SIZER: A Dataset and Model for Parsing 3D Clothing and Learning Size Sensitive 3D Clothing

Garvita Tiwari[1], Bharat Lal Bhatnagar[1], Tony Tung[2], and Gerard Pons-Moll[1]

[1] MPI for Informatics, Saarland Informatics Campus, Germany
[2] Facebook Reality Labs, Sausalito, USA
{gtiwari,bbhatnag,gpons}@mpi-inf.mpg.de, tony.tung@fb.com

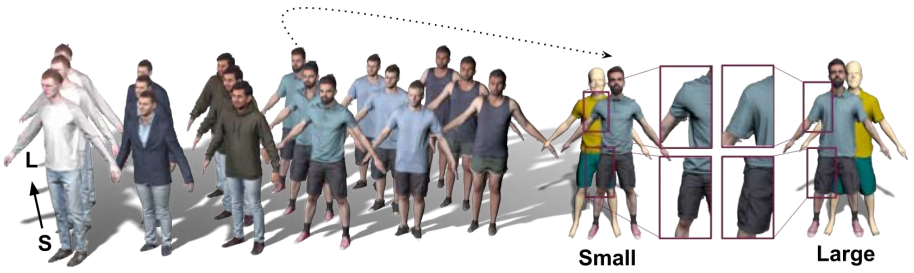


Fig. 1: *SIZER* dataset of people with clothing size variation. (*Left*): 3D Scans of people captured in different clothing styles and *sizes*. (*Right*): T-shirt and short pants for sizes small and large, which are registered to a common template.

Abstract. While models of 3D clothing learned from real data exist, no method can predict clothing deformation as a function of garment size. In this paper, we introduce SizerNet to predict 3D clothing conditioned on human body shape and garment size parameters, and ParserNet to infer garment meshes and shape under clothing with personal details in a single pass from an input mesh. SizerNet allows to estimate and visualize the dressing effect of a garment in various sizes, and ParserNet allows to edit clothing of an input mesh directly, removing the need for scan segmentation, which is a challenging problem in itself. To learn these models, we introduce the *SIZER* dataset of clothing size variation which includes 100 different subjects wearing casual clothing items in various sizes, totaling to approximately 2000 scans. This dataset includes the scans, registrations to the SMPL model, scans segmented in clothing parts, garment category and size labels. Our experiments show better parsing accuracy and size prediction than baseline methods trained on *SIZER*. The code, model and dataset will be released for research purposes at: https://virtualhumans.mpi-inf.mpg.de/sizer/.

1 Introduction

Modeling how 3D clothing fits on the human body as a function of size has numerous applications in 3D content generation (e.g., AR/VR, movie, video games, sport), clothing size recommendation (e.g., e-commerce), computer vision for fashion, and virtual try-on. It is estimated that retailers lose up to $600 billion each year due to sales returns as it is currently difficult to purchase clothing online without knowing how it will fit [3,2].

Predicting how clothing fits as a function of body shape and garment size is an extremely challenging task. Clothing interacts with the body in complex ways, and fit is a non-linear function of size and body shape. Furthermore, *clothing fit differences with size are subtle*, but they can make a difference when purchasing clothing online. Physics based simulation is still the most commonly used technique because it generalizes well, but unfortunately, it is difficult to adjust its parameters to achieve a realistic result, and it can be computationally expensive.

While there exist several works that learn how clothing deforms as a function of pose [30], or pose and shape [30,43,22,37,34], there are few works modeling how garments drape as a function of size. Recent works learn a space of styles [50,37] from physics simulations, but their aim is plausibility, and therefore they can not predict how a *real garment* will deform on a real body.

What is lacking is (1) a 3D dataset of people wearing the same garments in different sizes and (2) a data-driven model *learned from real scans* which varies with sizing and body shape. In this paper, we introduce the *SIZER* dataset, the first dataset of scans of people in different garment sizes featuring approximately 2000 scans, 100 subjects and 10 garments worn by subjects in four different sizes. Using the *SIZER* dataset we learned a Neural Network model, which we refer to as *SizerNet*, which given a body shape and a garment, can predict how the garment drapes on the body as a function of size. Learning *SizerNet* requires to map scans to a registered *multi-layer meshes* – separate meshes for body shape, and top and bottom garments. This requires segmenting the 3D scans, and estimating their body shape under clothing, and registering the garments across the dataset, which we obtain using the method explained in [14,38]. From the multi-layer meshes, we learn an encoder to map the input mesh to a latent code, and a decoder which additionally takes the body shape parameters of SMPL [33], the size label (S, M, L, XL) of the input garment, and the desired size of the output, to predict the output garment as a displacement field to a template.

Although visualizing how an existing garment fits on a body as a function of size is already useful for virtual try-on applications, we would also like to change the size of garments in existing 3D scans. Scans however, are just pointclouds, and parsing them into a multi-layer representation at test time using [14,38] requires segmentation, which sometimes requires manual intervention. Therefore, we propose *ParserNet*, which automatically maps a single mesh registration (SMPL deformed to the scan) to multi-layer meshes with a single feed-forward pass. *ParserNet*, not only segments the single mesh registration, but it reparam-

eterizes the surface so that it is coherent with common garment templates. The output multi-layer representation of *ParserNet* is powerful as it allows simulation and editing meshes separately. Additionally, the tandem of *SizerNet* and *ParserNet* allows us to edit the size of clothing directly on the mesh, allowing shape manipulation applications never explored before.

In summary, our contributions are:

- *SIZER* dataset: A dataset of clothing size variation of approximately 2000 scans including 100 subjects wearing 10 garment classes in different sizes, where we make available, scans, clothing segmentation, SMPL+G registrations, body shape under clothing, garment class and size labels.
- SizerNet: The first model learned from real scans to predict how clothing drapes on the body as a function of size.
- ParserNet: A data-driven model to map a single mesh registration into a multi-layered representation of clothing without the need for segmentation or non-linear optimization.

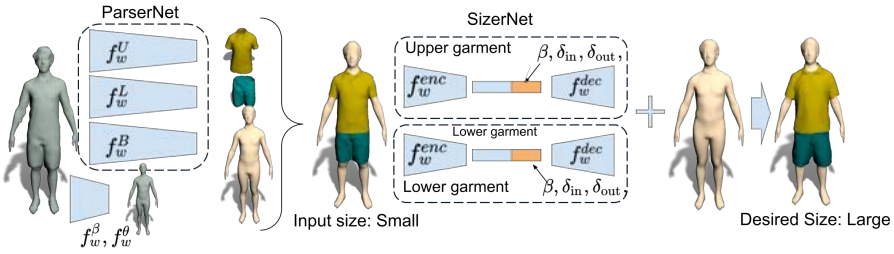


Fig. 2: We propose a model to estimate and visualize the dressing effect of a garment conditioned on body shape and garment size parameters. For this we introduce *ParserNet* (f_w^U, f_w^L, f_w^B), which takes a SMPL registered mesh $M(\boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{D})$ as input and predicts the SMPL parameters $(\boldsymbol{\theta}, \boldsymbol{\beta})$, parsed 3D garments using predefined templates $T^g(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{0})$ and predicts body shape under clothing while preserving the personal details of the subject. We also propose *SizerNet*, an encoder-decoder $(f_w^{\text{enc}}, f_w^{\text{dec}})$ based network, that resizes the garment given as input with the desired size label $(\delta_{\text{in}}, \delta_{\text{out}})$ and drapes it on the body shape under clothing.

2 Related Work

Clothing modeling. Accurate reconstruction of 3D cloth with fine structures (e.g., wrinkles) is essential for realism while being notoriously challenging. Methods based on multi-view stereo can recover global shape robustly but struggle with high frequency details in non-textured regions [51,44,16,6,47,32]. The pioneering work of [9,8] demonstrated for the first time detailed body and clothing

reconstruction from monocular video using a displacement from SMPL, which spearheaded recent developments [23,7,10,42,24,25]. These approaches do not separate body from clothing. In [38,30,14,26], the authors propose to reconstruct clothing as a layer separated from the body. These models are trained on 3D scans of real clothed people data and produce realistic models. On the other hand, physics based simulation methods have also been used to model clothing [48,49,35,21,45,46,37,43,22]. Despite the potential gap with real-world data, they are a great alternative to obtain clean data, free of acquisition noise and holes. However, they still require manual parameter tuning (e.g., time step for better convergence, sheer and stretch for better deformation effects, etc.), and can be slow or unstable. In [43,22,21] a pose and shape dependent clothing model is introduced, and [37,50] also model garment style dependent clothing using a lower-dimensional representation for style and size like PCA and garment sewing parameters, however there is no direct control on the size of clothing generated for given body shape. In [53], authors model the garment fit on different body shapes from images. Our model *SizerNet* automatically outputs realistic 3D cloth models conditioned on desired features (e.g., shape, size).

Shape under clothing. In [11,60,57], the authors propose to estimate body shape under clothing by fitting a 3D body model to 3D reconstructions of people. An objective function typically forces the body to be inside clothing while being close to the skin region. These methods cannot generalize well to complex or loose clothing without additional prior or supervision [17]. In [27,36,54,29,28,52], the authors propose learned models to estimate body shape from 2D images of clothed people, but shape accuracy is limited due to depth ambiguity. Our model *ParserNet* takes as input a 3D mesh and outputs 3D bodies under clothing with high fidelity while preserving subject identity (e.g., face details).

Cloth parsing. The literature has proposed several methods for clothed human understanding. In particular, efficient cloth parsing in 2D has been achieved using supervised learning and generative networks [55,56,58,18,19,20]. 3D clothing parsing of 3D scans has also been investigated [38,14]. The authors propose techniques based on MRF-GrabCut [41] to segment 3D clothing from 3D scans and transfer them to different subjects. However the approach requires several steps, which is not optimal for scalability. We extend previous work with *SIZER*, a fully automatic data-driven pipeline. In [13], the authors jointly predict clothing and inner body surface, with semantic correspondences to SMPL. However, it does not have semantic clothing information.

3D datasets. To date, only a few datasets consist of 3D models of subjects with segmented clothes. 3DPeople [40], Cloth3D [12] consists of a large dataset of synthetic 3D humans with clothing. None of the synthetic datasets contains realistic cloth deformations like the SIZER dataset. THUman [61] consists of sequences of clothed 3D humans in motion, captured with a consumer RGBD sensor (Kinectv2), and are reconstructed using volumetric SDF fusion [59]. How-

ever, 3D models are rather smooth compared to our 3D scans and no ground truth segmentation of clothing is provided. Dyna and D-FAUST [39,15] consist of high-res 3D scans of 10 humans in motion with different shape but the subjects are only wearing minimal clothing. BUFF [60] contains high-quality 3D scans of 6 subjects with and without clothing. The dataset is primarily designed to train models to estimate body shape under clothing and doesn't contain garments segmentation. In [14], the authors create a digital wardrobe with 3D templates of garments to dress 3D bodies. In [26], authors propose a mixture of synthetic and real data, which contains garment, body shape and pose variations. However, the fraction of real dataset (∼300 scans) is fairly small. DeepFahsion3D [62] is a dataset of real scans of clothing containing various garment styles. None of these datasets contain garment sizing variation. Unlike our proposed *SIZER* dataset, no dataset contains a large amount of pre-segmented clothing from 3D scans at different sizes, with corresponding body shapes under clothing.

3 Dataset

In this paper, we address a very challenging problem of modeling garment fitting as a function of body shape and garment size. As explained in Sec. 2, one of the key bottlenecks that hinder progress in this direction is the lack of real-world datasets that contain calibrated and well-annotated garments in different sizes draped on real humans. To this end, we present *SIZER* dataset, a dataset of over 2000 scans containing people in diverse body shapes in various garments styles and sizes. We describe our dataset in Sec. 3.1 and 3.2.

3.1 SIZER dataset: Scans

We introduce the *SIZER* dataset that contains 100 subjects, wearing the same garment in 2 or 3 garment sizes (S, M, L, XL). We include 10 garment classes, namely shirt, dress-shirt, jeans, hoodie, polo t-shirt, t-shirt, shorts, vest, skirt, and coat, which amounts to roughly 200 scans per garment class. We capture the subjects in a relaxed A-pose to avoid stretching or tension due to pose in the garments. Figure 1 shows some examples of people wearing a fixed set of garments in different sizes. We use a Treedy's static scanner [5] which has 130+ cameras, and reconstruct the scans using Agisoft's Metashape software [1]. Our scans are high resolution and are represented by meshes, which have different underlying graph connectivity across the dataset, and hence it is challenging to use this dataset directly in any learning framework. We preprocess our dataset, by registering them to SMPL [33]. We explain the structure of processed data in the following section.

3.2 SIZER dataset: SMPL and Garment registrations

To improve general usability of the *SIZER* dataset, we provide SMPL+G registrations [31,14] registrations. Registering our scans to SMPL, brings all our

scans to correspondence, and provides more control over the data via pose and shape parameters from the underlying SMPL. We briefly describe the SMPL and SMPL+G formulations below.

SMPL represents the human body as a parametric function $M(\cdot)$, of pose ($\boldsymbol{\theta}$) and shape ($\boldsymbol{\beta}$). We add per-vertex displacements ($\mathbf{D}$) on top of SMPL to model deformations corresponding to hair, garments, etc. thus resulting in the SMPL model. SMPL applies standard skinning $W(\cdot)$ to a base template $\mathbf{T}$ in T-pose. Here, $\mathbf{W}$ denotes the blend weights and $B_p(\cdot)$ and $B_s(\cdot)$ models pose and shape dependent deformations respectively.

$$M(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{D}) = W(T(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{D}), J(\boldsymbol{\beta}), \boldsymbol{\theta}, \mathbf{W}) \tag{1}$$

$$T(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{D}) = \mathbf{T} + B_s(\boldsymbol{\beta}) + B_p(\boldsymbol{\theta}) + \mathbf{D} \tag{2}$$

SMPL+G is a parametric formulation to represent the human body and garments as separate meshes.To register the garments we first segment scans into garments and skin parts [14]. We refine the scan segmentation step used in [14] by fine-tuning the Human Parsing network [20] with a multi-view consistency loss. We then use the multi-mesh registration approach from [14] to register garments to the SMPL+G model. For each garment class, we obtain a template mesh which is defined as a subset of the SMPL template, given by $T^g(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{0}) = \mathbf{I}^g T(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{0})$, where $\mathbf{I}^g \in \mathbb{Z}_2^{m_g \times n}$ is an indicator matrix, with $\mathbf{I}^g_{i,j} = 1$ if garment g vertex $i \in \{1 \dots m_g\}$ is associated with body shape vertex $j \in \{1 \dots n\}$. m_g and n denote the number of vertices in the garment template and the SMPL mesh respectively. Similarly, we define a garment function $G(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{D}^g)$ using Eq. (3), where $\mathbf{D}^g$ are the per-vertex offsets from the template

$$G(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{D}^g) = W(T^g(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{D}^g), J(\boldsymbol{\beta}), \boldsymbol{\theta}, \mathbf{W}). \tag{3}$$

For every scan in the *SIZER* dataset, we will release the scan, segmented scan, and SMPL+G registrations, garment category and garment size label.

This dataset can be used in several applications like virtual try-on, character animation, learning generative models, data-driven body shape under clothing, size and(or) shape sensitive clothing model, etc. To stimulate further research in this direction, we will release the dataset,code and baseline models, which can be used as a benchmark in 3D clothing parsing and 3D garment resizing. We use this dataset to build a model for the task of garment extraction from single mesh (*ParserNet*) and garment resizing (*SizerNet*), which we describe in the next section.

4 Method

We introduce *ParserNet* (Sec. 4.2), the first method for extracting garments directly from SMPL registered meshes. For parsing garments, we first predict the underlying body SMPL parameters using a pose and shape prediction network (Sec. 4.1) and use *ParserNet* to extract garment layers and personal features

like hair, facial features to create body shape under clothing. Next, we present *SizerNet* (Sec. 4.3), an encoder-decoder based deep network for garment resizing. An overview of the method is shown in Fig. 2.

4.1 Pose and shape prediction network

To estimate body shape under clothing, we first create the undressed SMPL body for a given clothed input single layer mesh $M(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{D})$, by predicting $\boldsymbol{\theta}, \boldsymbol{\beta}$ using $f_w^{\boldsymbol{\theta}}$ and $f_w^{\boldsymbol{\beta}}$ respectively. We train $f_w^{\boldsymbol{\theta}}$ and $f_w^{\boldsymbol{\beta}}$ with L_2 loss over parameters and per-vertex loss between predicted SMPL body and clothed input mesh, as shown in Eq. (4) and (5). Since the reference body under clothing parameters $\boldsymbol{\theta}, \boldsymbol{\beta}$ obtained via instance specific optimization (Sec. 3.2) can be inaccurate, we add an additional per-vertex loss between our predicted SMPL body vertices $M(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\beta}}, \mathbf{0})$ and the input clothed mesh $M(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{D})$. This brings the predicted undressed body closer to the input clothed mesh. We observe more stable results training $f_w^{\boldsymbol{\theta}}$ and $f_w^{\boldsymbol{\beta}}$ separately initially, using the reference $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ respectively. Since the $\boldsymbol{\beta}$ components in SMPL are normalized to have $\sigma = 1$, we un-normalize them by scaling by their respective standard deviations $[\sigma_1, \sigma_2, \ldots, \sigma_{10}]$ as given in Eq. (5).

$$\mathcal{L}_{\boldsymbol{\theta}} = w_{\text{pose}} ||\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}||_2^2 + w_v ||M(\boldsymbol{\beta}, \hat{\boldsymbol{\theta}}, \mathbf{0}) - M(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{D})|| \tag{4}$$

$$\mathcal{L}_{\boldsymbol{\beta}} = w_{\text{shape}} \sum_{i=1}^{10} \sigma_i (\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i)^2 + w_v ||M(\hat{\boldsymbol{\beta}}, \boldsymbol{\theta}, \mathbf{0}) - M(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{D})|| \tag{5}$$

Here, w_{pose}, w_{shape} and w_v are weights for the loss on pose, shape and predicted SMPL surface. $(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\beta}})$ denote predicted parameters. The output is a *smooth* (SMPL model) body shape under clothing.

4.2 ParserNet

Parsing garments. Parsing garments from a single mesh $(\mathbf{M})$ can be done by segmenting it into separate garments for each class $(\mathbf{G}_{\text{seg}}^{g,k})$, which leads to different underlying graph connectivity $(\mathcal{G}_{\text{seg}}^{g,k} = (\mathbf{G}_{\text{seg}}^{g,k}, \mathbf{E}_{\text{seg}}^{g,k}))$ across all the instances (k) of a garment class g, shown in Fig. 3 (right). Hence, we propose to parse garments by deforming vertices of a template $T^g(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{0})$ with fixed connectivity $\mathbf{E}^g$, obtaining vertices $\mathbf{G}^{g,k} \in \mathcal{G}^{g,k}$, where $\mathcal{G}^{g,k} = (\mathbf{G}^{g,k}, \mathbf{E}^g)$, shown in Fig. 3 (middle).

Our key idea is to predict the deformed vertices $\mathbf{G}^g$ directly as a convex combination of vertices of the input mesh $\mathbf{M} = M(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{D})$ with a learned sparse regressor matrix $\mathbf{W}^g$, such that $\mathbf{G}^g = \mathbf{W}^g \mathbf{M}$. Specifically, *ParserNet* predicts the sparse matrix $(\mathbf{W}^g)$ as a function of input mesh features (vertices and normals) and a predefined per-vertex neighborhood $(\mathcal{N}_i)$ for every vertex i of garment class g. We will henceforth drop $(.)^{g,k}$ unless required. In this way,

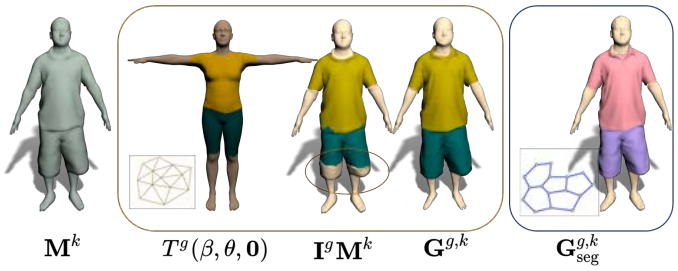


Fig. 3: Left to right: Input single mesh ($\mathbf{M}^k$), garment template ($T^g(\boldsymbol{\beta},\boldsymbol{\theta},\mathbf{0}) = \mathbf{I}^g T(\boldsymbol{\beta},\boldsymbol{\theta},\mathbf{0})$), garment mesh extracted using $\mathbf{G}^{g,k} = \mathbf{I}^g\mathbf{M}^k$, multi-layer meshes ($\mathbf{G}^{g,k}$) registered to SMPL+G, all with garment class specific edge connectivity $\mathbf{E}^g$, and segmented scan $\mathbf{G}^{g,k}_{\text{seg}}$ with instance specific edge connectivity $\mathbf{E}^{g,k}_{\text{seg}}$.

the output vertices $\mathbf{G}_i \in \mathbb{R}^3$, where $i \in \{1,\ldots,m_g\}$, are obtained as a convex combination of input mesh vertices $\mathbf{M}_j \in \mathbb{R}^3$ in a predefined neighborhood ($\mathcal{N}_i$).

$$\mathbf{G}_i = \sum_{j\in\mathcal{N}_i} \mathbf{W}_{ij}\mathbf{M}_j. \tag{6}$$

Parsing detailed body shape under clothing. For generating detailed body shape under clothing, we first create a *smooth body mesh*, using SMPL parameters $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ predicted from $f_w^{\boldsymbol{\theta}}, f_w^{\boldsymbol{\beta}}$ (Sec. 4.1). Using the same aforementioned convex combination formulation, *Body ParserNet* transfers the visible skin vertices from the input mesh to the smooth body mesh, obtaining hair and facial features. We parse the input mesh into upper, lower garments and detailed shape under clothing using 3 sub-networks (f_w^U, f_w^L, f_w^B) of *ParserNet*, as shown in Fig. 2.

4.3 SizerNet

We aim to edit the garment mesh based on garment size labels such as S, M, L, etc, to see the dressing effect of the garment for a new size. For this task, we propose an encoder-decoder based network, which is shown in Fig. 2 (right). The network f_w^{enc}, encodes the garment mesh $\mathbf{G}_{in}$ to a lower-dimensional latent code $\boldsymbol{x}_{\text{gar}} \in \mathbb{R}^d$, shown in Eq. (7). We append $(\boldsymbol{\beta}, \delta_{\text{in}}, \delta_{\text{out}})$ to the latent space, where $\delta_{\text{in}}, \delta_{\text{out}}$ are one-hot encodings of input and desired output sizing and $\boldsymbol{\beta}$ is the SMPL $\boldsymbol{\beta}$ parameter for underlying body shape.

$$\boldsymbol{x}_{\text{gar}} = f_w^{\text{enc}}(\mathbf{G}_{in}), \quad f_w^{\text{enc}}(.) : \mathbb{R}^{m_g\times 3} \rightarrow \mathbb{R}^d \tag{7}$$

The decoder network, $f_w^{\text{dec}}(.) : \mathbb{R}^{|\boldsymbol{\beta}|} \times \mathbb{R}^d \times \mathbb{R}^{2|\boldsymbol{\delta}|} \rightarrow \mathbb{R}^{m_g\times 3}$ predicts the displacement field $\mathbf{D}^g = f_w^{\text{dec}}(\boldsymbol{\beta}, \boldsymbol{x}_{\text{gar}}, \boldsymbol{\delta}_{\text{in}}, \boldsymbol{\delta}_{\text{out}})$ on top on template. We obtain the output garment $\mathbf{G}_{out}$ in the new desired size δ_{out} using Eq. (3).

4.4 Loss functions

We train the networks, *ParserNet* and *SizerNet* with training losses given by Eq. (8) and (9) respectively, where w_{3D}, w_{norm}, w_{lap}, w_{interp} and w_{w} are weights for the loss on vertices, normal, Laplacian, interpenetration and weight regularizer term respectively. We explain each of the loss terms in this section.

$$\mathcal{L}_{\text{parser}} = w_{\text{3D}}\mathcal{L}_{\text{3D}} + w_{\text{norm}}\mathcal{L}_{\text{norm}} + w_{\text{lap}}\mathcal{L}_{\text{lap}} + w_{\text{interp}}\mathcal{L}_{\text{interp}} + w_{\text{w}}\mathcal{L}_{\text{w}} \tag{8}$$

$$\mathcal{L}_{\text{sizer}} = w_{\text{3D}}\mathcal{L}_{\text{3D}} + w_{\text{norm}}\mathcal{L}_{\text{norm}} + w_{\text{lap}}\mathcal{L}_{\text{lap}} + w_{\text{interp}}\mathcal{L}_{\text{interp}} \tag{9}$$

- **3D vertex loss for garments.** We define $\mathcal{L}_{\text{3D}}$ as L_1 loss between predicted and ground truth vertices

$$\mathcal{L}_{\text{3D}} = ||\mathbf{G}_{\text{P}} - \mathbf{G}_{\text{GT}}||_1. \tag{10}$$

- **3D vertex loss for shape under clothing.** For training f_w^B (ParserNet for the body), we use the input mesh skin as supervision for predicting personal details of subject. We define a garment class specific geodesic distance weighted loss term, as shown in Eq. (11), where $\mathbf{I}^s$ is the indicator matrix for skin region and $\boldsymbol{w}_{\text{geo}}$ is a vector containing the *sigmoid* of the geodesic distances from vertices to the boundary between skin and non-skin regions. The loss term is high when the prediction is far from the input mesh $\mathbf{M}$ for the visible skin region, and lower for the cloth region, with a smooth transition regulated by the geodesic term. Let $\text{abs}_{ij}(\cdot)$ denote an element-wise absolute value operator. Then the loss is computed as

$$\mathcal{L}_{\text{3D}}^{\text{body}} = \|\boldsymbol{w}_{\text{geo}}^T \cdot \text{abs}_{ij}(\mathbf{G}_{\text{P}}^{\text{s}} - \mathbf{I}^s\mathbf{M})\|_1. \tag{11}$$

- **Normal Loss.** We define $\mathcal{L}_{\text{norm}}$ as the difference in angle between ground truth face normal ($\mathbf{N}_{\text{GT}}^i$) and predicted face normal ($\mathbf{N}_P^i$).

$$\mathcal{L}_{\text{norm}} = \frac{1}{N_{\text{faces}}} \sum_{i}^{N_{\text{faces}}} (1 - (\mathbf{N}_{\text{GT},i})^T\mathbf{N}_{P,i}). \tag{12}$$

- **Laplacian smoothness term.** This enforces the Laplacian of predicted garment mesh to be close to the Laplacian of ground truth mesh. Let $\mathbf{L}^g \in \mathbb{R}^{m_g \times m_g}$ be the graph Laplacian of the garment mesh $\mathbf{G}_{\text{GT}}$, and $\boldsymbol{\Delta}_{\text{init}} = \mathbf{L}^g\mathbf{G}_{\text{GT}} \in \mathbb{R}^{m_g \times 3}$ be the differential coordinates of the $\mathbf{G}_{\text{GT}}$, then we compute the Laplacian smoothness term for a predicted mesh $\mathbf{G}_{\text{P}}$ as

$$\mathcal{L}_{\text{lap}} = ||\boldsymbol{\Delta}_{\text{init}} - \mathbf{L}^g\mathbf{G}_{\text{P}}||_2. \tag{13}$$

- **Interpenetration loss.** Since minimizing per-vertex loss does not guarantee that the predicted garment lies outside the body surface, we use the interpenetration loss term in Eq. (14) proposed in GarNet [22]. For every vertex $\mathbf{G}_{\text{P},j}$, we find the nearest vertex in the predicted body shape under clothing ($\mathbf{B}_i$) and define the body-garment correspondences as $\mathcal{C}(\mathbf{B}, \mathbf{G}_{\text{P}})$.

Let $\mathbf{N}_i$ be the normal of the i^{th} body vertex $\mathbf{B}_i$. If the predicted garment vertex $\mathbf{G}_{\mathrm{P},j}$ penetrates the body, it is penalized with the following loss

$$\mathcal{L}_{\text{interp}} = \sum_{(i,j)\in\mathcal{C}(\mathbf{B},\mathbf{G}_{\mathrm{P}})} \mathbb{1}_{d(\mathbf{G}_{\mathrm{P},j},\boldsymbol{G}_{\mathrm{GT},j})<d_{tol}} ReLU(-\mathbf{N}_i(\mathbf{G}_{\mathrm{P},j}-\mathbf{B}_i))/m_g, \quad (14)$$

where notice that $\mathbb{1}_{d(\mathbf{G}_{\mathrm{P},j},\boldsymbol{G}_{\mathrm{GT},j})<d_{tol}}$ activates the loss when the distance between predicted garment mesh vertices and ground truth mesh vertices is small *i.e.* $< d_{tol}$.

- **Weight regularizer.** To preserve the fine details when parsing the input mesh, we want the weights predicted by the network to be sparse and confined in a local neighborhood. Hence, we add a regularizer which penalizes large values for $\mathbf{W}_{ij}$ if the distance between of $\mathbf{M}_j$ and the vertex $\mathbf{M}_k$ with largest weight $k = \arg\max_j \mathbf{W}_{ij}$ is large. Let $d(\cdot,\cdot)$ dennote Euclidean distance between vertices, then the regularizer equals

$$\mathcal{L}_w = \sum_{i=1}^{m_g} \sum_{j\in\mathcal{N}_i} \mathbf{W}_{ij} d(\mathbf{M}_k, \mathbf{M}_j), \;\; k = \arg\max_j \mathbf{W}_{ij}. \quad (15)$$

4.5 Implementation Details

We implement $f_w^{\boldsymbol{\theta}}$ and $f_w^{\boldsymbol{\beta}}$ networks with 2 fully connected and a linear output layer. We implement *ParserNet* f_w^U, f_w^L, f_w^B with 3 fully connected layers. We use neighborhood ($\mathcal{N}_i$) size of $|\mathcal{N}_i| = 50$, for our experiments. We first train the network for garment classes which share the same garment template and then fine-tune separately for each garment class g. To speed up training for *ParserNet*, we train the network to predict $\mathbf{W}^g = \mathbf{I}^g$, where $\mathbf{I}^g$ is the indicator matrix for garment class g, explained in Sec. 3.2. This initializes the network to parse the garment by cutting out a part of the input mesh based on the constant per-garment indicator matrix, shown in Fig. 3.

For *SizerNet* we use $d = 30$ and we implement f_w^{enc}, f_w^{dec} with fully connected layers and skip connections between encoder and decoder network. We held out 40 scans for testing in each garment class, which includes some cases with unseen subjects and some with unseen garment size for seen subjects. For pose-shape prediction network, *ParserNet* and *SizerNet* we use batch-size of 8 and learning rate of 0.0001.

5 Experiments and Results

5.1 Results of 3D garment parsing and shape under clothing

To validate the choice of parsing the garments using a sparse regressor matrix ($\mathbf{W}$), we compare the results of *ParserNet* with two baseline approaches: 1) A linearized version of *ParserNet* implemented with LASSO, and 2) A naive FC

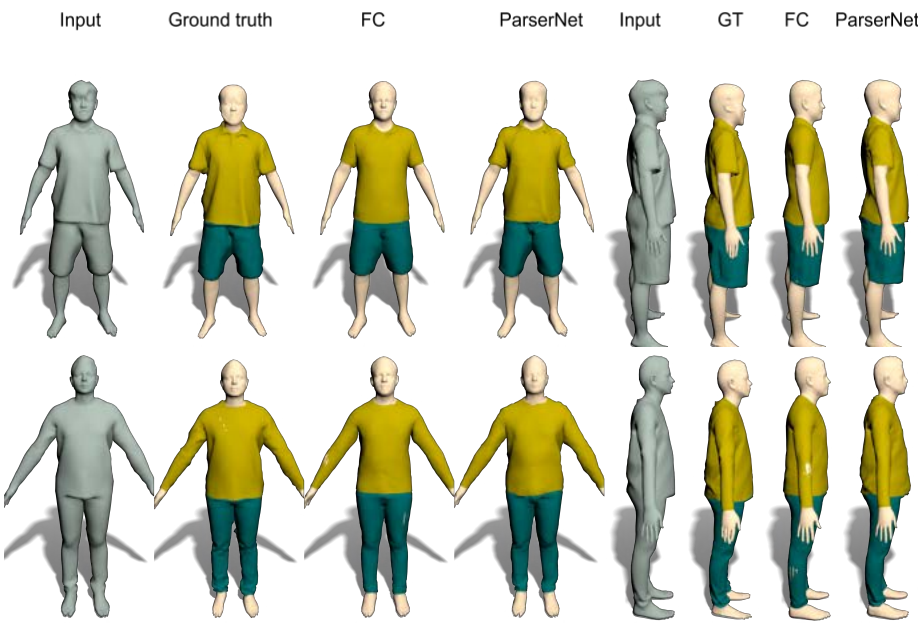


Fig. 4: Comparison of *ParserNet* with a FC network from front and lateral view.

network, which has the same architecture as *ParserNet*. However, instead of predicting the weight matrix ($\mathbf{W}$), the FC network directly predicts the deformation ($\mathbf{D}^g$) from the garment template ($T^g(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{0})$) for a given input ($\mathbf{M}$).

We compare the per-vertex error of *ParserNet* with the aforementioned baselines in Tab. 1. Figure 4 shows that *ParserNet* can produce details, fine wrinkles, and large garment deformations, which is not possible with a naive FC network. This is attainable because *ParserNet* reconstructs the output garment mesh as a localized sparse weighted sum of input vertex locations, and hence preserves the geometry details present in the input mesh. However, in the case of naive FC network, the predicted displacement field ($\mathbf{D}^g$) is smooth and does not explain large deformations. Hence, naive FC network is not able to predict loose garments and does not preserve fine details. We show results of *ParserNet* for more garment classes in Fig. 5 and add more results in the supplementary material.

5.2 Results of garment resizing

Editing garment meshes based on garment size label is an unexplored problem and, hence there are no well defined quantitative metrics. We introduce two quantitative metrics, namely change in mesh *surface area* (A_{err}) and *per-vertex error* (V_{err}) for evaluating the resizing task. *Surface area* accounts for the scale of a garment, which only changes with the garment size, and *per-vertex error* accounts for details and folds created due to the underlying body shape and looseness/tightness of the garment. Moreover, subtle changes in garment shape

Garment	Linear Model	FC	*ParserNet*	Garment	Linear Model	FC	*ParserNet*
Polo	32.21	17.25	**14.33**	Shorts	29.78	20.12	**16.07**
Shirt	27.63	19.35	**14.56**	Pants	34.82	18.2	**17.24**
Vest	28.17	18.56	**15.89**	Coat	41.27	22.19	**15.34**
Hoodies	37.34	23.69	**15.76**	Shorts2	31.38	23.45	**16.23**
T-Shirt	26.94	15.98	**13.77**				

Table 1: Average per-vertex error V_{err} of proposed method for parsing garment meshes for different garment class (in mm).

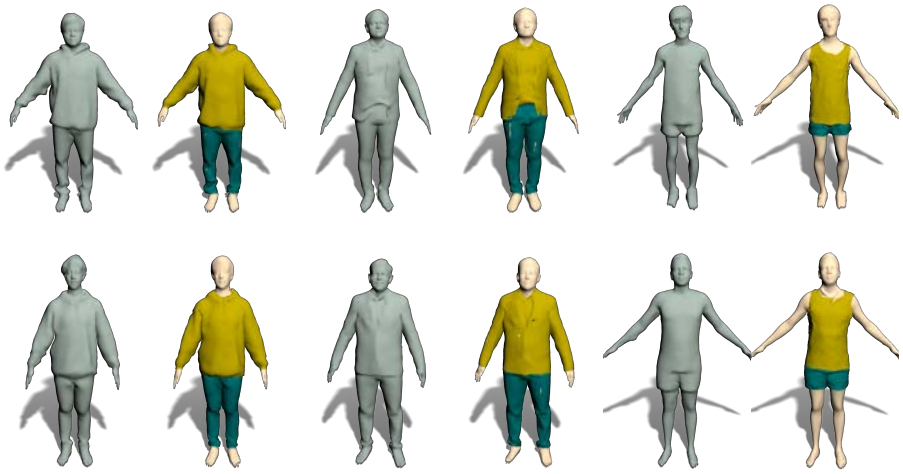


Fig. 5: Input single mesh and *ParserNet* results for more garments.

with respect to size are difficult to evaluate. Hence, we use heat map visualizations for qualitative analysis of the results.

Since there is no other existing work for garment resizing task to compare with, we evaluate our method against the following three baselines.

1. *Error margin* in data: We define error margin as the change in *per-vertex location* (V_{err}) and *surface area* (A_{err}) between garments of two consecutive size for a subject in the dataset. Our model should ideally produce a smaller error than this margin.
2. *Average prediction*: For every subject in the dataset, we create the average garment (G_{avg}), by averaging over all the available sizes for a subject.
3. *Linear scaling + Alignment*: We linearly scale the garment mesh, according to desired size label, and then align the garment to the underlying body.

Table 2 shows the errors for each experiment. *SizerNet* results in lower errors, as compared to the linear scaling method, which reflects the need for modelling

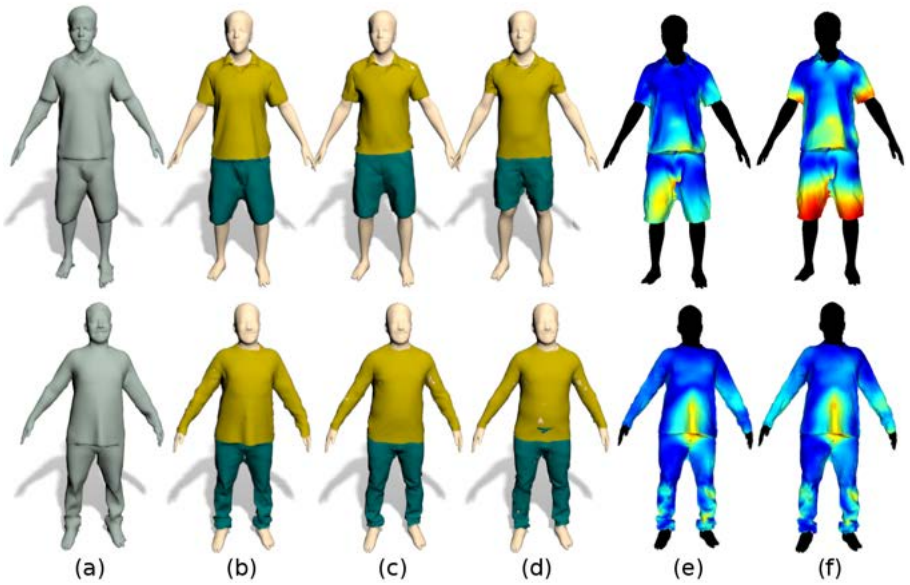
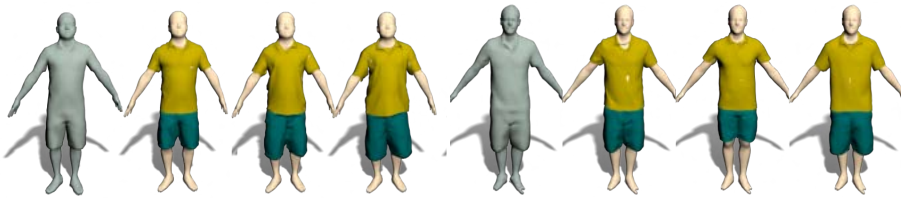


Fig. 6: (a) Input single mesh. (b) Parsed multi-layer mesh from ParserNet. (c),(d) Resized garment in two subsequent smaller sizes. (e), (f) Heatmap of change in per vertex error on original parsed garment for two new sizes.

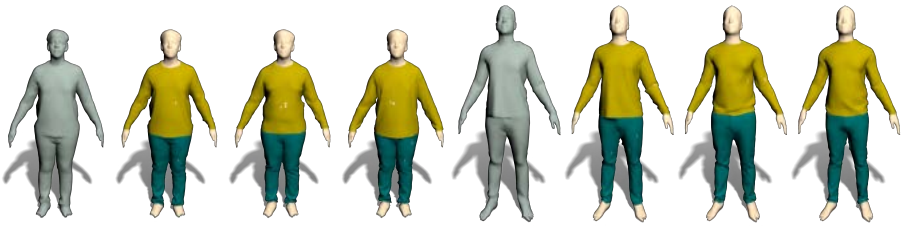
the non-linear relationship between garment shape, underlying body shape and garment size. We also see that network predictions yield lower error as compared to average garment prediction, which suggests that the model is learning the size variation, even though the differences in the ground truth itself are subtle. We present the results of *SizerNet* for common garment classes in Tab. 2, Fig. 6, 7 and add more results in the supplementary material.

Garment	Error-margin		Average-pred		Linear Scaling		Ours	
	V_{err}	A_{err}	V_{err}	A_{err}	V_{err}	A_{err}	V_{err}	A_{err}
Polo t-shirt	33.25	24.56	23.86	3.63	35.05	8.45	**16.42**	**1.79**
Shirt	36.52	19.57	21.95	2.76	34.53	7.01	**15.54**	**1.41**
Shorts	43.21	27.21	24.79	5.41	35.77	4.99	**16.71**	**2.38**
Pants	30.83	15.15	21.54	4.73	38.16	7.13	**19.26**	**2.43**

Table 2: Average per vertex error (V_{err} in mm) and surface area error(A_{err} in %) of proposed method for garment resizing.



(a) Small(input, parsed), Medium, Large (b) Medium(input, parsed), Small, Large



(c) Large(input, parsed), Medium, XLarge (d) XLarge(input,parse), Large, Medium

Fig. 7: Results of *ParserNet* + *SizerNet*, where we parse the garments from input single mesh and change the size of garment to visualise dressing effect.

6 Conclusion

We introduce *SIZER*, a clothing size variation dataset and model, which is the first real dataset to capture clothing size variation on different subjects. We also introduce *ParserNet*: a 3D garment parsing network and *SizerNet*: a size sensitive clothing model. With this method, one can change the single mesh registration to multi-layer meshes of garments and body shape under clothing, without the need for scan segmentation and can use the result for animation, virtual try-on, etc. *SizerNet* can drape a person with garments in different sizes. Since our dataset only consists of roughly aligned A-poses, we are limited to A-pose. We only exploit geometry information (vertices and normals) for 3D clothing parsing. In future work, we plan to use the color information in *ParserNet* via texture augmentation, to improve the accuracy and generalization of the proposed method. We will release the model, dataset, and code to stimulate research in the direction of 3D garment parsing, segmentation, resizing and predicting body shape under clothing.

Acknowledgements. This work is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - 409792180 (Emmy Noether Programme, project: Real Virtual Humans) and a Facebook research award. We thank Tarun, Navami, and Yash for helping us with the data capture and RVH team members [4], for their meticulous feedback on this manuscript.

References

1. Agisoft metashape, https://www.agisoft.com/
2. The high cost of retail returns, https://www.thebalancesmb.com/the-high-cost-of-retail-returns-2890350
3. Ihl group, https://www.ihlservices.com/
4. Real virtual humans, max planck institute for informatics, https://virtualhumans.mpi-inf.mpg.de/people.html
5. Treedy's scanner, https://www.treedys.com
6. de Aguiar, E., Stoll, C., Theobalt, C., Ahmed, N., Seidel, H., Thrun, S.: Performance capture from sparse multi-view video. ACM Trans. Graph. **27**(3), 98:1–98:10 (2008)
7. Alldieck, T., Magnor, M., Bhatnagar, B.L., Theobalt, C., Pons-Moll, G.: Learning to reconstruct people in clothing from a single RGB camera. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (jun 2019)
8. Alldieck, T., Magnor, M., Xu, W., Theobalt, C., Pons-Moll, G.: Detailed human avatars from monocular video. In: International Conference on 3D Vision (3DV) (sep 2018)
9. Alldieck, T., Magnor, M., Xu, W., Theobalt, C., Pons-Moll, G.: Video based reconstruction of 3d people models. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018)
10. Alldieck, T., Pons-Moll, G., Theobalt, C., Magnor, M.: Tex2shape: Detailed full human body geometry from a single image. In: IEEE International Conference on Computer Vision (ICCV). IEEE (oct 2019)
11. Bălan, A.O., Black, M.J.: The naked truth: Estimating body shape under clothing. In: European Conf. on Computer Vision. pp. 15–29. Springer (2008)
12. Bertiche, H., Madadi, M., Escalera, S.: CLOTH3D: clothed 3d humans. vol. abs/1912.02792 (2019)
13. Bhatnagar, B.L., Sminchisescu, C., Theobalt, C., Pons-Moll, G.: Combining implicit function learning and parametric models for 3d human reconstruction. In: European Conference on Computer Vision (ECCV). Springer (August 2020)
14. Bhatnagar, B.L., Tiwari, G., Theobalt, C., Pons-Moll, G.: Multi-garment net: Learning to dress 3d people from images. In: IEEE International Conference on Computer Vision (ICCV). IEEE (oct 2019)
15. Bogo, F., Romero, J., Pons-Moll, G., Black, M.J.: Dynamic FAUST: Registering human bodies in motion. In: IEEE Conf. on Computer Vision and Pattern Recognition (2017)
16. Bradley, D., Popa, T., Sheffer, A., Heidrich, W., Boubekeur, T.: Markerless garment capture. In: ACM Transactions on Graphics. vol. 27, p. 99. ACM (2008)
17. Chen, X., Pang, A., Zhu, Y., Li, Y., Luo, X., Zhang, G., Wang, P., Zhang, Y., Li, S., Yu, J.: Towards 3d human shape recovery under clothing. CoRR **abs/1904.02601** (2019)
18. Dong, H., Liang, X., Wang, B., Lai, H., Zhu, J., Yin, J.: Towards multi-pose guided virtual try-on network. International Conference on Computer Vision (ICCV) (2019)
19. Dong, H., Liang, X., Zhang, Y., Zhang, X., Xie, Z., Wu, B., Zhang, Z., Shen, X., Yin, J.: Fashion editing with adversarial parsing learning. Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
20. Gong, K., Liang, X., Li, Y., Chen, Y., Yang, M., Lin, L.: Instance-level human parsing via part grouping network. In: ECCV (2018)

21. Guan, P., Reiss, L., Hirshberg, D., Weiss, A., Black, M.J.: DRAPE: DRessing Any PErson. ACM Trans. on Graphics (Proc. SIGGRAPH) **31**(4), 35:1–35:10 (Jul 2012)
22. Gundogdu, E., Constantin, V., Seifoddini, A., Dang, M., Salzmann, M., Fua, P.: Garnet: A two-stream network for fast and accurate 3d cloth draping. In: IEEE International Conference on Computer Vision (ICCV). IEEE (oct 2019)
23. Habermann, M., Xu, W., , Zollhoefer, M., Pons-Moll, G., Theobalt, C.: Livecap: Real-time human performance capture from monocular video (oct 2019)
24. Habermann, M., Xu, W., , Zollhoefer, M., Pons-Moll, G., Theobalt, C.: Deepcap: Monocular human performance capture using weak supervision. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE (jun 2020)
25. Huang, Z., Xu, Y., Lassner, C., Li, H., Tung, T.: Arch: Animatable reconstruction of clothed humans. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3093–3102 (2020)
26. Jiang, B., Zhang, J., Hong, Y., Luo, J., Liu, L., Bao, H.: Bcnet: Learning body and cloth shape from a single image. arXiv preprint arXiv:2004.00214 (2020)
27. Kanazawa, A., Black, M.J., Jacobs, D.W., Malik, J.: End-to-end recovery of human shape and pose. In: Computer Vision and Pattern Regognition (CVPR) (2018)
28. Kolotouros, N., Pavlakos, G., Black, M.J., Daniilidis, K.: Learning to reconstruct 3D human pose and shape via model-fitting in the loop. In: International Conference on Computer Vision (Oct 2019)
29. Kolotouros, N., Pavlakos, G., Daniilidis, K.: Convolutional mesh regression for single-image human shape reconstruction. In: CVPR (2019)
30. Laehner, Z., Cremers, D., Tung, T.: Deepwrinkles: Accurate and realistic clothing modeling. In: European Conference on Computer Vision (ECCV) (September 2018)
31. Lazova, V., Insafutdinov, E., Pons-Moll, G.: 360-degree textures of people in clothing from a single image. In: International Conference on 3D Vision (3DV) (sep 2019)
32. Leroy, V., Franco, J., Boyer, E.: Multi-view dynamic shape refinement using local temporal integration. In: IEEE International Conference on Computer Vision, ICCV. pp. 3113–3122. Venice, Italy (oct 2017)
33. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: SMPL: A skinned multi-person linear model. ACM Trans. Graphics (Proc. SIGGRAPH Asia) **34**(6), 248:1–248:16 (Oct 2015)
34. Ma, Q., Yang, J., Ranjan, A., Pujades, S., Pons-Moll, G., Tang, S., Black, M.: Learning to dress 3d people in generative clothing. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE (jun 2020)
35. Miguel, E., Bradley, D., Thomaszewski, B., Bickel, B., Matusik, W., Otaduy, M.A., Marschner, S.: Data-driven estimation of cloth simulation models. Comput. Graph. Forum **31**(2), 519–528 (2012)
36. Omran, M., Lassner, C., Pons-Moll, G., Gehler, P., Schiele, B.: Neural body fitting: Unifying deep learning and model based human pose and shape estimation. In: International Conf. on 3D Vision (2018)
37. Patel, C., Liao, Z., Pons-Moll, G.: The virtual tailor: Predicting clothing in 3d as a function of human pose, shape and garment style. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE (Jun 2020)
38. Pons-Moll, G., Pujades, S., Hu, S., Black, M.: ClothCap: Seamless 4D clothing capture and retargeting. ACM Transactions on Graphics **36**(4) (2017)
39. Pons-Moll, G., Romero, J., Mahmood, N., Black, M.J.: Dyna: a model of dynamic human shape in motion. ACM Transactions on Graphics **34**, 120 (2015)

40. Pumarola, A., Sanchez, J., Choi, G., Sanfeliu, A., Moreno-Noguer, F.: 3DPeople: Modeling the Geometry of Dressed Humans. In: International Conference in Computer Vision (ICCV) (2019)
41. Rother, C., Kolmogorov, V., Blake, A.: Grabcut: Interactive foreground extraction using iterated graph cuts. vol. 23 (2004)
42. Saito, S., Huang, Z., Natsume, R., Morishima, S., Kanazawa, A., Li, H.: Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2304–2314 (2019)
43. Santesteban, I., Otaduy, M.A., Casas, D.: Learning-Based Animation of Clothing for Virtual Try-On. Computer Graphics Forum (Proc. Eurographics) (2019)
44. Starck, J., Hilton, A.: Surface capture for performance-based animation. IEEE Computer Graphics and Applications **27**(3), 21–31 (2007)
45. Stuyck, T.: Cloth Simulation for Computer Graphics. Synthesis Lectures on Visual Computing, Morgan & Claypool Publishers (2018)
46. Tao, Y., Zheng, Z., Zhong, Y., Zhao, J., Quionhai, D., Pons-Moll, G., Liu, Y.: Simulcap : Single-view human performance capture with cloth simulation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (jun 2019)
47. Tung, T., Nobuhara, S., Matsuyama, T.: Complete multi-view reconstruction of dynamic scenes from probabilistic fusion of narrow and wide baseline stereo. In: IEEE 12th International Conference on Computer Vision, ICCV. pp. 1709–1716. Kyoto, Japan (Sep 2009)
48. Wang, H., Hecht, F., Ramamoorthi, R., O'Brien, J.F.: Example-based wrinkle synthesis for clothing animation. ACM Transactions on Graphics (Proceedings of SIGGRAPH) **29**(4), 107:1–8 (Jul 2010)
49. Wang, H., Ramamoorthi, R., O'Brien, J.F.: Data-driven elastic models for cloth: Modeling and measurement. ACM Transactions on Graphics (Proceedings of SIGGRAPH) **30**(4), 71:1–11 (Jul 2011)
50. Wang, T.Y., Ceylan, D., Popovic, J., Mitra, N.J.: Learning a shared shape space for multimodal garment design. ACM Trans. Graph. **37**(6), 1:1–1:14 (2018)
51. White, R., Crane, K., Forsyth, D.A.: Capturing and animating occluded cloth. ACM Trans. Graph. **26**(3), 34 (2007)
52. Xiang, D., Joo, H., Sheikh, Y.: Monocular total capture: Posing face, body, and hands in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 10965–10974 (2019)
53. Xu, H., Li, J., Lu, G., Zhang, D., Long, J.: Predicting ready-made garment dressing fit for individuals based on highly reliable examples. Computers & Graphics (2020)
54. Xu, Y., Zhu, S.C., Tung, T.: Denserac: Joint 3d pose and shape estimation by dense render and compare. In: International Conference on Computer Vision (2019)
55. Yamaguchi, K.: Parsing clothing in fashion photographs. In: Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). p. 3570–3577. CVPR '12, IEEE Computer Society, USA (2012)
56. Yamaguchi, K., Kiapour, M.H., Berg, T.L.: Paper doll parsing: Retrieving similar styles to parse clothing items. In: IEEE International Conference on Computer Vision, ICCV 2013, Sydney, Australia, December 1-8, 2013. pp. 3519–3526. IEEE Computer Society (2013)
57. Yang, J., Franco, J.S., Hétroy-Wheeler, F., Wuhrer, S.: Analyzing clothing layer deformation statistics of 3d human motions. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) Computer Vision – ECCV 2018. pp. 245–261. Springer International Publishing, Cham (2018)

58. Yang, W., Luo, P.and Lin, L.: Clothing co-parsing by joint image segmentation and labeling (2014)
59. Yu, T., Zheng, Z., Guo, K., Zhao, J., Dai, Q., Li, H., Pons-Moll, G., Liu, Y.: Doublefusion: Real-time capture of human performances with inner body shapes from a single depth sensor. In: The IEEE International Conference on Computer Vision and Pattern Recognition(CVPR). IEEE (June 2018)
60. Zhang, C., Pujades, S., Black, M., Pons-Moll, G.: Detailed, accurate, human shape estimation from clothed 3D scan sequences. In: IEEE CVPR (2017)
61. Zheng, Z., Yu, T., Wei, Y., Dai, Q., Liu, Y.: Deephuman: 3d human reconstruction from a single image. In: The IEEE International Conference on Computer Vision (ICCV) (October 2019)
62. Zhu, H., Cao, Y., Jin, H., Chen, W., Du, D., Wang, Z., Cui, S., Han, X.: Deep fashion3d: A dataset and benchmark for 3d garment reconstruction from single images. arXiv preprint arXiv:2003.12753 (2020)