computer Vision (ICCV)*, 2019.

[42] X. Xu, H. Chen, F. Moreno-Noguer, L. A. Jeni, and F. D. la Torre, "3d human shape and pose from a single low-resolution image," in *European Conference on Computer Vision (ECCV)*, 2020.

[43] T. Alldieck, M. Magnor, W. Xu, C. Theobalt, and G. Pons-Moll, "Detailed human avatars from monocular video," in *International Conference on 3D Vision (3DV)*. IEEE, 2018.

[44] H. Onizuka, Z. Hayirci, D. Thomas, A. Sugimoto, H. Uchiyama, and R.-i. Taniguchi, "Tetratsdf: 3d human reconstruction from a single image with a tetrahedral outer shell," in *Computer Vision and Pattern Recognition Conference (CVPR)*, 2020.

[45] V. Lazova, E. Insafutdinov, and G. Pons-Moll, "360-degree textures of people in clothing from a single image," in *International Conference on 3D Vision (3DV)*, sep 2019.

[46] A. Sayo, H. Onizuka, D. Thomas, Y. Nakashima, H. Kawasaki, and K. Ikeuchi, "Human shape reconstruction with loose clothes from partially observed data by pose specific deformation," in *Pacific-Rim Symposium on Image and Video Technology*. Springer, 2019.

[47] H. Zhu, X. Zuo, S. Wang, X. Cao, and R. Yang, "Detailed human shape estimation from a single image by hierarchical mesh deformation," in *Computer Vision and Pattern Recognition Conference (CVPR)*, 2019.

[48] G. Varol, D. Ceylan, B. Russell, J. Yang, E. Yumer, I. Laptev, and C. Schmid, "Bodynet: Volumetric inference of 3d human body shapes," in *European Conference on Computer Vision (ECCV)*, 2018.

[49] A. Pumarola, J. Sanchez-Riera, G. Choi, A. Sanfeliu, and F. Moreno-Noguer, "3dpeople: Modeling the geometry of dressed humans," in *International Conference on Computer Vision (ICCV)*, 2019.

[50] V. Gabeur, J.-S. Franco, X. Martin, C. Schmid, and G. Rogez, "Moulding humans: Non-parametric 3d human shape estimation from single images," in *International Conference on Computer Vision (ICCV)*, 2019.

[51] R. Natsume, S. Saito, Z. Huang, W. Chen, C. Ma, H. Li, and S. Morishima, "Siclope: Silhouette-based clothed people," in *Computer Vision and Pattern Recognition Conference (CVPR)*, 2019.

[52] S. Saito, Z. Huang, R. Natsume, S. Morishima, A. Kanazawa, and H. Li, "Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization," in *International Conference on Computer Vision (ICCV)*, 2019.

[53] J. Chibane, T. Alldieck, and G. Pons-Moll, "Implicit functions in feature space for 3d shape reconstruction and completion," in *Computer Vision and Pattern Recognition Conference (CVPR)*. IEEE, jun 2020.

[54] E. Corona, M. Zanfir, T. Alldieck, E. G. Bazavan, A. Zanfir, and C. Sminchisescu, "Structured 3d features for reconstructing controllable avatars," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 16 954–16 964.

[55] Z. Zheng, T. Yu, Y. Wei, Q. Dai, and Y. Liu, "Deephuman: 3d human reconstruction from a single image," in *International Conference on Computer Vision (ICCV)*, 2019.

[56] Z. Huang, Y. Xu, C. Lassner, H. Li, and T. Tung, "Arch: Animatable reconstruction of clothed humans," in *Computer Vision and Pattern Recognition Conference (CVPR)*, 2020.

[57] T. Issenhuth, J. Mary, and C. Calauzènes, "Do not mask what you do not need to mask: a parser-free virtual try-on," *European Conference on Computer Vision (ECCV)*, 2020.

[58] H. Yang, R. Zhang, X. Guo, W. Liu, W. Zuo, and P. Luo, "Towards photo-realistic virtual try-on by adaptively generating-preserving image content," in *Computer Vision and Pattern Recognition Conference (CVPR)*, June 2020.

[59] A. Neuberger, E. Borenstein, B. Hilleli, E. Oks, and S. Alpert, "Image based virtual try-on network from unpaired data," in *Computer Vision and Pattern Recognition Conference (CVPR)*, June 2020.

[60] Z. Zhu, Z. Xu, A. You, and X. Bai, "Semantically multi-modal image synthesis," in *Computer Vision and Pattern Recognition Conference (CVPR)*, 2020.

[61] R. Yu, X. Wang, and X. Xie, "Vtnfp: An image-based virtual try-on network with body and clothing feature preservation," in *International Conference on Computer Vision (ICCV)*, October 2019.

[62] X. Han, X. Hu, W. Huang, and M. R. Scott, "Clothflow: A flow-based model for clothed person generation," in *International Conference on Computer Vision (ICCV)*, 2019.

[63] B. Wang, H. Zheng, X. Liang, Y. Chen, L. Lin, and M. Yang, "Toward characteristic-preserving image-based virtual try-on network," in *European Conference on Computer Vision (ECCV)*, 2018.

[64] A. Raj, P. Sangkloy, H. Chang, J. Hays, D. Ceylan, and J. Lu, "Swapnet: Image based garment transfer," in *European Conference on Computer Vision (ECCV)*. Springer, 2018.

[65] X. Han, Z. Wu, Z. Wu, R. Yu, and L. S. Davis, "Viton: An image-based virtual try-on network," in *Computer Vision and Pattern Recognition Conference (CVPR)*, 2018.

[66] A. Mir, T. Alldieck, and G. Pons-Moll, "Learning to transfer texture from clothing images to 3d humans," in *Computer Vision and Pattern Recognition Conference (CVPR)*. IEEE, June 2020.

[67] Z. Su, W. Wan, T. Yu, L. Liu, L. Fang, W. Wang, and Y. Liu, "Mulaycap: Multi-layer human performance capture using a monocular video camera," *arXiv preprint arXiv:2004.05815*, 2020.

[68] F. Zhao, Z. Xie, M. Kampffmeyer, H. Dong, S. Han, T. Zheng, T. Zhang, and X. Liang, "M3d-vton: A monocular-to-3d virtual try-on network," in *International Conference on Computer Vision (ICCV)*, 2021.

[69] W. E. Lorensen and H. E. Cline, "Marching cubes: A high resolution 3d surface construction algorithm," *SIGGRAPH*, 1987.

[70] Z. Chen and H. Zhang, "Learning implicit fields for generative shape modeling," in *Computer Vision and Pattern Recognition Conference (CVPR)*, 2019.

[71] L. Mescheder, M. Oechsle, M. Niemeyer, S. Nowozin, and A. Geiger, "Occupancy networks: Learning 3d reconstruction in function space," in *Computer Vision and Pattern Recognition Conference (CVPR)*, 2019.

[72] A. Pumarola, E. Corona, G. Pons-Moll, and F. Moreno-Noguer, "D-nerf: Neural radiance fields for dynamic scenes," in *Computer Vision and Pattern Recognition Conference (CVPR)*, 2021.

[73] M. Runz, K. Li, M. Tang, L. Ma, C. Kong, T. Schmidt, I. Reid, L. Agapito, J. Straub, S. Lovegrove, and R. Newcombe, "Frodo: From detections to 3d objects," in *Computer Vision and Pattern Recognition Conference (CVPR)*, 2020.

[74] J. Chibane, G. Pons-Moll *et al.*, "Neural unsigned distance fields for implicit function learning," *Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

[75] G. Tiwari, B. L. Bhatnagar, T. Tung, and G. Pons-Moll, "Sizer: A dataset and model for parsing 3d clothing and learning size sensitive 3d clothing," in *European Conference on Computer Vision (ECCV)*, 2020.

[76] L. Yang, Q. Song, Z. Wang, M. Hu, C. Liu, X. Xin, W. Jia, and S. Xu, "Renovating parsing r-cnn for accurate multiple human parsing," in *European Conference on Computer Vision (ECCV)*, 2020.

[77] V. Lazova, E. Insafutdinov, and G. Pons-Moll, "360-degree textures of people in clothing from a single image," in *International Conference on 3D Vision (3DV)*. IEEE, 2019.

[78] K. Gong, X. Liang, Y. Li, Y. Chen, M. Yang, and L. Lin, "Instance-level human parsing via part grouping network," in *European Conference on Computer Vision (ECCV)*, 2018.

[79] N. Mahmood, N. Ghorbani, N. F. Troje, G. Pons-Moll, and M. J. Black, "Amass: Archive of motion capture as surface shapes," in *International Conference on Computer Vision (ICCV)*, 2019.

[80] B. Cheng, M. D. Collins, Y. Zhu, T. Liu, T. S. Huang, H. Adam, and L.-C. Chen, "Panoptic-deeplab: A simple, strong, and fast baseline for bottom-up panoptic segmentation," in *Computer Vision and Pattern Recognition Conference (CVPR)*, 2020.

[81] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European Conference on Computer Vision (ECCV)*. Springer, 2014.

[82] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *International Conference on Computer Vision (ICCV)*, 2017.

[83] Y. Ge, R. Zhang, X. Wang, X. Tang, and P. Luo, "Deepfashion2: A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images," in *Computer Vision and Pattern Recognition Conference (CVPR)*, 2019.

[84] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *European Conference on Computer Vision (ECCV)*. Springer, 2016.

[85] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Computer Vision and Pattern Recognition Conference (CVPR)*, 2017.

[86] C. Zhang, S. Pujades, M. J. Black, and G. Pons-Moll, "Detailed, accurate, human shape estimation from clothed 3d scan sequences," in *Computer Vision and Pattern Recognition Conference (CVPR)*, 2017.

[87] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, "2d human pose estimation: New benchmark and state of the art analysis," in *Computer Vision and Pattern Recognition Conference (CVPR)*, 2014.

[88] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[89] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Computer Vision and Pattern Recognition Conference (CVPR)*, 2016.

[90] T. Salimans and D. P. Kingma, "Weight normalization: A simple reparameterization to accelerate training of deep neural networks," in *Conference on Neural Information Processing Systems (NeurIPS)*, 2016.

[91] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

[92] Blender Online Community, *Blender - a 3D modelling and rendering package*, Blender Foundation. [Online]. Available: http://www.blender.org

[93] L. Hu, C. Ma, L. Luo, and H. Li, "Single-view hair modeling using a hairstyle database," *ACM Transactions on Graphics (ToG)*, 2015.

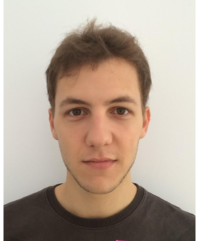
[94] "Renderpeople," http://renderpeople.com/.

[95] "Axyz," http://secure.axyz-design.com/.

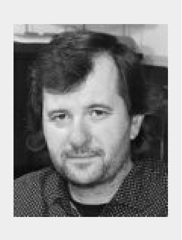
[96] P.-P. Sloan, J. Kautz, and J. Snyder, "Precomputed radiance transfer for real-time rendering in dynamic, low-frequency lighting environments," in *CGIT*, 2002.

[97] "Remove background," https://www.remove.bg/.

[98] R. Alp Güler, N. Neverova, and I. Kokkinos, "Densepose: Dense human pose estimation in the wild," in *Computer Vision and Pattern Recognition Conference (CVPR)*, 2018.



Enric Corona received the Bachelor and MSc degrees from the Technical University of Catalonia (UPC) in 2016 and Universitat Pompeu Fabra (UPF) in 2017 respectively. He is currently a PhD student at the Institut de Robòtica i Informàtica Industrial in Barcelona. His doctoral research investigates computer vision applications for sensing human body pose and motion, that are convenient for applications for Human-Robot-Interaction applications.



Guillem Alenyà is Researcher and Director at the Institut de Robotica i Informàtica Industrial (IRI), a joint centre of the Spanish Scientific Research Council (CSIC) and Polytechnic University of Catalonia (UPC). He received a PhD degree from UPC in 2007 and has been visitor at KIT-Karlsruhe (2007), INRIA-Grenoble (2008) and BRL-Bristol(2016). He has participated in numerous scientific projects involving image understanding, next-best-view, rule learning, and plan execution tasks. He is currently a local PI of the EU MSCA training network SOCRATES, a local co-PI of the EU project IMAGINE, and a Leader of numerous technological transfer projects.

Gerard Pons-Moll is Professor at the University of Tuebingen, head of the Emmy Noether independent research group "Real Virtual Humans", senior researcher at the Max Planck for Informatics (MPII) in Saarbrucken, Germany. His research lies at the intersection of computer vision, computer graphics and machine learning – with special focus on analyzing people in videos, and creating virtual human models by "looking" at real ones. His work has received several awards including the prestigious Emmy Noether Grant (2018), a Google Faculty Research Award (2019), a Facebook Reality Labs Faculty Award (2018), and the German Pattern Recognition Award (2019)His work got Best Papers Awards and nomminations at CVPR'20, CVPR'21, ECCV'22. He serves regularly as area chair for the top conferences in vision and graphics (CVPR, ICCV, ECCV, Siggraph).

Francesc Moreno-Noguer is a Research Scientist of the Spanish National Research Council at the Institut de Robotica i Informatica Industrial. His research interests are in Computer Vision and Machine Learning, with topics including human shape and motion estimation, 3D reconstruction of rigid and nonrigid objects and camera calibration. He received the Polytechnic University of Catalonia's Doctoral Dissertation Extraordinary Award, several best paper awards (ECCV 2018 Honorable mention, ICCV 2017 workshop in Fashion, Intl. Conf. on Machine Vision applications 2016), outstanding reviewer awards at ECCV 2012 and CVPR 2014, 2021, and Google and Amazon Faculty Research Awards in 2017 and 2019, respectively. He has (co)authored over 150 publications in refereed journals and conferences (including 10 IEEE Transactions on PAMI, 5 Intl. Journal of Computer Vision, 28 CVPR, 11 ECCV and 9 ICCV).