SMPLicit: Topology-aware Generative Model for Clothed People

Enric Corona[1] Albert Pumarola[1] Guillem Alenyà[1] Gerard Pons-Moll[2,3] Francesc Moreno-Noguer[1]
[1]Institut de Robòtica i Informàtica Industrial, CSIC-UPC, Barcelona, Spain
[2]University of Tübingen, Germany, [3]Max Planck Institute for Informatics, Germany

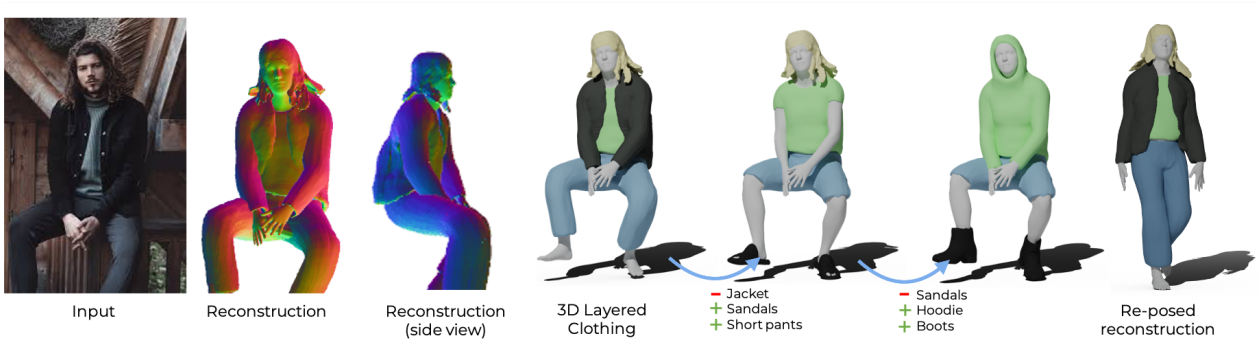

Figure 1: We introduce SMPLicit, a fully differentiable generative model for clothed bodies, capable of representing garments with different topology. The four figures on the left show the application of the model to the problem of 3D body and cloth reconstruction from an input image. We are able to 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.

Abstract

*In this paper we introduce SMPLicit, a novel generative model to jointly represent body pose, shape and clothing geometry. 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 to open jackets), while controlling other properties like the garment size or tightness/looseness. We show our model to be applicable to a large variety of garments including T-shirts, hoodies, jackets, shorts, pants, skirts, shoes and even hair. The representation flexibility of SMPLicit builds upon an implicit model conditioned with the SMPL human body parameters and a learnable latent space which is semantically interpretable and aligned with the clothing attributes. The proposed model is fully differentiable, allowing for its use into larger end-to-end trainable systems. In the experimental section, we demonstrate SMPLicit can be readily used for fitting 3D scans and for 3D reconstruction in images of dressed people. In both cases we are able to go beyond state of the art, by retrieving complex garment geometries, handling situations with multiple clothing layers and providing a tool for easy outfit editing. To stimulate further research in this direction, we will make our code and model publicly available at* http://www.iri.upc.edu/people/ecorona/smplicit/.

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 [35, 45], 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 [4, 26, 37, 44]. 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 ar-

chitecture 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 [37, 44]. 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.

To summarize, the main contributions of our work are: (1) A generative model that is capable of representing clothes under different topology; (2) A low-dimensional and semantically interpretable latent vector for controlling clothing style and cut; (3) A model that can be conditioned on human pose, shape and garment style/cut; (4) A fully differentiable model for easy integration with deep learning; (5) A versatile approach that can be applied to both 3D scan fitting and 3D shape reconstruction from images in the wild; (6) A 3D reconstruction algorithm that produces controllable and editable surfaces.

2. Related work

Cloth modeling 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* and *3D reconstruction of clothed humans*, the two 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 [7, 47, 16, 27, 35, 45, 51], a number of approaches aim to learn clothing models from real data, obtained using multiple images [1, 3, 4, 9, 21], 3D scans [40, 46, 8] or RGBD sensors [68, 69]. 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 [15, 66, 32] have problems to generalize beyond the deformation patterns of the training data. [37] 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.

Method	Body Pose Variations	Body Shape Variations	Topology	Low-Dimension Latent Vector	Model is public
Santesteban [57]	✓	✓			
DRAPE [18]	✓	✓		✓	
Wang [64]		✓		✓	✓
GarNet [20]	✓	✓			✓
TailorNet [44]	✓	✓		✓	✓
BCNet [26]	✓	✓		✓	✓
Vidaurre [63]	✓	✓		✓	
Shen [59]	✓	✓	✓		✓
SMPLicit	✓	✓	✓	✓	✓

Table 1: **Comparison of our method with other works.**

An alternative to the use of real data is to learn clothing models using data from physics simulation engines [18, 20, 44, 57, 64]. 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, [26, 59, 63] have proposed strategies to model garments with topologies departing from the SMPL body mesh, like skirts or dresses. [26] does so by predicting generic skinning weights for the garment, independent from those of the body mesh. In [63], 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. [59] 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 3D body meshes, but not clothing [11, 19, 28, 30, 31, 33, 41, 45, 52, 60, 65]. To reconstruct clothed people, a standard practice is to represent clothing geometry as an offset over the SMPL body mesh [1, 2, 3, 42, 4, 9, 34, 58, 71]. However, these approaches are prone to fail for loose garments that

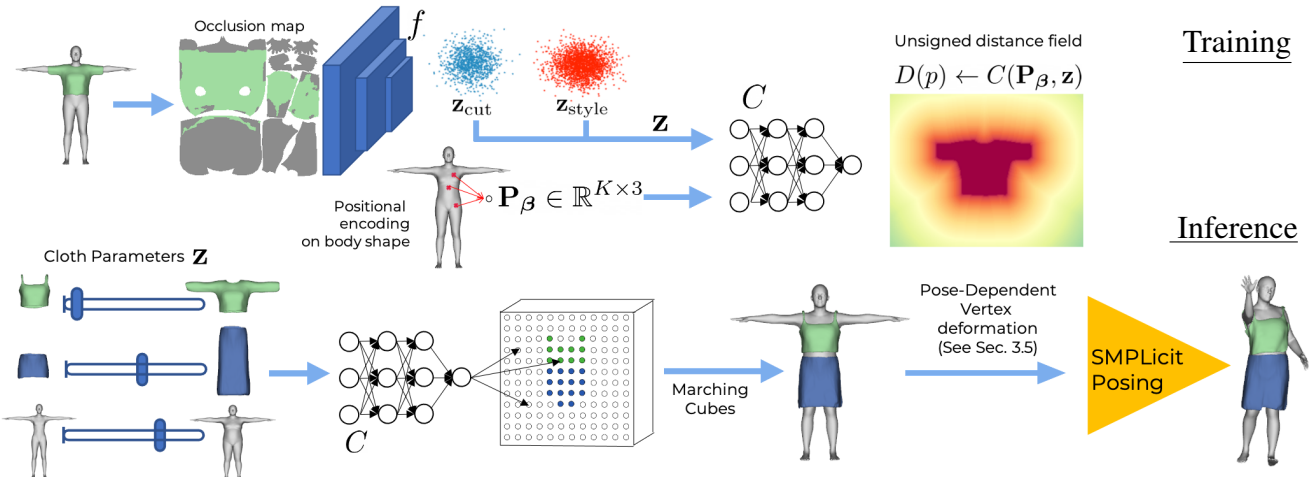

Figure 2: **Architecture of SMPLicit during training (top row) and inference (bottom row).** At the core of SMPLicit lies an implicit-function network C that predicts unsigned distance from the query point $\mathbf{p}$ to the cloth iso-surface. The input $\mathbf{P}_{\boldsymbol{\beta}}$ is encoded from $\mathbf{p}$ 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 [43]. 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 [35] of the closest SMPL vertex.

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 [62], geometry images [49], bi-planar depth maps [17] or visual hulls [39]. Certainly, the most powerful model-free representations are those based on implicit functions [54, 55, 13]. Recent approaches have also combined parametric and model-free representations, like SMPL plus voxels [70] and SMPL plus implicit functions [8, 24].

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.

3. SMPLicit

We next describe the SMPLicit 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 [35] 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 [1, 3] on top of the body, or learns garment category-specific vertex-based models [20, 44]. 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 [36]. 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 [43].

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 [12, 38, 43], 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.

Output representation: One of the main challenges in learning a 3D generative clothing model is registering training garments [9, 46] (known to be a hard problem), which is necessary for vertex-based models [37, 44]. Implicit surface representations do not require registration, but necessitate closed surfaces for learning occupancies [38, 48] or signed distances [43, 53]. Since garments are open surfaces, we follow recent work [14] 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:

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

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 [36]. 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 [14], 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 [44], 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 [44] 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

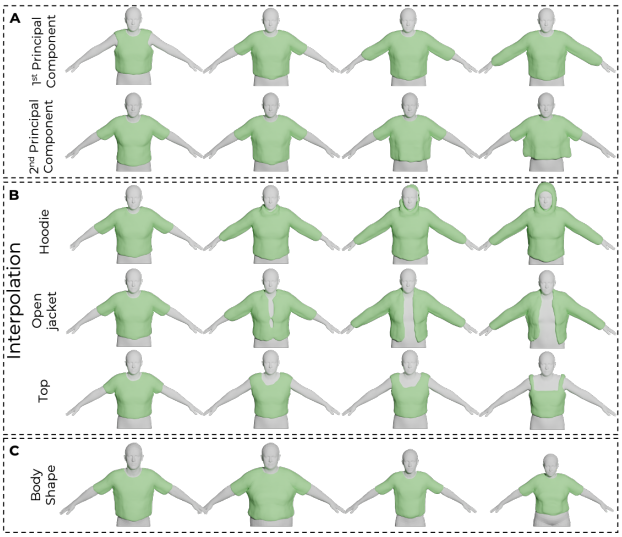

Figure 3: **Overview of interpolations on latent space.** (A) effect of the two first principal components in the garment geometry. (B) SMPLicit can be used to interpolate from T-shirts to more complex clothes like hoodies, jackets or tops. (C) examples of retargeting an upper-body cloth to different human body shapes.

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 [35] 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. Applications of SMPLicit

In this section, we show the potential of SMPLicit for several computer vision and graphics applications. We demonstrate how to interpolate garments in the latent space and edit their cut and style. We then 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. Generative properties

To provide control to the user, we perform PCA on the latent space to discover directions which vary intuitive cloth properties, 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. 3-(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.

	Distance to surface (mm)			
	Short Sleeves		Long Sleeves	
Method	Lower-Body	Upper-Body	Lower-Body	Upper-Body
Cape [37]	1.15	0.87	1.09	1.35
TailorNet [44]	-	0.32	0.48	0.41
SMPLicit	0.78	0.46	0.58	0.52

Table 2: **Capacity of SMPLicit for fitting 3D scans in comparison with TailorNet [44] and CAPE [37].** 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.

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. 3-(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. 3-(C), we show results of re-targeting a single T-Shirt to significantly different body shapes.

4.2. Fitting SMPLicit to 3D scans of dressed people

Here we show how to fit SMPLicit to 3D scans of the Sizer dataset [61] 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:

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

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 [11], and a

Figure 4: **Fitting SMPLicit to 3D Scans of the Sizer Dataset [61]**. 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.

regularization term for $\boldsymbol{z}$. We use scheduling and first optimize the pose and shape, and finally all parameters jointly. See the Supp. Mat. for more details.

4.3. Fitting SMPLicit to images

Similar to SMPL for undressed bodies, SMPLicit provides the robustness and semantic knowledge to reconstruct 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 [52], as well as a 2D cloth semantic segmentation [67]. 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 $\mathbf{p}$:

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

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 reconstruct 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. Implementation details

We next describe the main implementation details. Further information is provided in the Suppl. Material and in the code that will be made publicly available.

For the cloth latent space, 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 [29], 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 $BS = 12$,

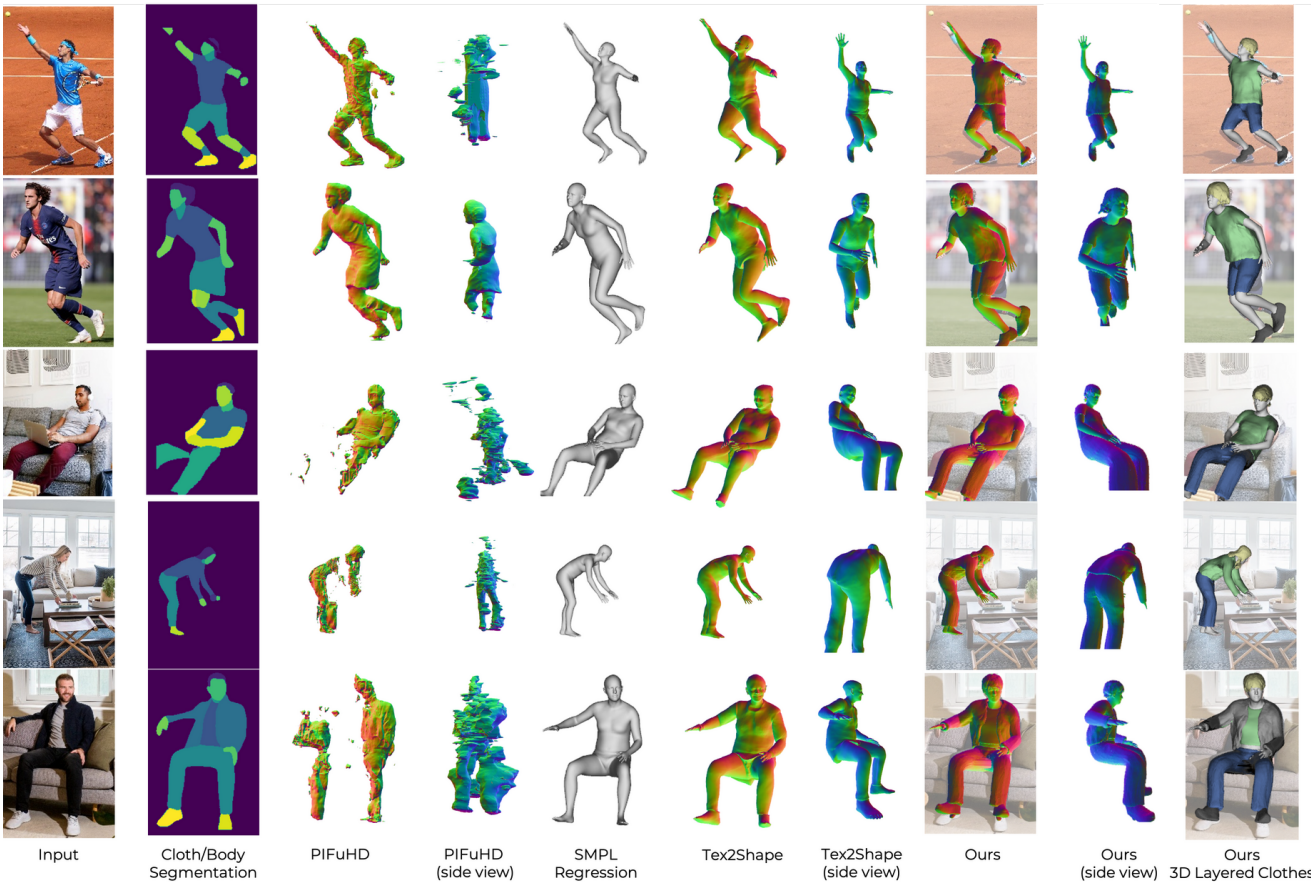

Figure 5: **3D reconstruction of clothed humans,** in comparison to PIFuHD [55] and Tex2Shape [4]. SMPL regression is from [52].

$\sigma_n = 10^{-2}$ and refine a pre-trained ResNet-18 [22] as image encoder f. As [43], we use weight normalization [56] instead of batch normalization [25].

6. Experiments

This section first describes the datasets used to train SMPLicit, and then we show results for fitting 3D scans and 3D reconstruction of dressed people from images.

6.1. 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 [26]. This data is augmented by manually cutting different sleeve sizes on Blender [10], 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 [23], which contains 343 highly dense hair point-clouds, 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 the pose-dependent deformation model of Sec. 3.5, we use cloth simulations from TailorNet [44], which consist of 200 shirt and pants instances. For the remaining garments, except for shoes (which we do not further deform), we train a deformation model parameterized only by $\mathbf{z}_\theta$, given manually warps generated using Blender.

6.2. 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 [61], comparing against the recent TailorNet [44] and CAPE [37]. 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 2 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 bet-

Input Reconstruction 3D Layered Clothing Input Reconstruction 3D Layered Clothing

Figure 6: **Fitting SMPLicit in multi-person images from the MPII [6] 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.

ter 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).

6.3. 3D reconstruction of clothed humans

Finally, using the optimization pipeline detailed in Sec. 4.3, 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 use [52] to estimate SMPL parameters and [67] to obtain a pixel-wise segmentation of gross clothing labels (*i.e.* upper-clothes, coat, hair, pants, skirts and shoes).

In Fig. 5 we show the results of this fitting on several images in-the-wild with a single person under arbitrary poses. We compare against PIFuHD [55] and Tex2Shape [4]. Before applying PIFuHD, we automatically remove the background using [50], as PIFuHD was trained with no- or simple backgrounds. Tex2Shape requires DensePose [5] 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). 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 Fig. 6 we go a step further, and show that SMPLicit can also be applied on challenging scenarios with multi-persons, taken from the MPII Dataset [6]. For this purpose we iterate over all SMPL detections [52], 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 [52], and cloth segmentation algorithms, like [50], makes this task feasible. Of course, the overall success depends on each individual algorithm. For instance, in the bottom-right example of Fig. 6, errors in the segmentation labels are propagated to our reconstruction algorithm which incorrectly predicts two upper-body garments for certain individuals.

7. Conclusion

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. Ad-

ditionally, our generative model can be used in geometric content edition tasks to *e.g*. dynamically change the type of garment attributes, opening the door to build novel virtual try-on systems.

Acknowledgements: This work is supported in part by an Amazon Research Award and by the Spanish government with the projects HuMoUR TIN2017-90086-R and María de Maeztu Seal of Excellence MDM-2016-0656. Gerard Pons-Moll is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - 409792180 (Emmy Noether Programme, project: Real Virtual Humans)

References

[1] Thiemo Alldieck, Marcus Magnor, Bharat Lal Bhatnagar, Christian Theobalt, and Gerard Pons-Moll. Learning to reconstruct people in clothing from a single rgb camera. In *CVPR*, 2019. 2, 3

[2] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Detailed human avatars from monocular video. In *3DV*. IEEE, 2018. 2

[3] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Video based reconstruction of 3d people models. In *CVPR*, 2018. 2, 3

[4] Thiemo Alldieck, Gerard Pons-Moll, Christian Theobalt, and Marcus Magnor. Tex2shape: Detailed full human body geometry from a single image. In *ICCV*, 2019. 1, 2, 7, 8

[5] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *CVPR*, 2018. 8

[6] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *CVPR*, 2014. 8

[7] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. Scape: shape completion and animation of people. *SIGGRAPH*, 2005. 2

[8] Bharat Lal Bhatnagar, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Combining implicit function learning and parametric models for 3d human reconstruction. In *European Conference on Computer Vision (ECCV)*. Springer, August 2020. 2, 3

[9] Bharat Lal Bhatnagar, Garvita Tiwari, Christian Theobalt, and Gerard Pons-Moll. Multi-garment net: Learning to dress 3d people from images. In *ICCV*, 2019. 2, 4

[10] Blender Online Community. *Blender - a 3D modelling and rendering package*. Blender Foundation, Blender Institute, Amsterdam, 2020. 7

[11] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *ECCV*. Springer, 2016. 2, 5

[12] Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. In *CVPR*, 2019. 4

[13] Julian Chibane, Thiemo Alldieck, and Gerard Pons-Moll. Implicit functions in feature space for 3d shape reconstruction and completion. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2020. 3

[14] Julian Chibane, Aymen Mir, and Gerard Pons-Moll. Neural unsigned distance fields for implicit function learning. *NeurIPS*, 2020. 4

[15] Enric Corona, Guillem Alenya, Antonio Gabas, and Carme Torras. Active garment recognition and target grasping point detection using deep learning. *Pattern Recognition*, 74:629–641, 2018. 2

[16] Enric Corona, Albert Pumarola, Guillem Alenya, Francesc Moreno-Noguer, and Grégory Rogez. Ganhand: Predicting human grasp affordances in multi-object scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5031–5041, 2020. 2

[17] Valentin Gabeur, Jean-Sébastien Franco, Xavier Martin, Cordelia Schmid, and Gregory Rogez. Moulding humans: Non-parametric 3d human shape estimation from single images. In *ICCV*, 2019. 3

[18] Peng Guan, Loretta Reiss, David A Hirshberg, Alexander Weiss, and Michael J Black. Drape: Dressing any person. *ToG*, 2012. 2

[19] Peng Guan, Alexander Weiss, Alexandru O Balan, and Michael J Black. Estimating human shape and pose from a single image. In *ICCV*. IEEE, 2009. 2

[20] Erhan Gundogdu, Victor Constantin, Amrollah Seifoddini, Minh Dang, Mathieu Salzmann, and Pascal Fua. Garnet: A two-stream network for fast and accurate 3d cloth draping. In *ICCV*, 2019. 2, 3

[21] Marc Habermann, Weipeng Xu, Michael Zollhoefer, Gerard Pons-Moll, and Christian Theobalt. Livecap: Real-time human performance capture from monocular video. *ACM Transactions on Graphics (TOG)*, 2019. 2

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 7

[23] Liwen Hu, Chongyang Ma, Linjie Luo, and Hao Li. Single-view hair modeling using a hairstyle database. *ToG*, 2015. 7

[24] Zeng Huang, Yuanlu Xu, Christoph Lassner, Hao Li, and Tony Tung. Arch: Animatable reconstruction of clothed humans. In *CVPR*, 2020. 3

[25] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015. 7

[26] Boyi Jiang, Juyong Zhang, Yang Hong, Jinhao Luo, Ligang Liu, and Hujun Bao. Bcnet: Learning body and cloth shape from a single image. In *ECCV*, 2020. 1, 2, 7

[27] Hanbyul Joo, Tomas Simon, and Yaser Sheikh. Total capture: A 3d deformation model for tracking faces, hands, and bodies. In *CVPR*, 2018. 2

[28] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *CVPR*, 2018. 2

[29] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[30] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *ICCV*, 2019. 2

[31] Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. In *CVPR*, 2019. 2

[32] Zorah Lahner, Daniel Cremers, and Tony Tung. Deepwrinkles: Accurate and realistic clothing modeling. In *ECCV*, 2018. 2

[33] Christoph Lassner, Javier Romero, Martin Kiefel, Federica Bogo, Michael J Black, and Peter V Gehler. Unite the people: Closing the loop between 3d and 2d human representations. In *CVPR*, 2017. 2

[34] Verica Lazova, Eldar Insafutdinov, and Gerard Pons-Moll. 360-degree textures of people in clothing from a single image. In *International Conference on 3D Vision (3DV)*, sep 2019. 2

[35] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ToG*, 2015. 1, 2, 3, 5

[36] William E Lorensen and Harvey E Cline. Marching cubes: A high resolution 3d surface construction algorithm. *SIGGRAPH*, 21(4), 1987. 3, 4

[37] Qianli Ma, Jinlong Yang, Anurag Ranjan, Sergi Pujades, Gerard Pons-Moll, Siyu Tang, and Michael J Black. Learning to dress 3d people in generative clothing. In *CVPR*, 2020. 1, 2, 4, 5, 6, 7

[38] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *CVPR*, 2019. 4

[39] Ryota Natsume, Shunsuke Saito, Zeng Huang, Weikai Chen, Chongyang Ma, Hao Li, and Shigeo Morishima. Siclope: Silhouette-based clothed people. In *CVPR*, 2019. 3

[40] Alexandros Neophytou and Adrian Hilton. A layered model of human body and garment deformation. In *3DV*. IEEE, 2014. 2

[41] Mohamed Omran, Christoph Lassner, Gerard Pons-Moll, Peter Gehler, and Bernt Schiele. Neural body fitting: Unifying deep learning and model based human pose and shape estimation. In *3DV*. IEEE, 2018. 2

[42] Hayato Onizuka, Zehra Hayirci, Diego Thomas, Akihiro Sugimoto, Hideaki Uchiyama, and Rin-ichiro Taniguchi. Tetratsdf: 3d human reconstruction from a single image with a tetrahedral outer shell. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6011–6020, 2020. 2

[43] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *CVPR*, 2019. 3, 4, 7

[44] Chaitanya Patel, Zhouyingcheng Liao, and Gerard Pons-Moll. Tailornet: Predicting clothing in 3d as a function of human pose, shape and garment style. In *CVPR*, 2020. 1, 2, 3, 4, 5, 6, 7

[45] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *CVPR*, 2019. 1, 2

[46] Gerard Pons-Moll, Sergi Pujades, Sonny Hu, and Michael Black. ClothCap: Seamless 4D clothing capture and retargeting. *SIGGRAPH*, 36(4), 2017. 2, 4

[47] Gerard Pons-Moll, Javier Romero, Naureen Mahmood, and Michael J. Black. Dyna: A model of dynamic human shape in motion. *ACM Transactions on Graphics, (Proc. SIGGRAPH)*, 34(4):120:1–120:14, aug 2015. 2

[48] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. *arXiv preprint arXiv:2011.13961*, 2020. 4

[49] Albert Pumarola, Jordi Sanchez-Riera, Gary Choi, Alberto Sanfeliu, and Francesc Moreno-Noguer. 3dpeople: Modeling the geometry of dressed humans. In *ICCV*, 2019. 3

[50] Remove background. https://www.remove.bg/. 8

[51] Javier Romero, Dimitrios Tzionas, and Michael J Black. Embodied hands: Modeling and capturing hands and bodies together. *ToG*, 2017. 2

[52] Yu Rong, Takaaki Shiratori, and Hanbyul Joo. Frankmocap: Fast monocular 3d hand and body motion capture by regression and integration. *arXiv preprint arXiv:2008.08324*, 2020. 2, 6, 7, 8

[53] Martin Runz, Kejie Li, Meng Tang, Lingni Ma, Chen Kong, Tanner Schmidt, Ian Reid, Lourdes Agapito, Julian Straub, Steven Lovegrove, and Richard Newcombe. Frodo: From detections to 3d objects. In *CVPR*, 2020. 4

[54] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *ICCV*, 2019. 3

[55] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *CVPR*, 2020. 3, 7, 8

[56] Tim Salimans and Durk P Kingma. Weight normalization: A simple reparameterization to accelerate training of deep neural networks. In *Advances in neural information processing systems*, pages 901–909, 2016. 7

[57] Igor Santesteban, Miguel A Otaduy, and Dan Casas. Learning-based animation of clothing for virtual try-on. In *Computer Graphics Forum*, volume 38. Wiley Online Library, 2019. 2

[58] Akihiko Sayo, Hayato Onizuka, Diego Thomas, Yuta Nakashima, Hiroshi Kawasaki, and Katsushi Ikeuchi. Human shape reconstruction with loose clothes from partially observed data by pose specific deformation. In *Pacific-Rim Symposium on Image and Video Technology*, pages 225–239. Springer, 2019. 2

[59] Yu Shen, Junbang Liang, and Ming C. Lin. Gan-based garment generation using sewing pattern images. In *ECCV*, 2020. 2

[60] David Smith, Matthew Loper, Xiaochen Hu, Paris Mavroidis, and Javier Romero. Facsimile: Fast and accurate scans from an image in less than a second. In *ICCV*, 2019. 2

[61] Garvita Tiwari, Bharat Lal Bhatnagar, Tony Tung, and Gerard Pons-Moll. Sizer: A dataset and model for parsing 3d clothing and learning size sensitive 3d clothing. In *ECCV*, 2020. 5, 6, 7

[62] Gul Varol, Duygu Ceylan, Bryan Russell, Jimei Yang, Ersin Yumer, Ivan Laptev, and Cordelia Schmid. Bodynet: Volumetric inference of 3d human body shapes. In *ECCV*, 2018. 3

[63] Raquel Vidaurre, Igor Santesteban, Elena Garces, and Dan Casas. Fully Convolutional Graph Neural Networks for Parametric Virtual Try-On. *Computer Graphics Forum*, 2020. 2

[64] Tuanfeng Y Wang, Duygu Ceylan, Jovan Popovic, and Niloy J Mitra. Learning a shared shape space for multimodal garment design. *arXiv preprint arXiv:1806.11335*, 2018. 2

[65] Xiangyu Xu, Hao Chen, Francesc Moreno-Noguer, Laszlo Attila Jeni, and Fernando De la Torre. 3d human shape and pose from a single low-resolution image. In *ECCV*, 2020. 2

[66] Jinlong Yang, Jean-Sébastien Franco, Franck Hétroy-Wheeler, and Stefanie Wuhrer. Analyzing clothing layer deformation statistics of 3d human motions. In *ECCV*, 2018. 2

[67] Lu Yang, Qing Song, Zhihui Wang, Mengjie Hu, Chun Liu, Xueshi Xin, Wenhe Jia, and Songcen Xu. Renovating parsing r-cnn for accurate multiple human parsing. In *ECCV*, 2020. 6, 8

[68] Tao Yu, Zerong Zheng, Kaiwen Guo, Jianhui Zhao, Qionghai Dai, Hao Li, Gerard Pons-Moll, and Yebin Liu. Doublefusion: Real-time capture of human performances with inner body shapes from a single depth sensor. In *CVPR*, 2018. 2

[69] Tao Yu, Zerong Zheng, Yuan Zhong, Jianhui Zhao, Qionghai Dai, Gerard Pons-Moll, and Yebin Liu. Simulcap: Single-view human performance capture with cloth simulation. In *CVPR*. IEEE, 2019. 2

[70] Zerong Zheng, Tao Yu, Yixuan Wei, Qionghai Dai, and Yebin Liu. Deephuman: 3d human reconstruction from a single image. In *ICCV*, 2019. 3

[71] Hao Zhu, Xinxin Zuo, Sen Wang, Xun Cao, and Ruigang Yang. Detailed human shape estimation from a single image by hierarchical mesh deformation. In *CVPR*, 2019. 2