Towards 3D Human Shape Recovery Under Clothing

Xin Chen[1] Anqi Pang[1] Yu Zhu[2] Yuwei Li[1] Xi Luo[1]
Ge Zhang[1] Peihao Wang[1] Yingliang Zhang[1] Shiying Li[1] Jingyi Yu[1]
[1]ShanghaiTech University [2]DGene Digital
{chenxin2,liyw,luoxi,lishy1,yujingyi}@shanghaitech.edu.cn yu.zhu@plex-vr.com

Abstract

We present a learning-based scheme for robustly and accurately estimating clothing fitness as well as the human shape on clothed 3D human scans. Our approach maps the clothed human geometry to a geometry image that we call clothed-GI. To align clothed-GI under different clothing, we extend the parametric human model and employ skeleton detection and warping for reliable alignment. For each pixel on the clothed-GI, we extract a feature vector including color/texture, position, normal, etc. and train a modified conditional GAN network for per-pixel fitness prediction using a comprehensive 3D clothing. Our technique significantly improves the accuracy of human shape prediction, especially under loose and fitted clothing. We further demonstrate using our results for human/clothing segmentation and virtual clothes fitting at a high visual realism.

1. Introduction

With the availability of commodity 3D scanners such as Microsoft Kinect and, most recently, mobile 3D scanners based on structured light and time-of-flight imaging, it has become increasing common to create 3D human models in place of traditional 3D images. For example, KinectFusion [24] and DoubleFusion [18] produce high quality 3D scans using a single 3D sensor whereas more sophisticated dome systems [5] acquire dynamic models with textures. However, nearly all existing approaches conduct reconstruction without considering the effects of clothing, or more precisely the fitness of clothing. In reality, human body geometry and clothing geometry covering the body can exhibit significant variations: borrowing jargon from clothing manufactures, clothing can be loose - large clothing-body gaps to allow a full range motion, fitted - a slimmer, athletic cut eliminating the bulk of extra fabric, and compression - ultra-tight, second-skin fit.

The focus of this paper is to robustly and accurately estimate clothing fitness from the acquired 3D human models. Figure 1 shows our estimated fitness, underlying body shape with its clothing of a human. Applications are numerous, ranging from more accurate body shape estimation under clothing, clothing-body segmentation, clothing simulations, etc. Previous approaches have focused on approximating human shape from a single or multiple viewpoints, preferable with fitted or compression clothing. These approaches adopt optimization schemes to estimate the best model that fits the imagery, pose, and motion data, by assuming clothing a thin and fit layer over the skin. In reality, the looseness of clothing greatly affects the accuracy and robustness of the measure. Figure 2 shows an example of body shape of the same human body, but under jacket, robes, t-shirt. The results exhibit strong variations while ignoring the fitness of clothing.

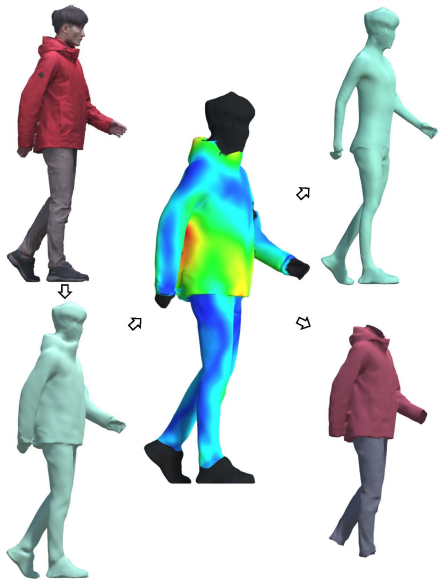

Figure 1. Our fitness measure and body shape recovery scheme. Using a clothed 3D scan (upper left) as input, we fit a parametric model (bottom left) using geometry images and measure clothing fitness (center). The results can be used to automatically segment clothing from the 3D scan as well as to predict the underlying human body shape.

Figure 2. Results using direct parametric body fitting. Top row: the same subject wearing clothing with drastically different fitness. Bottom row: the parametric body fitting results using SMPL[22]. The body estimation exhibits severe inconsistency caused by clothing.

Different from previous approaches, we focus on simultaneously modeling the fitness of clothing and human body shape. We observe that humans can quickly identify clothing fitness (loose vs. fit vs. compression) as important prior to shape estimation and seek to develop a similar learning-based pipeline. Specifically, we set out to combine global and local inferences: the former includes clothing styles and types and the latter includes shape deformations such as folds and puffiness.

We first present a data-driven clothing fitness estimation scheme that considers clothing type and geometry as well as human pose. The input to our scheme is a 3D mesh of clothed human and we set out to map it to a geometry image [9] that we call clothed-GI. We extend the parametric human model SMPL [22] by aligning key geometry features (e.g., joints, hands, head, feet, etc) so that human models of different shapes and under different clothing can be uniformly analyzed via clothed-GI. By employing skeleton and joint warping, our alignment scheme supports a large variety of poses in addition to the neutral A or T pose. For each pixel on the clothed-GI, we extract a feature vector including color/texture, curvature, position, normal, etc and set out to predict its fitness measure. We use the vectors between each corresponding vertex pair of body and clothing geometry as the fitness measure and train a modified conditional GANs network for per-pixel fitness prediction. Finally, we use the fitness measure to predict human shape under clothing.

We collect a large 3D data set that consists of a large variety of clothing: t-shirt, jacket, down jacket, jocker pants, trousers, etc. The data set provides the ground truth for training and testing. Comprehensive experiments show that, compared with the state-of-the-art, our technique significantly improves the accuracy of human shape prediction especially under loose and fitted clothing. We further demonstrate how the recovered human geometry can also be used to automatically segment clothing from human body on 3D meshes as well as virtual clothes fitting at a high visual realism.

2. Related Work

The literature on 3D human body shape estimation is vast and we only review the most relevant ones. Most works can be categorized as multi-view stereo vs. depth fusion based approaches. The former employs correspondence matching and triangulation [8, 33, 25], assisted by visual SLAM. The most notable work is the multi-view dome setting from the CMU group composed of 600 cameras that can reconstruct realistic single or multiple 3D humans [14, 15, 38]. The latter uses active sensors such as structured light and time-of-flight range scanning (e.g, Microsoft Kinect I and II, respectively) and are of a much lower cost [3, 23, 41, 6]. Newcombe *et al*. [23] compensate geometric changes due to motion captured from a single RGB-D sensor. Yu *et al*. [42] present a single view system to reconstruct cloth geometry and inner body shape based on the parametric body model. Their approach allows the subject to wear casual clothing and separately treat the inner body shape and the outer clothing geometry.

Estimating body shape under clothing is more challenging. Existing methods employ a statistical or parametric 3D body model, e.g., SCAPE [1] and SMPL [22], and require the subject wearing minimal or fitted clothing. The earlier work by [2] builds on the concept of visual hull under an assumption that the clothing becomes looser or tighter on different body parts as a person moves. They estimate a maximal silhouette-consistent parametric shape (MSCPS) from several images of a person with both minimal and normal clothing. Wuhrer *et al*. [37] estimate body shape from static scans or motion sequences by modeling body shape variation with a skeleton-based deformation. Their method requires fitted clothing. In general, human body shape estimation in wide and puffy clothing is significantly more difficult than in fited clothing since, even for humans. More recent approaches attempt to align clothing on the human body model [11, 43, 29]. Our approach also aims to align a parametric model but we employ the geometry image analysis and exploit fitness prediction for more reliable shape prediction.

[28, 36, 26] learn articulated body poses of humans from their occluded body parts via sequential convolutional networks (ConvNets). [20] predicts body segments and landmarks from annotated human pose datasets, and conducts body pose estimation with clothing and 3D body fitting. Lassner *et al*. [19] present a generative model of full body in clothing, but their work focuses more on appearance generation than body shape estimation. Pavlakos *et al*. [27] propose a ConvNet based method with parameterized body model to generate a detailed 3D mesh from a single color image, and refine the mesh by projecting it back to the 2D

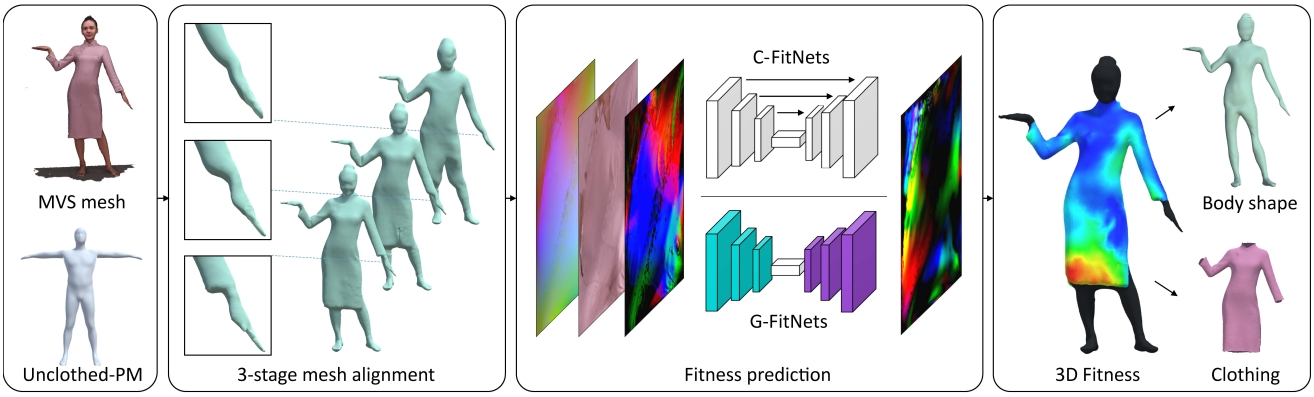

Figure 3. Our learning based pipeline fitness prediction and body shape estimation under clothing.

image for full body pose and shape estimation. Their technique relies on parameter prediction from the body model and body pose training data. Most recently, [40] exploits the relation between clothing and the underlying body movement with data collected as the same actor in the same clothing, the same actor in different clothing, and different actors in the same type of clothing. Pons-Moll *et al*. [29] estimate a minimally clothed shape and uses retargeting to map the body shape in different sizes and poses. Lähner *et al*. [16] propose a data-driven framework based on body motion. Instead of modeling the variations of clothing and underlying body, our approach builds a GAN based technique to learn the fitness between the human body and the clothing layer. We then estimate both inner body shape and clothing segmentation from our new dataset of different subjects in different clothing styles.

3. Body Shape Alignment

Our goal is to estimate fitness for human shape estimation along with clothing segmentation. Figure 3 illustrates the pipeline of our proposed approach for each body mesh. The major difference from the previous approaches such as [40] is that we directly predict the fitness level to represent the relation between underlying human body and its clothing layer. To align an unclothed (template) model to the scanned clothed one, we apply shape deformation and then compare the geometry images between the two for training and prediction.

3.1. Input 3D Meshes

We also use the multi-view dome system to acquire the initial 3D meshes, although we can also use state-of-the-art KinectFusion or DoubleFusion to generate the input. Our dome system consists of 80 calibrated video cameras. We have recruited 18 subjects, 9 males and 9 females. 10 subjects are reconstructed under the canonical "A" or "T" static poses and 8 subjects are captured in free movements. We use a total of 227 pieces of clothing. On average each static subject wears around 20 different pieces of clothing, ranging from fit to puffy, light to heavy. For each dynamic subject, we capture 400 500 frames. We also captured their corresponding unclothed models.

We reconstruct the 3D mesh for each human body using [30] to construct a new dataset. We further extract the skeleton points of human body and hands in each 2D image using [4, 31], and conduct 3D point triangulation [35] to recover a 3D skeletal structure for each human body[35].

3.2. Deformation-based Alignment

Our goal is to map a parametric model to the recovered real human shapes. We use the SMPL parametric model. Since we need later deform the shape to the target real human, we first smooth out the detail geometry on the head, feet and hands, e.g., ears, nose and fingers and at the same time we make these parts more densely sampled so that we can warp corresponding vertices to the target with details, as shown in Figure 4 (a). We bind the genus zero body model with the skeleton from OpenPose [4, 31], 23 joints for the body and 21 joints for each hand (in Figure 4 (b)). This results in a remeshed body model $\mathcal{M}_T$, which we call unclothed parametric model(unclothed-PM), with $N_{\mathbf{M}} = 14985$ vertices, $N_{\mathbf{F}} = 29966$ facets and $N_{\mathbf{J}} = 65$ joints, as defined in Equation (1).

$$\mathcal{M}_{\mathrm{T}} = \left\{\mathbf{M} \in \mathbb{R}^{N_{\mathbf{M}} \times 3}, \mathbf{F} \in \mathbb{R}^{N_{\mathbf{F}} \times 3}, \mathbf{J} \in \mathbb{R}^{N_{\mathbf{J}} \times 4+3}\right\} \quad (1)$$

where we use capital and bold letters to denote matrices $\mathbf{F}$ for facets, $\mathbf{J}$ for joints and $\mathbf{M}$ for vectors of mesh vertices, and capital letters to denote function (e.g. $M(\cdot)$). The joints in the unclothed-PM are:

$$\mathbf{J} = \left\{\mathbf{\Theta} \in \mathbb{R}^{N_{\mathbf{J}} \times 3}, \mathbf{S} \in \mathbb{R}^{N_{\mathbf{J}}}, \mathbf{m} \in \mathbb{R}^{3}\right\} \quad (2)$$

where we parameterize each rotation in $\mathbf{\Theta}$ using the axis-angle representation. $\mathbf{S}$ is the scaling of each joint along the skeleton direction, $\mathbf{m}$ the global translation.

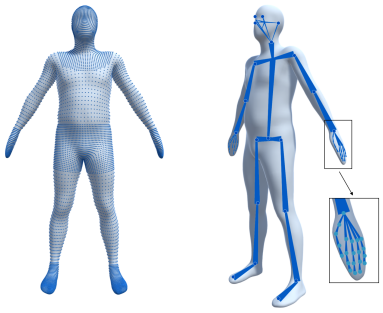

Figure 4. (a) We refine the classical parametric body template by applying remeshing to make respective body parts more densely sample for more reliably shape alignment with the clothed ones. (b) The skeleton and joints on the parametric body model.

We then optimize the mesh of the unclothed-PM following the embedded deformation (ED) graph [34].

$$\mathcal{G} = \left\{ \mathbf{R} \in \mathbb{R}^{N_\mathbf{G} \times 3}, \mathbf{t} \in \mathbb{R}^{N_\mathbf{G} \times 3} \right\} \tag{3}$$

As shown in Equation (3), ED graph of the body mesh has $N_\mathbf{G}$ nodes. The warping field G_k of each node consists of rotation $\mathbf{R}_k \in \mathbf{SO}(3)$ and translate $\mathbf{t_k} \in \mathbb{R}^\mathbf{3}$, where $k \in [0, N_G)$, and can be formulated as

$$G_k(\mathbf{v}) = \mathbf{R}_k(\mathbf{v} - \hat{\mathbf{g}}_k) + \hat{\mathbf{g}}_k + \mathbf{t}_k, \tag{4}$$

where $\hat{\mathbf{g}}_k \in \mathbb{R}^3$ indicates the canonical position at node k. A deformed vertex $\mathbf{v}_i, i \in [0, N_\mathbf{M})$ on ED graph $\mathcal{G}$ is denoted by

$$\mathbf{v}_i(\mathcal{G}) = G(\hat{\mathbf{v}}_i) = \sum_{k \in N_\mathbf{G}} \mathbf{w}^\mathbf{G}_{i,k} G_k(\hat{\mathbf{v}}_i). \tag{5}$$

where $\hat{\mathbf{v}}_i$ indicates the canonical position at vertex i, $\mathbf{w}^\mathbf{G}_{i,k}$ the blending weight between vertex i and graph node k. We compute these weights based on Euclidean metric of n-nearest nodes as in [34].

We further warp the unclothed-PM $\mathcal{M}_T$ with the skeletal structure $\mathbf{J}_{mv}$ from multi-view images and compute its vertex matrix as $\mathbf{M}_{warp} = M(\mathbf{J}_{mv})$ for the initial deformation in a specific pose.

We adopt a 3-stage deformation scheme, as shown in Figure 5.

Silhouette deformation. For a 3D mesh of human body in our dataset, we set up a virtual system $\mathcal{C}$ in Equation (6) with $N_\mathcal{C} = 30$ virtual cameras to view different parts of the mesh. For each of five parts at head, feet and hands, we set two cameras orthogonal to each other. And for the front and back sides of the torso mesh, we arrange five cameras at equal intervals for the upper and lower parts, respectively.

$$\mathcal{C} = \left\{ \left(\mathbf{c}_j \in \mathbb{R}^6, w^\mathcal{C}_j \in \mathbb{R}^1 \right) | j \in [0, N_\mathcal{C}) \right\}, \tag{6}$$

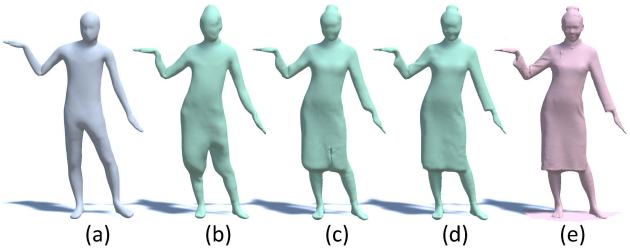

Figure 5. Silhouette-guided alignment. We deform the parametric model (a) to match the target clothed model (e). (b) to (d) show the intermediate results. The deformation based approach preserves topological consistency on their corresponding geometry images.

where $\mathbf{c}_j$ denotes extrinsic parameters of a camera. $w^\mathcal{C}_j \in [0.5, 1]$ represents the camera positions, eg., 0.5 for the torso and 1 for the left hand.

We then render the 3D mesh and generate a high-quality silhouette of the body mesh at each virtual view. Similar to [39], we implement the multi-view silhouette generation with energy function as shown in Equation (7):

$$E_\mathrm{S}(\mathcal{G}) = E^\mathrm{S}_\mathrm{mv}(\mathcal{G}) + \lambda^\mathrm{S}_\mathrm{reg} E^\mathrm{S}_\mathrm{reg}(\mathcal{G}) \tag{7}$$

where E_mv represents the data term for the multi-view silhouette generation, E_reg a regularization term as defined in [32]. For the data term,

$$E^\mathrm{S}_\mathrm{mv}(\mathcal{G}) = \sum_{j \in \mathcal{C}} \frac{w^\mathcal{C}_j}{|\mathbf{v}^\mathrm{S}_j|} \sum_{k \in \mathbf{v}^\mathrm{S}_j} |\mathbf{n}^\mathrm{T}_k (P_j(\mathbf{v}_i(\mathcal{G})) - \mathbf{p}_k)|^2, \tag{8}$$

where $\mathbf{v}^\mathrm{S}_j$ is the vertex set of virtual silhouettes, $P_j(\cdot)$ is the projection function of camera j. For each deformed vertex $\mathbf{v}_i$, its corresponding silhouette point is $\mathbf{p}_k \in \mathbb{R}^2$ with normal $\mathbf{n}_k \in \mathbb{R}^2$. We search for the corresponding points via the Iterative Closest Point (ICP) algorithm. And for the regularization term,

$$E^\mathrm{S}_\mathrm{reg}(\mathcal{G}) = \sum_{k \in \mathcal{G}} \sum_{n \in \mathcal{N}_k} w^\mathcal{N}_{k,n} \| (\mathbf{g}_k - \mathbf{g}_n) - \mathbf{R}_k (\hat{\mathbf{g}}_k - \hat{\mathbf{g}}_n) \|^2_2 \tag{9}$$

where $\mathcal{N}_k \in \mathcal{G}$ is the 1-ring neighbourhood of graph node k, $w^\mathrm{N}_{k,n}$ the weight between the nodes k, n.

ED graph-based none-rigid deformation. we exploit a global non-rigid deformation with ED graph to deform our silhouette deformation result $\mathbf{M}_S$ to the 3D mesh. We resample the ED graph of $\mathcal{M}_T$ with more nodes, but remain $\mathcal{G}$ for clarity. The energy function is defined as

$$E_\mathrm{D}(\mathcal{G}) = E^\mathrm{D}_\mathrm{data}(\mathcal{G}) + \lambda^D_\mathrm{reg} E^\mathrm{D}_\mathrm{reg}(\mathcal{G}) \tag{10}$$

where data term $E^\mathrm{D}_\mathrm{data}(\mathcal{G})$ is for the deformation and regularization term $E^\mathrm{D}_\mathrm{reg}(\mathcal{G})$ the same as in Equation (9). The data

term in Equation (10) is as

$$E^{\mathrm{D}}_{\mathrm{data}}(\mathcal{G}) = \lambda^{\mathrm{D}}_{\mathrm{point}} \sum_{i \in \mathbf{M}} \|\mathbf{v}_i(\mathcal{G}) - \mathbf{v}^c_i\|^2 + \lambda^{\mathrm{D}}_{\mathrm{plane}} \sum_{i \in \mathbf{M}} (\mathbf{n}^{\mathrm{T}}_i(\mathcal{G})(\mathbf{v}_i(\mathcal{G}) - \mathbf{v}^c_i)) \tag{11}$$

where $\lambda^{\mathrm{D}}_{\mathrm{point}}$ is the weight for point-to-point distance, $\lambda^{\mathrm{D}}_{\mathrm{plane}}$ the weight for point-to-plane distance. $\mathbf{n}_i(\mathcal{G})$ represents the normal of the deformed vertex $\mathbf{v}_i(\mathcal{G})$, and its corresponding point $\mathbf{v}^c_i$ similar to ICP.

Per-vertex non-rigid deformation. We finally refine the deformation from ED graph-based none-rigid result $\mathbf{M_D}$ to the 3D mesh via per-vertex optimization as in Equation (12), with its data term in Equation (13).

$$E_{\mathrm{V}}(\mathbf{M}) = E^{\mathrm{V}}_{\mathrm{data}}(\mathbf{M}) + \lambda^{\mathrm{V}}_{\mathrm{reg}} E^{\mathrm{V}}_{\mathrm{reg}}(\mathbf{M}) \tag{12}$$

$$E^{\mathrm{V}}_{\mathrm{data}}(\mathbf{M}) = \lambda^{\mathrm{V}}_{\mathrm{point}} \sum_{i \in \mathbf{M}} \|\mathbf{v}_i - \mathbf{v}^c_i\|^2 + \lambda^{\mathrm{V}}_{\mathrm{plane}} \sum_{i \in \mathbf{M}} (\mathbf{n}^{\mathrm{T}}_i(\mathbf{v}_i - \mathbf{v}^c_i)) \tag{13}$$

After refinement, we obtain a clothed body model, namely clothed-PM. The clothed-PM is a pose-dependent 3D mesh of human body in clothing, we can calculate the fitness if given its underlying body shape of the mesh or given its clothing segment. We show later that clothed-PM enables reliable shape estimation under clothing and clothing segmentation.

3.3. Clothed Geometry Images

Our deformation schemes enable direct alignment between the source and the target models. However, to measure the fitness and subsequently conduct training, we need to represent the two models suitable for training. We exploit the geometry images [9]. There are different ways of forming geometry images. In our case, our 3D clothed-PM mesh should maintain shape while preserving fitness measures. We first compare the fitness at vertices or a patch of the same body parts on several meshes, eg., shoulders and waists, then manually cut the mesh along a path where the fitness variation on its geometry image becomes smaller. Most recently, Li *et al*. [21] propose an optimization approach to automatically seek a satisfactory seamline with low discontinuity artifacts. We also use their OptCuts algorithm to cut our clothed-PM mesh into its clothed-GI for comparisons.

We adopt a similar notation of [7]. For each $i \in \{1, \ldots, n\}$, choose any set of real numbers $\lambda_{i,j}$ for $j = 1, \ldots, N$ such that

$$\lambda_{i,j} = 0, \quad (i,j) \notin E, \qquad \lambda_{i,j} > 0, \quad (i,j) \in E, \quad \sum_{j=1}^{N} \lambda_{i,j} = 1 \tag{14}$$

where $(i,j) \in E$ represents neighbouring vertices at i and j, and different $\lambda_{i,j}$ corresponding to different mappings. We then immobilize vertices at the boundary, and define $\mathbf{u}_1, \ldots, \mathbf{u}_n$ to be the solutions of the linear system of equations as in Equation (15).

$$\mathbf{u}_i = \sum_{i=1}^{N} \lambda_{i,j} \mathbf{u}_j, \quad i = 1, \ldots, n \tag{15}$$

Our clothed-GI preserves key information in its 3D mesh at each vertex. From the clothed-GI, we build a 2D feature map with 9 channels of positions, normals and RGB colors at a vertex, and refine this map via linear interpolation as input to our FitNets.

4. Fitness Prediction

We explore a new concept, fitness, to describe the relation between the naked body shape and its clothing, and propose a ConvNets based and GAN based architecture to predict the fitness on our new dataset for both body shape estimation and clothing segmentation.

4.1. Fitness Measure

We define the fitness $\mathcal{F}$ as a vector in three channels in Equation (16). In Euclidean geometry, the direction of $\mathcal{F}$ is from an unclothed body model $\mathbf{M}$ with $N_{\mathbf{M}}$ vertices to a clothed body mesh $\mathbf{M^c}$ with $N_{\mathbf{M^c}}$ vertices, and its magnitude the Euclidean distance between the corresponding vertices on the two models. More specifically, the fitness $\mathbf{F}_i$ at a vertex i is calculated in Equation (17).

$$\mathcal{F} = \left\{ \mathbf{F} \in \mathbb{R}^{N_{\mathbf{M}} \times 3} \right\} \tag{16}$$

$$\mathbf{F}_i = \mathbf{v}^c_i - \mathbf{v}_i \tag{17}$$

where $\mathbf{v}^c_i$ and $\mathbf{v}_i$ represent the corresponding vertices at i from the clothed and unclothed body models, $\mathbf{v}_i, i \in [0, N_{\mathbf{M}})$.

This issue results in a critical problem of correspondence matching. In practice, however, it is challenging since humans generally wear more casual clothing than compression or fit. The relation between the body shape and its clothing layer also relies on many factors, eg., body pose, motion, clothing materials, thickness, etc.

Our clothed-PM enables calculating the fitness between its underlying body shape and clothing layer of the mesh, given either ground truth of the two. We calculate the fitness $\mathcal{F}_i$ at a vertex i of body model as in Equation (18).

$$\mathbf{F_i} = \frac{\sum_{\mathbf{v_r^c} \in \mathcal{N}_1^c} K_G(\mathbf{v_r^c} - \mathbf{v_i}) + \sum_{\mathbf{v_s^c} \in \mathcal{N}_2^c} K_G(\mathbf{v_s^c} - \mathbf{v_i})}{||\mathcal{N}_1^c|| + ||\mathcal{N}_2^c||} \tag{18}$$

where v_i refers to a vertex at i of the ground truth of inner body shape. $\mathcal{N}_1^c$ is the set of closest vertices ray-traced forward along the normal direction of v_i with apex angle of 15 degrees, and $\mathcal{N}_2^c$ the set of closest vertices on the clothed-PM mesh ray-traced backward.

In this work, we construct a learning based approach to automatically predict fitness from our clothed-PM mesh of a human body when unknown its ground truth of inner body shape or clothing segment.

4.2. FitNets Architecture

We construct two network architectures, ConvNets based and GANs based, to predict fitness, which we name C-FitNets and G-FitNets for short, respectively.

Our C-FitNets consists of both encoder and decoder for a high-resolution fitness map. We improve the fully-convolutional regression networks (FCRN) [10] by combining the ResNet-50 [12] . We also modify its setting of input data to fit our feature map and its up-projection [17] to simplify the structural complexity. The loss function for regression is usually to minimize a squared Euclidean distance between predicted and observed. Similar to the FCRN, we use BerHu loss for a best pixelwise fitness map.

$$\mathcal{B}(x) = \begin{cases} |x| & |x| \leq c \\ \frac{x^2+c^2}{2c} & |x| > c \end{cases} \tag{19}$$

Alternatively, we modify a conditional GANs [13] to construct our G-FitNets. We change the input channels and recursive depth of the network. The conditional GANs learns a mapping to generate an output from an input image x and random noise vector z to output image y $G : \{x, z\} \rightarrow y$ which is indistinguishable by the adversarial discriminator D, which is trained to detect the output fake from the input. Following [13], the objective function is defined as

$$\begin{aligned} \mathcal{L}_{GAN}(G, D) =& \mathbb{E}_{x,y}[\log D(x, y)] + \\ & \mathbb{E}_{x,z}[\log(1 - D(x, G(x, z))] \end{aligned} \tag{20}$$

where G minimizes $\mathcal{L}_{GAN}$ while the adversarial D maximizes it. We use L1 distance, rather than L2, for fewer blurring artifacts.

$$\mathcal{L}_{L1}(G) = \mathbb{E}_{x,y,z}\left[\|y - G(x, z)\|_1\right] \tag{21}$$

$$G^* = \arg\min_G \max_D \mathcal{L}_{GAN}(G, D) + \lambda \mathcal{L}_{L1}(G) \tag{22}$$

Figure 6. Sample training data from our dataset: three real human models (human body models individually scanned with and without clothing; the clothes model is manually segmented) and one synthetic model (right most).

Fitness between the underlying body shape and its clothing layer is originally at each vertex of the 3D mesh, we therefore transform the 2D fitness map from our FitNets back to 3D via the inverse mapping of Equation (15) in Sec. 3.3, and calculate the displacement $\mathbf{d}_i$ between the clothed-PM and its former unclothed-PM in Equation (23).

$$\mathbf{d}_i = \mathbf{v}_i^{\mathbf{V}} - \mathbf{v}_i^{\mathbf{W}}, \mathbf{v}_i^{\mathbf{V}} \in \mathbf{M}_{\mathbf{V}}, \mathbf{v}_i^{\mathbf{w}} \in \mathbf{M}_{\text{warp}} \tag{23}$$

where $\mathbf{M}_{\text{warp}}$ is the vertex matrix of unclothed-PM $\mathcal{M}_{\text{warp}}$, $\mathbf{M}_{\text{V}}$ the vertex matrix of clothed-PM $\mathcal{M}_{\text{V}}$. We assign the fitness at the vertices whose direction is opposite to zero, and refine the 3D fitness via a local Gaussian filter.

5. Experimental Results

We carry out comprehensive experiments to validate our method on a new dataset and the Bodies Under Flowing Fashion (BUFF) dataset in [43], and evaluate the performance of our networks.

5.1. Dataset

Figure 6 shows our dataset that consists of 4k+ 3D meshes reconstructed via the MVS method [30] from multi-view images (some of which undone from dynamic sequences) of 18 subjects in different clothing, as described in Sec. 3.1. We also generate 2k+ 3D meshes of synthetic avatars in different clothing with Adobe Fuse CC. The clothing for both the subjects and the avatars varies in styles, materials, length, size and heaviness, eg., long or short shirts, long or short down coat, long or short pants, etc. The 3D meshes in the dataset is pose-dependent, each has 50,000 vertices and 100,000 facets. We categorize 80% of the 3D meshes, reconstructed or synthesized, for training set and 20% for testing set. Note that we manually segment the head, feet and hands of the 3D meshes when input to our FitNets since these parts are uncovered in the clothing.

For ground truth body shape, we further capture images of the subjects in minimal clothing, and manually adjust

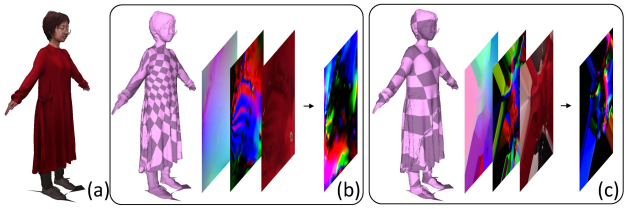

Figure 7. Feature maps and fitness measures on the clothed-GI of a sample 3D clothed model (a). We show the results using two types of GI generation schemes: regular texture mapping (b) and conformal geomtry images (c). Their corresponding feature maps are based on position, normal, and color.

their poses to the same ones when they are in other clothing styles. We ask 5 artists to annotate clothing segments of a 3D mesh and average them as its ground truth clothing segment. We then generate the fitness between the body shape and clothing segment of ground truth by employing the algorithm described in Sec. 4.1.

5.2. Evaluations

Using our networks, we qualitatively and quantitatively evaluate our experimental results in fitness prediction, underlying body shape estimation and clothing segmentation, in comparison with the original FCRN [10] on our dataset and the BUFF dataset [43].

Fitness prediction Our input feature map and output fitness map are 224×224×9 and 224×224×3 in size, respectively. We visualize these maps in colors. Figure 7 shows the clothed-GI manually cut by an artist or with the OptCut algorithm from a 3D mesh in our dataset, the feature maps of positions, normals, and RGB colors, and the fitness map predicted from our networks. The directions and magnitudes of fitness in intensity match with the feature map, visually similar to the feature map of normals.

We employ four error metrics, such as $L1$ and $L2$, to quantitatively evaluate the fitness map, and normalize the values of fitness ranging from -10 to 10. The mean-squared error (MSE) from $L1$ and $L2$ norms between the predicted fitness map y and ground truth y is computed as:

$$\text{MSE}(y, \hat{y}) = \frac{1}{wh} \sum_{i=0}^{wh} (\hat{y}_i - y_i)^2 \tag{24}$$

Average $\log_{10}$ error (log10) and the average relative error (ARE) are often used for depth prediction, computed as:

$$\log_{10} \text{error}(y, \hat{y}) = \frac{1}{wh} \sum_{i=1}^{wh} |\log_{10} y_i - \log_{10} \hat{y}_i| \tag{25}$$

$$\text{rel}(y, \hat{y}) = \frac{1}{wh} \sum_{i=1}^{wh} \frac{|y_i - \hat{y}_i|}{y_i} \tag{26}$$

Table 1 shows four metrics in three networks, C-FitNets, G-FitNets, and the original FCRN. "regular" and "optcut" denote the input clothed-GI cut manually by an artist and with the OptCut algorithm, respectively. Our G-FitNets perform best amongst the three networks.

Table 1. Comparison of four metrics in different networks withj regular and conformal clothed-GI.

Methods	L1	log10	ARE	L2
G-FitNets regular	0.109	**0.0096**	**0.0093**	0.102
C-FitNets regular	0.415	0.0342	0.0346	0.489
FCRN regular	0.189	0.0160	0.0159	0.142
G-FitNets optcut	**0.071**	0.0062	0.0060	**0.075**
C-FitNets optcut	0.126	0.0107	0.0105	0.120
FCRN optcut	0.117	0.0098	0.0096	0.111

Body shape estimation under clothing. We have investigated physical body shapes statistically in our 3D mesh dataset, and observed that body shape has a close relation with the clothing covering it subject to fitness. In this work, we initially set the fitness of human skin to zero, and use solely one mesh to estimate its body shape. With the fitness predicted, we estimate the body shape of a mesh along with its clothing segment from the clothed-PM. Table 2 demonstrates the per pixel accuracy for the segmentation in three networks, C-FitNets, G-FitNets, and FCRN. The G-FitNets again provides best performance, highly beyond the other two. The results also show that the input clothed-GI cut manually outperform that with the OptCut technique since the latter attempts to minimize the seam length while satisfying distortion bound. And after we get the segment result on 2D, we can then inverse map it to either MVS 3D mesh or DPM.

Table 2. Prediction accuracy comparisons using three different deep networks and two different geometry image inputs (regular and conformal optcut).

Methods	Per-pixel accuracy
G-FitNets regular	**0.9033**
C-FitNets regular	0.8730
FCRN regular	0.8761
G-FitNets optcut	**0.7608**
C-FitNets optcut	0.6015
FCRN optcut	0.6797

We have also conducted experiments on two body meshes in the BUFF dataset [43], as shown in Figure 8. The results as well as those from our dataset show that our new method enables accurate and robust estimation of fitness, body shape and clothing segmentation of a human body in a variety of clothing styles. This may fail when clothes on a body are extremely heavy and puffy and when a body is unshaped, eg., the underlying legs of the third mesh in Figure

Figure 8. Sample results using our technique for fitness prediction, shape estimation and clothes segmentation. From left to right: input 3D meshes, clothed-PM, fitness measure, estimated body shape under clothing, and automatically segmented clothes.

2, as well as in the case body parts such as armpits are occluded. In the supplementary materials, we further demonstrate the recovered shape can be used for virtual clothing fitting.

Computation time and parameters We conduct our experiments on a commonplace setting with CPU Intel Core i7-8700K 3.7GHZ, GPU NVIDIA GeForce GTX 1080Ti 8G. The parameters assigned for stage 1, 2 and 3 are as follows: $N_{\mathbf{G}^S} = 1407$, $\lambda^S_{reg} = 10$; $N_{\mathbf{G}^D} = 2103$, $\lambda^V_{reg} = 7$ $\lambda^D_{point} = 0.5$, $\lambda^D_{plane} = 1.5$; $\lambda^V_{reg} = 1$, $\lambda^V_{point} = 1$, $\lambda^V_{plane} = 1.5$.

For the computation time in our algorithm, training occupies the most, about 5 hours, but this only conducts once for the learning process. The 3-stage silhouette-based deformation costs up to 6 seconds, and fitness prediction about 10 seconds.

6. Conclusion and Future Work

We have presented a learning-based scheme for robustly and accurately estimating clothing fitness as well as human geometry on clothed 3D human scans. At the core of our approach is the use of geometry images to align clothed human geometry under various types of clothing. Specifically, we have collected a large 3D clothing dataset, tailored features, and trained a modified conditional GAN network to automatically determine clothing fitness and subsequently the underlying human shape. Comprehensive experiments have shown that our approach is reliable and accurate. We have further demonstrated using our scheme for automatic human/clothing segmentation, virtual clothes fitting, and clothing motion simulation. Since both clothing geometry and human shape geometry are obtained from real data, we manage to produce results at a high visual realism.

Since our model is based on geometry images, our technique by far can only handle genus 0 human geometry. In reality, human geometry can form much more complex topology and more sophisticated geometry image generation and subsequently alignment schemes should be developed and are our immediate future work. Our technique uses 3D clothed models as input. In the future, we plan to explore the possibility of directly using a single or a sparse set of 2D images. The challenge there is how to reliably infer 3D clothing geometry from images when accurate 3D reconstruction is unavailable. By augmenting our training as well as using lighting estimation, it may be possible to directly clothing fitness from a single 2D image.

References

[1] D. Anguelov, P. Srinivasan, D. Koller, S. Thrun, J. Rodgers, and J. Davis. Scape: Shape completion and animation of people. *ACM Transactions on Graphics (TOG)*, 24(3):408–416, 2005.

[2] A. O. Bălan and M. J. Black. The naked truth: Estimating body shape under clothing. In *Proceedings of the European Conference on Computer Vision*, pages 15–29. Springer, 2008.

[3] F. Bogo, M. J. Black, M. Loper, and J. Romero. Detailed full-body reconstructions of moving people from monocular rgb-d sequences. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2300–2308, 2015.

[4] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[5] A. Collet, M. Chuang, P. Sweeney, D. Gillett, D. Evseev, D. Calabrese, H. Hoppe, A. Kirk, and S. Sullivan. High-quality streamable free-viewpoint video. *ACM Transactions on Graphics (ToG)*, 34(4):69, 2015.

[6] M. Dou, S. Khamis, Y. Degtyarev, P. Davidson, S. R. Fanello, A. Kowdle, S. O. Escolano, C. Rhemann, D. Kim, J. Taylor, et al. Fusion4d: Real-time performance capture of challenging scenes. *ACM Transactions on Graphics (TOG)*, 35(4):114, 2016.

[7] M. S. Floater. Parametrization and smooth approximation of surface triangulations. *Computer Aided Geometric Design*, 14(3):231–250, 1997.

[8] Y. Furukawa, C. Hernández, et al. Multi-view stereo: A tutorial. *Foundations and Trends® in Computer Graphics and Vision*, 9(1-2):1–148, 2013.

[9] X. Gu, S. J. Gortler, and H. Hoppe. Geometry images. *ACM Transactions on Graphics (TOG)*, 21(3):355–361, 2002.

[10] A. Gupta, A. Vedaldi, and A. Zisserman. Synthetic data for text localisation in natural images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2016.

[11] N. Hasler, C. Stoll, B. Rosenhahn, T. Thormählen, and H.-P. Seidel. Estimating body shape of dressed humans. *Computers & Graphics*, 33(3):211–216, 2009.

[12] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the The IEEE Conference on Computer Vision and Pattern Recognition*, June 2016.

[13] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1125–1134, 2017.

[14] H. Joo, T. Simon, X. Li, H. Liu, L. Tan, L. Gui, S. Banerjee, T. Godisart, B. Nabbe, I. Matthews, et al. Panoptic studio: A massively multiview system for social interaction capture. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(1):190–204, 2019.

[15] H. Joo, T. Simon, and Y. Sheikh. Total capture: A 3d deformation model for tracking faces, hands, and bodies. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8320–8329, 2018.

[16] Z. Lahner, D. Cremers, and T. Tung. Deepwrinkles: Accurate and realistic clothing modeling. In *Proceedings of the European Conference on Computer Vision*, pages 667–684, 2018.

[17] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab. Deeper depth prediction with fully convolutional residual networks. In *2016 Fourth International Conference on 3D Vision (3DV)*, pages 239–248. IEEE, 2016.

[18] Z.-z. Lan, L. Bao, S.-I. Yu, W. Liu, and A. G. Hauptmann. Double fusion for multimedia event detection. In *Proceedings of the International Conference on Multimedia Modeling*, pages 173–185. Springer, 2012.

[19] C. Lassner, G. Pons-Moll, and P. V. Gehler. A generative model of people in clothing. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 853–862, 2017.

[20] C. Lassner, J. Romero, M. Kiefel, F. Bogo, M. J. Black, and P. V. Gehler. Unite the people: Closing the loop between 3d and 2d human representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6050–6059, 2017.

[21] M. Li, D. M. Kaufman, V. G. Kim, J. Solomon, and A. Sheffer. Optcuts: Joint optimization of surface cuts and parameterization. *ACM Transactions on Graphics (TOG)*, 37(6), 2018.

[22] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black. Smpl: A skinned multi-person linear model. *ACM Transactions on Graphics (TOG)*, 34(6):248, 2015.

[23] R. A. Newcombe, D. Fox, and S. M. Seitz. Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 343–352, 2015.

[24] R. A. Newcombe, S. Izadi, O. Hilliges, D. Molyneaux, D. Kim, A. J. Davison, P. Kohi, J. Shotton, S. Hodges, and A. Fitzgibbon. Kinectfusion: Real-time dense surface mapping and tracking. In *Proceedings of the IEEE International Symposium on Mixed and Augmented Reality*, pages 127–136. IEEE, 2011.

[25] R. A. Newcombe, S. J. Lovegrove, and A. J. Davison. Dtam: Dense tracking and mapping in real-time. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2320–2327. IEEE, 2011.

[26] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In *Proceedings of the European Conference on Computer Vision*, pages 483–499. Springer, 2016.

[27] G. Pavlakos, L. Zhu, X. Zhou, and K. Daniilidis. Learning to estimate 3d human pose and shape from a single color image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 459–468, 2018.

[28] L. Pishchulin, E. Insafutdinov, S. Tang, B. Andres, M. Andriluka, P. V. Gehler, and B. Schiele. Deepcut: Joint subset partition and labeling for multi person pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4929–4937, 2016.

[29] G. Pons-Moll, S. Pujades, S. Hu, and M. J. Black. Clothcap: Seamless 4d clothing capture and retargeting. *ACM Transactions on Graphics (TOG)*, 36(4):73, 2017.

[30] J. L. Schönberger, E. Zheng, J.-M. Frahm, and M. Pollefeys. Pixelwise view selection for unstructured multi-view stereo. In *Proceedings of the European Conference on Computer Vision*, pages 501–518. Springer, 2016.

[31] T. Simon, H. Joo, I. Matthews, and Y. Sheikh. Hand keypoint detection in single images using multiview bootstrapping. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[32] O. Sorkine and M. Alexa. As-rigid-as-possible surface modeling. In *Proceedings of the Symposium on Geometry Processing*, volume 4, pages 109–116, 2007.

[33] C. Strecha, W. Von Hansen, L. Van Gool, P. Fua, and U. Thoennessen. On benchmarking camera calibration and multi-view stereo for high resolution imagery. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. Ieee, 2008.

[34] R. W. Sumner, J. Schmid, and M. Pauly. Embedded deformation for shape manipulation. *ACM Transactions on Graphics (TOG)*, 26(3):80, 2007.

[35] B. Triggs, P. F. McLauchlan, R. I. Hartley, and A. W. Fitzgibbon. Bundle adjustment a modern synthesis. In *Proceedings of the International Workshop on Vision Algorithms*, pages 298–372. Springer, 1999.

[36] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. Convolutional pose machines. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4724–4732, 2016.

[37] S. Wuhrer, L. Pishchulin, A. Brunton, C. Shu, and J. Lang. Estimation of human body shape and posture under clothing. *Computer Vision and Image Understanding*, 127:31–42, 2014.

[38] D. Xiang, H. Joo, and Y. Sheikh. Monocular total capture: Posing face, body, and hands in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[39] W. Xu, A. Chatterjee, M. Zollhoefer, H. Rhodin, D. Mehta, H.-P. Seidel, and C. Theobalt. Monoperfcap: Human performance capture from monocular video. *ACM Transactions on Graphics (TOG)*, 37(2):27, 2018.

[40] J. Yang, J.-S. Franco, F. Hétroy-Wheeler, and S. Wuhrer. Analyzing clothing layer deformation statistics of 3d human motions. In *Proceedings of the European Conference on Computer Vision*, pages 237–253, 2018.

[41] T. Yu, K. Guo, F. Xu, Y. Dong, Z. Su, J. Zhao, J. Li, Q. Dai, and Y. Liu. Bodyfusion: Real-time capture of human motion and surface geometry using a single depth camera. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 910–919, 2017.

[42] T. Yu, Z. Zheng, K. Guo, J. Zhao, Q. Dai, H. Li, G. Pons-Moll, and Y. Liu. Doublefusion: Real-time capture of human performances with inner body shapes from a single depth sensor. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7287–7296, 2018.

[43] C. Zhang, S. Pujades, M. J. Black, and G. Pons-Moll. Detailed, accurate, human shape estimation from clothed 3d scan sequences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4191–4200, 2017.