Towards 3D Human Shape Recovery Under Clothing

Xin Chen, Anqi Pang, Wei Yang, Jingyi Yu†

Abstract—We present a learning-based scheme for estimating clothing tightness as well as the human shape on clothed 3D human scans robustly and accurately. Our approach maps the clothed human Chen: geometry/appearance 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 tightness 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, cloth retargeting and animations.

Index Terms—Human shape recovery, 3D cloth segmentation, parametric human model.

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 [33] and DoubleFusion [24] produce high quality 3D scans using a single 3D sensor whereas more sophisticated dome systems [12] acquire dynamic models with textures. However, nearly all existing approaches conduct reconstruction without considering the effects of clothing, or more precisely the tightness 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 tightness from the acquired 3D human models. Fig.1 shows our estimated tightness, 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. Fig.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 tightness of clothing.

Fig. 1: Our method takes a clothed 3D scan (Left) as input, and predict the clothing tightness (Second) with the assistance of parametric human model and geometry images. With the predicted tightness, we can predict the underlying human body shape from the 3D scan as well as segment the clothes (Thrid). This illustrates how we can support a range of applications related to multi-layer avatar generation (Right), cloth segmentation, and try-on.

Different from previous approaches, we focus on simultaneously modeling the tightness of clothing and human body shape. We observe that humans can quickly identify clothing tightness (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 tightness 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 [16] that we call clothed-GI. We extend our parametric human model modified from SMPL [29], and subdivide key geometry features around

- *Xin Chen, Anqi Pang are affiliated with the School of Information Science and Technology, ShanghaiTech University, Shanghai Institute of Microsystem and Information Technology, Chinese Academy of Science, as well as University of Chinese Academy of Sciences. E-mail: {chenxin2, pangaq}@shanghaitech.edu.cn.*
- *Wei Yang is with University of Delaware, DGene and Google LLC. E-mail: wyangcs@udel.edu.*
- *Jingyi Yu is with the School of Information Science and Technology, ShanghaiTech University, Shanghai, China. E-mail: yujingyi@shanghaitech.edu.cn.*
- † *Corresponding author: Jingyi Yu.*

Fig. 2: The effectiveness of the parametric fitting model, such as SMPL [29], is weakened by clothes. From left to right, SMPL generates worse results (more unfit to underlying human body) for ultra-tight, fitted and loss clothes.

cloth boundary(e.g. neck, wrist, waist, ankle, etc) so that human models of different shapes and under different clothing can be effectively analyzed via clothed-GI. By employing skeleton and joint warping, our alignment scheme supports a large variety of poses with a 3-stage(silhouette/point-cloud/per-vertex) deformation method. For each pixel on the clothed-GI, we extract a feature vector including color/texture, position, normal and set out to predict its tightness measure. We use the vectors between each corresponding vertex pair of body and clothing geometry as the tightness measure and train a modified conditional generative adversarial network (GAN) for per-pixel tightness prediction. Finally, we use the tightness measure to predict human shape under clothing and help to segment clothing.

We collect a 3D dataset that consists of a large variety of clothing: T and long shirt, short/long/down coat, hooded jacket, pants, skirt/dress etc, and 3D human shapes. Comprehensive experiments show that, compared with the state-of-the-art, with only one static model as input, 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 cloth retargeting and animation.

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 (MVS) vs. depth fusion based approaches. The former employs correspondence matching and triangulation [15], [34], [45], 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 [19], [20], [52]. 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 [7], [14], [32], [56]. Newcombe *et al*. [32] compensate geometric changes due to motion captured from a single RGB-D sensor. Yu *et al*. [57] 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 [4] and SMPL [29], and require the subject wearing minimal or fitted clothing. The earlier work by [6] 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*. [51] 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 fitted clothing since, even for humans. More recent approaches attempt to align clothing on the human body model [17], [39], [60]. Our approach also aims to align a parametric model but we employ the geometry image analysis and exploit tightness prediction for more reliable shape prediction.

[35], [38], [50] learn articulated body poses of humans from their occluded body parts via sequential convolutional networks (ConvNets). [26] predicts body segments and landmarks from annotated human pose datasets, and conducts body pose estimation with clothing and 3D body fitting. Lassner *et al*. [25] present a generative model of full body in clothing, but their work focuses more on appearance generation than body shape estimation. Pavlakos *et al*. [37] 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 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, [55] 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*. [39] estimate a minimally clothed shape and uses retargeting to map the body shape in different sizes and poses. Lähner *et al*. [22] 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 tightness between the human body and the cloth layer. We then estimate both inner body shape and clothing segmentation from our new dataset of different subjects in different clothing styles.

The main difference to the previous approaches, such as [55], [60], is that we introduce the clothing tightness to model the relations between clothing and the underlying human body. There are also several works, such as [1] and [2], propose to use per-vertex displacements to compensate the gap between clothed and unclothed human model. However their purpose is to reconstruct a clothed human rather than to extract body shape. And their offsets/displacements didn't consider the influence of various cloth types and human poses. Instead, we take the tightness into consideration for both mesh alignment and shape recovery. Chen: Moreover, with the extract the tightness our approach can recover clothing and body shape from a single static mesh. Compared to single image based [37] and mesh sequence based [21] methods, our approach can leverage the comercial 3d sensors, such as Kinect and ToF cameras on mobile phones, for reliable body and clothing recovery.

3 Overview

To recovery human body and clothing from a single mesh, we first prepare a dataset with both clothed human mesh and underlying body shape. We collect raw clothed human mesh with a the multi-

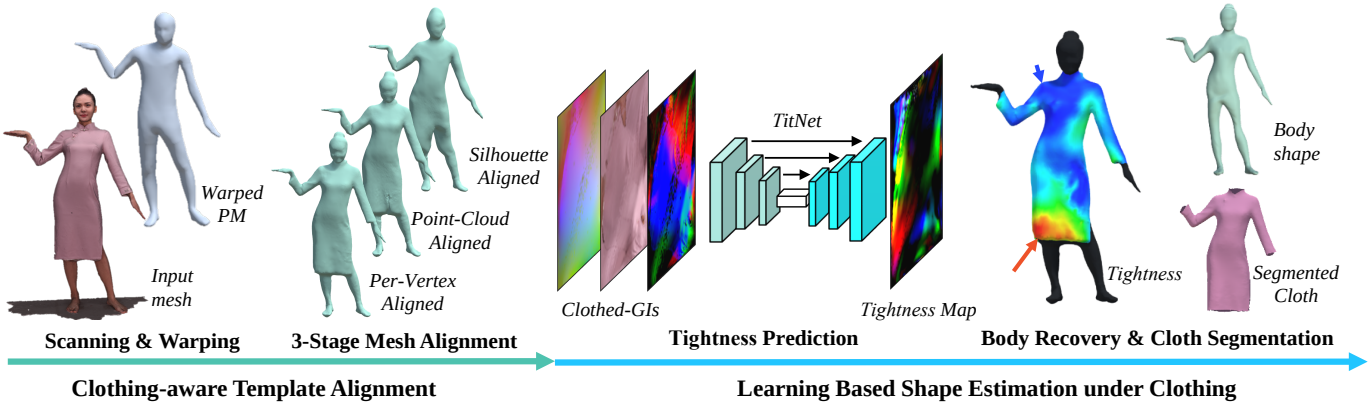


Fig. 3: The pipeline of our approach. The first step is to warp our CA-SMPL with scanned input mesh. Then we deform warped mesh using *3-Stag Mesh Alignment*. After we estimate the tightness map from mapped clothed-GI with *Tightness Prediction*. The final step is to recover body shape from predicted tightness on mesh, and to segment cloth.

camera dome system, as described in Sec.4.2. The ground-truth unclothed body shape are generated by artist (Sec.6.1).

The pipeline of our method is consist of three major steps, as shown in Figure 3:

Step 1: Clothed Human Body Alignment. We modify the SMPL [29] to generate a clothing adapted parametric model (CA-SMPL) for better deformation of the clothing part (Sec.4.1). We then deform the CA-SMPL to align with the input raw human mesh, and map the clothed deformed model to a 2D image, we call clothed geometry image (clothed-GI), for learning based tightness estimation (Sec.4.4).

Step 2: Tightness Prediction. We model the per-vertex looseness and tightness of the clothing as tightness (Sec.5.1), which is a vector from a vertex on clothing layer to the corresponding vertex on underlying human body. We develop a conditional GAN, namely TitNet, to predict the per-pixel tightness (Sec.5.2) in clothed-GI.

Step 3: Shape Recovery Under Clothing. In this step, we transform the predicted tightness map back to 3D mesh, and refine the 3D tightness with Gaussian filter.) We finally recover the underlying body shape and segment the clothing out with predicted tightness in Sec.5.3).

At the end, we present an exemplary application of our approach for avatar animation, which retargets the segmented clothing layer to a new body shape under a different pose.

4 Body Shape Alignment

Under canonical pose, tightness between various clothing and body shapes share a similar distribution, e.g. the clothes tend to be lose around human oxter and crotch, while tight on shoulder and chest. This implies we can predict the tightness of clothes with its underlying human body via a learning based approach. While learning directly on 3D shape using techniques such as Graph Conventional Network (GCN), instead we adopt the geometric image concept and map the 3D human mesh into a 2D image. Such mapping are more effective and can maintain the continuity of semantic body distribution as much as possible [16]. However, mapping a raw scan 3D model to a 2D image is generally infeasible due to noise and irregular geometries. Hence we propose to align the input model with a modified SMPL template and generate the geometric image from aligned template. In the next, we first introduce how we obtain the input clothed human meshes with a dome system. Then, we describe our 3-stage alignment method for aligning the template model with the input. Finally, we describe how we apply the geometry image, generated from aligned mesh via UV mapping, for tightness training and prediction.

4.1 Clothing Adapted Human Template Model

The most successful parametric human models, i.e. the SMPL and SMPL-X [36], focus on unclothed human body pose/shape modeling, which we demonstrate is not suitable for clothed human body as shown in Fig. 2. Hence We modify the SMPL model to better align with the clothing part. Specifically, we subdivide the mesh around the head, waist, feet and hands, places that are probably around the boundaries of clothing. We also simplify the mesh around ears, nose and fingers for efficiency purpose, as shown in Fig.4 (a). We rig the modified model with the skeleton defined by OpenPose [8], [43], i.e. 23 joints for the main body part and 21 joints for each hand (Fig.4 (b)).

Our clothing adapted SMPL (CA-SMPL) $\mathcal{M}_T$ contains $N_{\mathbf{M}} = 14985$ vertices, $N_{\mathbf{F}} = 29966$ facets and $N_{\mathbf{J}} = 65$ joints, as defined in Eq.(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 $\mathbf{F}$ denotes the facets, $\mathbf{J}$ is the joints and $\mathbf{M}$ is the vertices. $M(\hat{\mathbf{J}})$ means warp the CA-SMPL according to joints $\hat{\mathbf{J}}$. The result is vertices set $\hat{\mathbf{M}}$. The joints in the CA-SMPL 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 the rotation $\mathbf{\Theta}$ with axis-angle representation. $\mathbf{S}$ is the scaling of each joint along the bone direction, and $\mathbf{m}$ is the global translation.

We deform the CA-SMPL following the embedded deformation (ED) graph [46]:

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

where $N_{\mathbf{G}}$ is the number of nodes in ED graph. The warping function G_k for each node consists of rotation $\mathbf{R}_k \in \mathbf{SO}(3)$ and translate $\mathbf{t_k} \in \mathbb{R}^{\mathbf{3}}$, $k \in [0, N_G)$. G_k has the following form:

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

$\hat{\mathbf{g}}_k \in \mathbb{R}^3$ indicates the canonical position of node k. The vertex $\mathbf{v}_i, i \in [0, N_{\mathbf{M}})$ after deformation according to ED graph $\mathcal{G}$ is:

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

$\hat{\mathbf{v}}_i$ is the canonical position of vertex i, $\mathbf{w}^{\mathbf{G}}_{i,k}$ is the weight between vertex i and graph node k. We compute $\mathbf{w}^{\mathbf{G}}_{i,k}$ according to the Euclidean distance of n-nearest nodes as in [46].

4.2 Input 3D Mesh Acquisition

The input of our method are raw 3D human meshes, which we reconstruct using MVS approach [42] from multi-view human images captured by a dome system equipped with 80 cameras. With this system, we construct a new dataset consists of raw 3D human meshes in various poses and pre-calibrated camera parameters. We also estimate the 2D human joints for each image as in [8], [43] and obtain the 3D joints through triangulation as in [48].

4.3 3-Stage Alignment

To accurately align with the clothed human body, we propose a 3-stage alignment scheme, i.e. silhouette based-, point cloud based- and per-vertex deformation. Specifically, we first align our CA-SMPL with silhouette using a coarser ED-graph to handle error prone places due to holes or noise on raw mesh. Then we deform with a finer ED-graph according to the point cloud. Finally, we conduct a per-vertex deformation to generate the geometric details. The following sub-sections provide the details of each stage.

For prepossessing, we warp the CA-SMPL $\mathcal{M}_T$ with the triangulated 3D joints $\mathbf{J}_{mv}$ and compute vertex accordingly $\mathbf{M}_{warp} = M(\mathbf{J}_{mv})$.

Start with $\mathbf{M}_{warp}$, our 3-stage deformation scheme works as follows (see Fig.4).

Silhouette based non-rigid deformation. For each input 3D mesh with extracted joints, we set up a virtual system $\mathcal{C}$ in Eq.(6) with $N_{\mathcal{C}} = 30$ virtual cameras to view different parts of the mesh. For each of five parts capturing neck, ankles and wrists, we set two cameras orthogonal to each other. And for the front and back sides, we arrange five cameras (different angle of view) capturing the upper and lower body torso respectively.

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

where $\mathbf{c}_j$ denotes extrinsic parameters of a camera. $w^{\mathcal{C}}_j \in [0.5, 1]$ represents the weighting factor for two different camera positions, 0.5 for the camera capturing torso and 1 for the camera capturing limb. This weight factor can help the optimization to improve the alignment of clothing boundary.

We then render a high resolution mask of the body mesh at each virtual view. Following the idea of [53], we more focus on the multi-view silhouette and the balance of different view and define the following term:

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

where E_{mv} represents the data term for the multi-view silhouette generation, E_{reg} a regularization term as defined in [44]. For the data term,

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

where $\mathbf{v}^{\text{S}}_j$ 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,

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

where $\mathcal{N}_k \in \mathcal{G}$ is the 1-ring neighbourhood of graph node k, $w^{\mathbf{N}}_{k,n}$ the weight between the nodes k, n. We denote this vertices matrix of optimization result as $\mathbf{M}_S$.

Point cloud based non-rigid deformation. We exploit a point cloud based non-rigid deformation with ED graph to deform our silhouette deformation result $\mathbf{M}_S$ to the input 3D mesh. We re-sample the ED graph of $\mathcal{M}_T$ with more nodes, but remain $\mathcal{G}$ for clarity. The energy function is defined as

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

where data term $E^{\text{D}}_{\text{data}}(\mathcal{G})$ is for the deformation and regularization term $E^{\text{D}}_{\text{reg}}(\mathcal{G})$ the same as in Eq.(9). The data term in Eq.(10) is as

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

where $\lambda^{\text{D}}_{\text{point}}$ is the weight for point-to-point distance, $\lambda^{\text{D}}_{\text{plane}}$ 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. We denote this vertices matrix of optimization result as $\mathbf{M}_D$.

Per-vertex deformation. We finally refine the deformation from ED graph-based non-rigid result $\mathbf{M}_{\mathbf{D}}$ to the input 3D mesh via per-vertex optimization. This part of algorithm can improve the local details like clothing wrinkle and boundary. as in Eq.(12), with its data term in Eq.(13).

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

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

We denote this vertices matrix of optimization result as $\mathbf{M}_V$. Finally, we obtain a clothed deformed model, namely CDM, deformed from CA-SMPL. We can calculate the tightness if given its the UnClothed Body Model, namely UCBM, same deformed from CA-SMPL. We show this part later in 6.1. Following part is to establish the mapping from our template to a geometry image (UV map).

4.4 Geometry Image Representation of Clothed 3D Human

Though it's possible to predict the tightness directly on mesh data using techniques like GCN [28], it consumes much more computational resource and is generally harder to train. Instead, we map the clothed 3D body mesh to a 2D image, which has been proved to be effective in previous works [3], [23], [54].

There are many methods to generate a 2D mapping from a 3D mesh ways. We choose two representative methods for comparison. One is the mapping approach of the geometry image [16] with

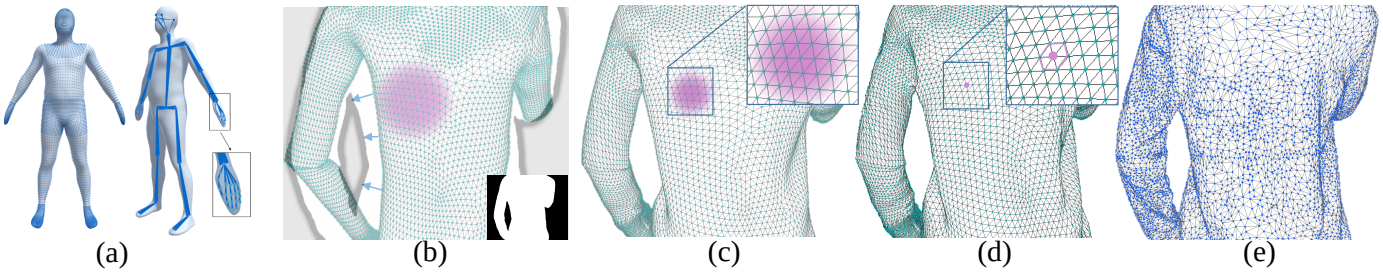


Fig. 4: Each stage of our alignment approach. (a) The template model for alignment. (b) The first stage, silhouette based deformation. (c) The second stage, point cloud based deformation. (d) The third stage, per-vertex deformation. (e) The referenced mesh (target mesh).

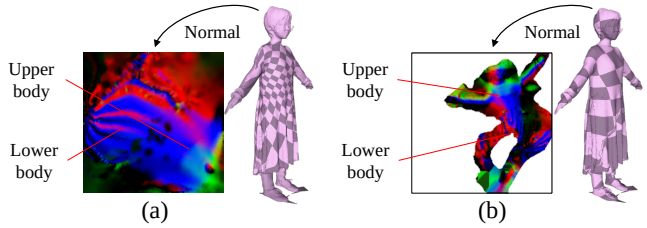


Fig. 5: The comparison of two feature map with different mapping methods. (a) The normal map using geometry image [16]. (b) The normal map using OptCuts algorithm [27].

Fig. 6: Illustration of tightness $\mathbf{T}_i$ of vertex i on mesh.

gapless filling but relative large distortions, denoted as $M_{GI}(\cdot)$. The other method is OptCuts [27], denoted as $M_{Opt}(\cdot)$, which automatically seek the best seam for cutting and generate an image with lower distortion but contains gap area. We map the positions, normals and RGB colors of each vertex into its 2D map and conduct a linear interpolation to fill the image, we call the generated clothed geometry image clothed-GI. We adopt both mapping methods in our experiment in Sec.6.4, and finally select the geometry image [16] for better performance.

5 Tightness Prediction

Previous human reconstruction methods, including methods based on scanned depth map(s) [12], [14], [32], [33], and silhouette(s) [5], [9], [10], [13], [30], [53], represent the human body as a single layer. Recently, [31], [40], [58], [59], [60] proposed the idea of multi-layer body shape recovery. We extend this idea and define the tightness, which describes the relation between the underlying body shape (body layer) and the garment (cloth layer). Subsequently, we propose a conditional GAN to predict the tightness.

5.1 Tightness measure

As mentioned in related work, tightness has same mathematical representation but different meaning with the existing displacement/offset in [1] and [2]. Tightness indicates the displacement between real body layer and cloth layer rather than the displacement between reconstructed/aligned mesh and template mesh.

We define the tightness on mesh with the topology of our CA-SMPL. As we mentioned, the tightness are the gap between two layers (body/cloth), so we take our UCBM (aligned with a unclothed body mesh) and CDM (aligned with a clothed mesh) as example. For a vertex $\mathbf{v}_i$ in CDM, we define the vertex tightness as following:

$$\mathbf{T}_i = \mathbf{v}_i - \mathbf{v}_i^c \tag{14}$$

where $\mathbf{v}_i^c$ is the corresponding vertex in the UCBM, The direction of $\mathbf{T}_i$is from the vertex in a CDM to the corresponding vertex in its UCBM. Its magnitude is the euclidean distance between the two corresponding vertices. Hence, we can define the tightness matrix of CA-SMPL as $\mathcal{T}$:

$$\mathcal{T} = \left\{ \mathbf{T} \in \mathbb{R}^{N_\mathbf{M} \times 3} \right\}, \tag{15}$$

where $\mathbf{M}$ is the vertex matrix of CDM with $N_\mathbf{M}$ vertices. This theoretical calculation of tightness has a critical problem of correspondence matching. It is challenging to align the CA-SMPL to the exact corresponding position between body layer and cloth layer.

In practice, we use the CDM (cloth layer) to enable calculating the ground truth tightness between its and the carved shape mesh (not a UCBM from our CA-SMPL) directly. Hence, we can calculate the approximate tightness $\mathcal{T}_i$ at a vertex v_i of the CDM as following:

$$\mathbf{F_i} = \frac{\sum_{\mathbf{v_r^c} \in \mathcal{N}_1^c} K_G(\mathbf{v_i} - \mathbf{v_r^c}) + \sum_{\mathbf{v_d^c} \in \mathcal{N}_2^c} K_G(\mathbf{v_i} - \mathbf{v_s^c})}{||\mathcal{N}_1^c|| + ||\mathcal{N}_2^c||} \tag{16}$$

where v_r^c and v_d^c refers to corresponding vertices in the body shape mesh. $\mathcal{N}_1^c$ is the set of vertices ray-traced along the normal direction of v_i within a double-cone with an aperture of 30 degrees, as shown in Fig.6, and $\mathcal{N}_2^c$ the set of 20 closest vertices of v_i in the body shape mesh. $K_G(\cdot)$ is a weighting function of the Gaussian kernel based on the angle between two vertex normals for de-noising.

After tracing from the vertices of CDM to the body shape mesh, we also apply same algorithm from UCBM to the clothed MVS mesh. We average the two results to extract the ground truth of 3D tightness on mesh. With the mapping function in Sec.4.4, we final generate the tightness map for following learning part.

5.2 TitNet architecture

For our problem is natural to explore network architecture designed to generate multi channel images. We treat the input as 9 channels input of positions, normals, and colors, the output will be 9 channel predicted tightness, top clothing mask and under clothing mask. And In our experiment, we adapt the architecture of Pix2Pix [18], which is based on cycle GAN [62].

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 [18], the objective function is defined as

$$\begin{aligned}\mathcal{L}_{GAN}(G, D) =& \mathbb{E}_{x,y}[\log D(x, y)]+ \\ & \mathbb{E}_{x,z}[\log(1 - D(x, G(x, z))]\end{aligned} \tag{17}$$

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

$$\mathcal{L}_{L1}(G) = \mathbb{E}_{x,y,z}\left[\|y - G(x, z)\|_1\right] \tag{18}$$

$$G^* = \arg\min_G \max_D \mathcal{L}_{GAN}(G, D) + \lambda\mathcal{L}_{L1}(G) \tag{19}$$

we adapt our generator and discriminator architectures from those in Pix2Pix [18]. Considering our input and output differ in appearance but have the same semantic structure meanings, the generator structure in the input is aligned with structure in the output. We treat different source of input as different input channels, so are the output. All the input and output are using the same 2D mapping method (GI [16] or OptCus [27]), so they keep aligned and have the same image structure. In our TitNet, the generator is U-Net [41] encoder-decoder structure with skip connections between mirrored layers in the encoder and decoder stacks. This kind of U-Net with connections can share information between the input and output. In our discriminator, we tried PatchGAN [18] discriminator and normal full image GAN discriminator. We find that the high-frequency structure is not important in our task, so we adapt normal full image GAN discriminator with pyramid structure. We run this discriminator conventionally across the real and fake image.

Benefiting from our tightness predictor, we can extract the hidden information between different clothing appearance and the tightness. The input positions and normals also help our predictor to consider the effect of current human pose. In the Sec.6.4, we introduce the training and testing in detail, and evaluate this TitNet with a proposed dataset. These experiment demonstrate the reliability of our method. In the following section, we introduce how to recovery body shape and segment clothing With the help of predicted tightness map and top/down clothing mask.

5.3 Shape recovery under clothing

Shape recovery As we defined in Sec.5.1, each tightness vector on template model starts from cloth layer and ends at body layer. With the help of tightness we can directly recover body shape. From our TitNet, we estimate the tightness map of each mapped clothed mesh. We use the inverse function of our mapping $M_{GI}^{-1}(\cdot)$ or $M_{Opt}^{-1}(\cdot)$, explained in Sec.4.4, to generate tightness $\hat{\mathcal{T}}$ on mesh:

$$\hat{\mathcal{T}} = \left\{\hat{\mathbf{T}} \in \mathbb{R}^{N_{\mathbf{M}} \times 3}\right\}. \tag{20}$$

With same Gaussian kernel in Eq.16, we use a simple least-squares optimization to generate final body shape:

$$\begin{aligned}E_{\text{body}}(\mathbf{M}) =& \lambda_{fit}(\mathbf{M} + \mathbf{T} - \mathbf{M}_{\mathbf{V}})+ \\ & \lambda_{smooth}(\mathbf{M} - K_G(\mathbf{M}))+ \\ & \lambda_{reg}(\mathbf{M}_{warp} - \mathbf{M}),\end{aligned} \tag{21}$$

where the first term uses tightness to restrict body shape. The second term is to smooth body surface. The third term is the regression term with original warped body shape. Vertex matrix $\mathbf{M}_{warp}$ and $\mathbf{M}_{\mathbf{V}}$ are from the warped mesh and aligned mesh, as mentioned in Sec.4.3. The parameters λ_{fit} is 1, while λ_{smooth} and λ_{reg} are 0.1 . Finally, we recover body shape mesh with vertex matrix $\hat{\mathbf{M}}_b$.

Fig. 7: Sample data from our dataset: three real human subjects with scanned body shape meshes, the segmented clothes and one synthetic model (right most) for pre-training.

Clothing segmentation As mentioned in Sec.5.2, our TitNet generate both tightness map and clothing mask for top/down clothing. Current clothing mask are not good enough to segment accurately, but still more effect than manual defined garment prior used in [40]. We follow the optimization in [40] with our data-driven garment prior. We use the following integer linear optimization to solve variable $v_i \in \mathbf{v}$ for every vertex in our CDM:

$$E_{\text{cloth}}(\mathbf{V}) = \sum_{i \in \mathcal{T}} \varphi_i\left(v_i\right) + \sum_{(i,j) \in \mathcal{T}} \psi_{ij}\left(v_i, v_j\right), \tag{22}$$

where the first part, the unary term includes the data likelihood and the data-driven garment prior. In this part, we fit a Gaussian mixture model (GMM) to the appearance of each garment with HSV space rather than RGB. The second term is a binary term for smoothness. It encourages the local nodes have the same label. As it is not the focus of our work, we can only briefly introduce this part. The details of each term and the design of fitted GMM is in [40].

6 Experimental Results

In this section, we conduct both qualitative and quantitative experiments to evaluate our method. First of all, we present a 3D human and clothing dataset, called the Clothing Tightness Dataset (CTD). All our experiment are conducted on our CTD and the Bodies Under Flowing Fashion (BUFF) dataset proposed in [60]. We first evaluate our clothing alignment module(4.3) and validate the effectiveness of different stages. Then, we compare the performance of our TitNet(4.3) with the proposed baseline. And we qualitatively analyze the recovered body shape and the segmented clothing on CTD. Finally, we exhibit some interesting applications of our approach including cloth fitting and retargeting.

6.1 Dataset

Clothing Tightness Dataset (CTD): Our CTD contains 880 clothed meshes with human body geometry and segmented individual pieces of clothing. Among them, 228 meshes are statically captured and 652 are from dynamic 3D human sequences (13 sequences in total). We have captured 18 subjects, 9 males and 9 females, 10 of them are with the canonical "A" or "T" poses and 8 subjects are under free movements. Our CTD contains

Method	Mean↓	RMS↓
Non-rigid [47]	0.348%	0.662%
Silhouette-based [53]	0.494%	0.779%
2-Stage + Point cloud	0.286%	0.585%
3-Stage (Ours)	**0.263%**	**0.521%**

TABLE 1: Comparison of alignment methods for clothed human mesh. *Non-rigid* [47] is the baseline to directly align a 3D mesh with the input point cloud. *Silhouette-based* [53] is the baseline to align a 3D mesh from silhouette only, which is also our first stage. ↓ means the smaller is better. *Mean* and *Root-Mean-Square (RMS)* are the average of Hausdorff distance [11], normalized with the bounding box diagonal of all clothed meshes, which is 2.9 in our case.

228 different garments for each static caption, including T/long shirt, short/long/down coat, hooded jacket, pants, and skirt/dress, ranging from ultra-tight to puffy. For dynamic sequence, we capture 400 500 frames under 30fps and evenly sample 40 50 frames for our dataset. As mentioned in Sec.4.2, we use multi-view stereo to reconstruct each 3D mesh in our dataset, with about 50,000 vertices, 100,000 facets and with 4K texture. We show an overview of our dataset in Fig. 7. To obtain the ground truth underlying body shape, 5 artists manually segment each piece of clothing and carve the body shape out from the raw mesh. We then generate the ground truth tightness trough the algorithm described in Sec.5.1. We use 80% of the dataset as training set, and the remain data as testing set. We also generate 800 hundred clothed human meshes with synthetic avatars using Adobe Fuse CC for the pre-training of our network.

Bodies Under Flowing Fashion (BUFF) dataset [60] is a relevant dataset for body shape estimation, which contains 3 males and 3 females models wearing 2 types of clothing, t-shirt/long pants and a soccer outfit. It focuses on dynamic sequence and only provides the vertex color rather than high-quality texture. BUFF also contains body shapes under general T pose without clothing as their ground truth. Since BUFF is not a dataset for learning method, we use a frame from their mesh sequence as the input, and predict tightness maps with trained model from our dataset.

6.2 Implementation details

We conduct our experiments on one GPU NVIDIA GTX 1080T GPU. For 3-stage alignment, we perform these optimizations, in which $N_{\mathbf{G}^S} = 1407$, $\lambda^S_{reg} = 10$; $N_{\mathbf{G}^D} = 2103$, $\lambda^V_{reg} = 7$ $\lambda^D_{point} = 0.5$, $\lambda^D_{plane} = 1.5$; $\lambda^V_{reg} = 1$, $\lambda^V_{point} = 1$ and $\lambda^V_{plane} = 1.5$. For tightness prediction, we set the resolution of clothed-GI as 224×224.

For the computation time in our algorithm, we implement a efficient Gauss-Newton solver based on CUDA for our optimizations. The 3-stage deformation costs 5-8 seconds, about 1.5/2/3 seconds in turn. We train the tightness predictor on CTD about 5 hours and the prediction need about 1 seconds for one tightness map.

6.3 3-Stage alignment results

We show our qualitative and quantitative results for a variety of garments and pose. In Fig.8 we present the result details of each stage on dress. The results show that the first stage (silhouette based) can only generally align clothing from the view of virtual camera, but the limbs have already fitted. Although the clothing

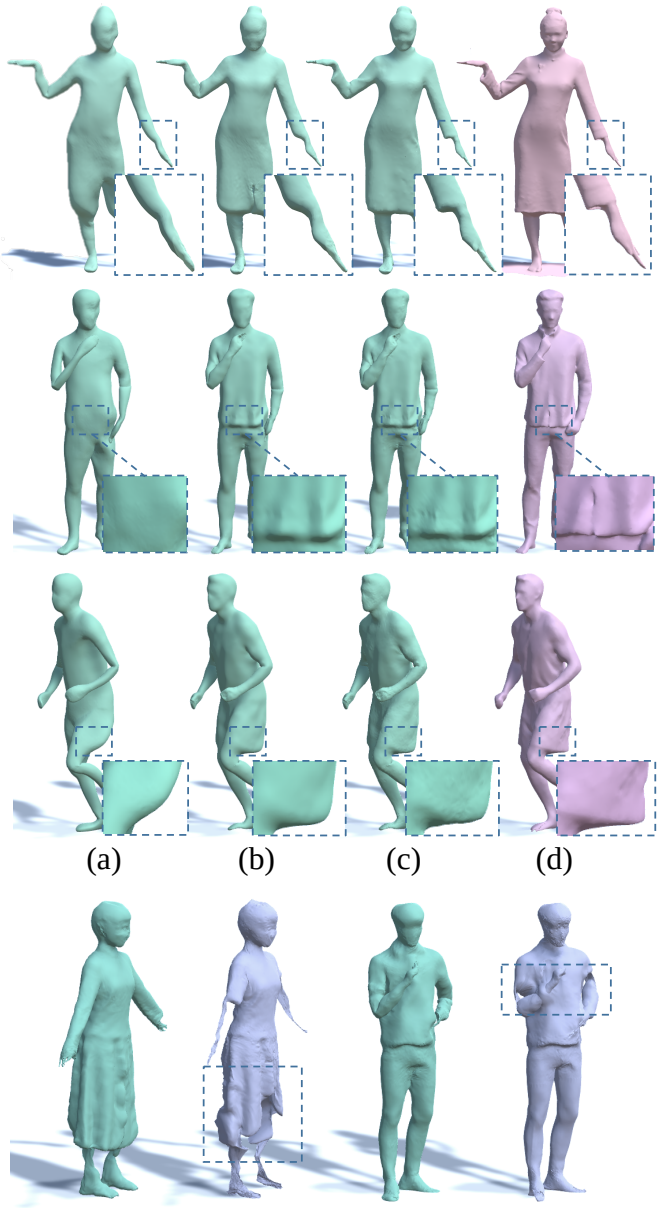


Fig. 8: The comparison of 3 stage alignment methods. (a) The results of silhouette-guided alignment. (b) The results of point cloud guided alignment. (c) The results of per-vertex alignment. (d) The input meshes from MVS (Target meshes).

details in the second stage (point cloud based) are satisfactory, the three stage (per-vertex) can even float the crack on skirt and match the clothing boundary around wrist. However, as shown in the bottom of Fig. 8, without the good initial state provided by silhouette based deformation, the optimization can not converge to a good result. For more different clothing type, we show the result in Fig.8, our method can adapt to various clothing type. we use Metro [11] as the metric of our quantitative evaluation. It is a popular measurement based on Hausdorff distance for comparing the difference of two meshes. In our case, we normalize these Hausdorff distances with the bounding box diagonal of all clothed meshes, which is 2.9. Hence, as shown in Tab.1, compared to directly non-rigid alignment [47] method, we improve 0.085% on the mean error of alignment.

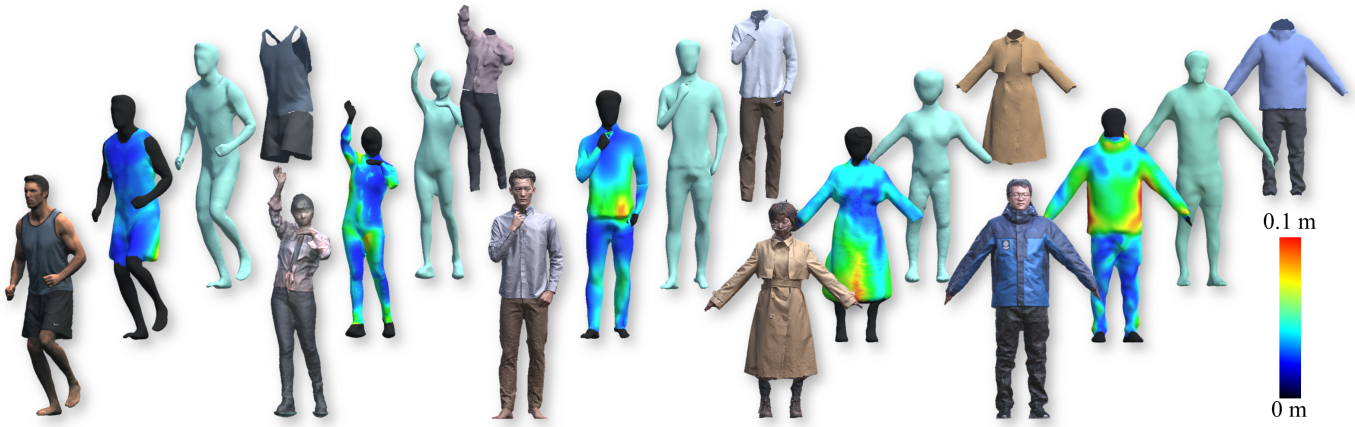


Fig. 9: The gallery of our results. From down to top, the input meshes, predicted tightness, recovered body shapes and segmented clothes.

Method	SSIM↑	L1/L2↓
Baseline our L2	62.27%	0.141%
OptCuts with mask L1	64.82%	0.1892%
OptCuts no mask L1	43.91%	0.4929%
Ours	**67.24%**	**0.2223%**

TABLE 2: SSIM [49] and L1 method for GAN evaluation. ↑ means the larger is better, and ↓ means the smaller is better.

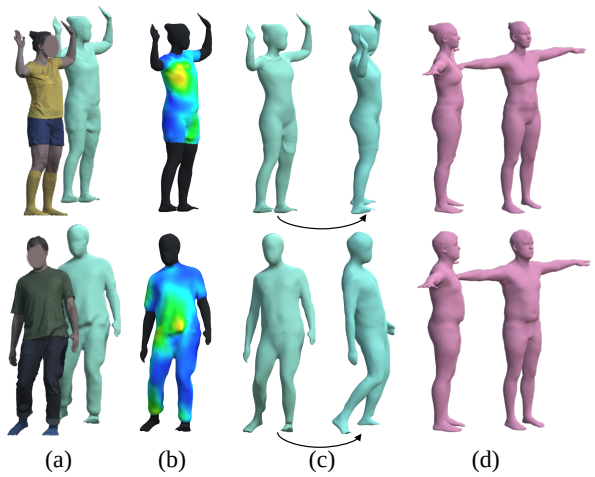


Fig. 10: Comparison with our recovered body shapes and the detailed method [60] in BUFF Dataset. (a) Input Meshes and aligned meshes. (b) Predicted tightness on meshes. (c) The recovered body shape of our results. (d) The results of the detailed method [60].

6.4 TitNet evaluation

For the evaluation of our TitNet, we use L1 norm and the structural similarity (SSIM) [49], a method for predicting the perceived quality of images. We set the window size as 11 to avoid the unreasonable effectiveness [61] of this metric. As the clothing mask can influence the results of tightness prediction, We also compare this part in our experiments. As shown in Tab.2, compared to the baseline using L2 loss, our final method using geometry image can improve 5.02% performance.

Method	Mean↓	RMS↓
Warped Template	1.316%	1.841%
SMPL Fitting	0.538%	0.764%
Our	**0.191%**	**0.451%**

TABLE 3: Comparison of recovered body shape in CTD. *Baseline* is the warped mesh from template model.

6.5 Body shape evaluation

We have also conducted experiments on two body meshes in the BUFF dataset [60], as shown in Fig.10. Without re-training our TitNet on BUFF, the results still show that our new method enables accurate and robust estimation of tightness, body shape for one static mesh of a human body in a variety of clothing styles. Considering the input of the detailed methods [60] are all mesh sequences, their results are quite close to the ground truth in BUFF dataset. However, our method still achieves good results, because our method only use one static mesh as input.

We evaluate our recovered body shapes with the ground truth body shapes in CTD. We still use Metro [11] for the metric of this evaluation. The metric has explained in Sec.6.3. In Tab.3, we show the recovered error between our body shapes and these ground truth body shape. For a fair compassion, we compare to a 3D SMPL fitting method rather than the SMPL fitting from single image, as our input is a 3D mesh. This 3D SMPL fitting method use point cloud as target to regress parameters of SMPL, which is generally used as part of human body optimization, like DoubleFusion [58]. With the help of tightness, our method can improve 0.347% performance than these fitting methods.

6.6 Cloth retargeting and avatar

Within our method, we have aligned the clothed mesh with a CDM (in Sec.4.3) and recovered the body shape with a UCBM (in Sec.5.3), and these two models have same mesh topology and riggied skeleton. The displacements between this CDM and its UCBM are our optimized tightness. We can still use this method to build the displacement between two UCBM from different person. Hence, we can transfer clothing between different UCBM with cloth-body-body displacements, as shown in Fig.12. We even fit our CA-SMPL with SMPL [29] under different shape parameters, and transfer clothing to shape-variant SMPL model directly.

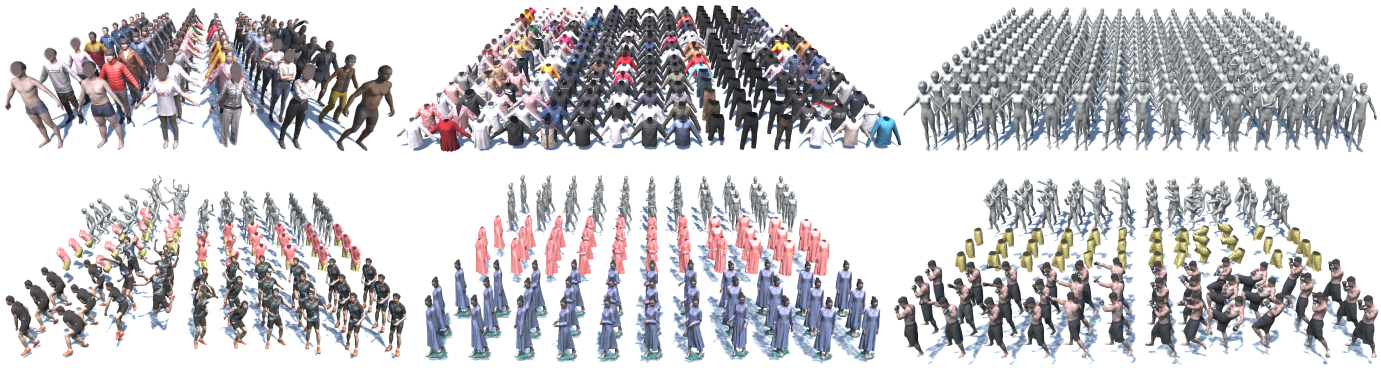


Fig. 11: The gallery of our Clothing Tightness Dataset (CTD). First row from left to right: 1) Sampling of various clothed human including synthetic models (rightmost). 2) Various segmented clothes with 'A' pose only. 3) Caver-ed body shapes with 'A' pose only. Second row are three typical dynamic sequences in our dataset including clothed human, segmented clothes, and carved body shapes.

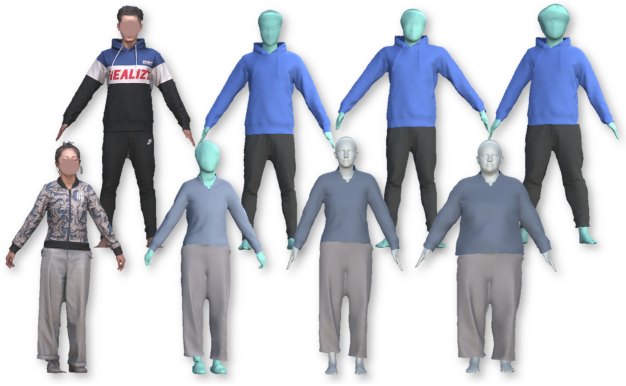


Fig. 12: First column are these MVS meshes. Second column are estimated body shape and segmented clothing. Third and fourth column are fitted clothing with slim body and fat body. Our estimated body (green), SMPL [29] (gray).

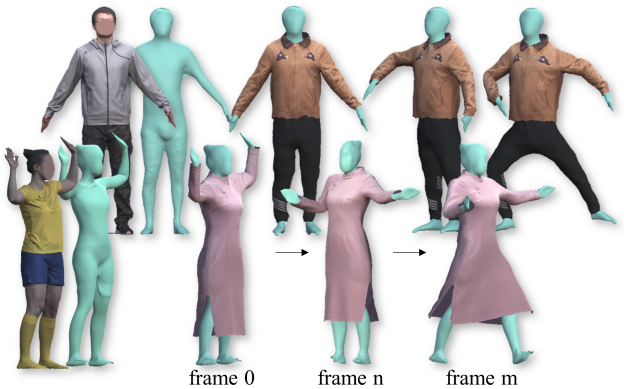


Fig. 13: Cloth retargeting results (second column) from source subject (left) and animated results from second column to right.

Benefiting from our CA-SMPL and tightness on each vertex, we can infer the body shape and the segmented clothing. Naturally, we expect to use these two parts to build a two-layer avatar for animation. Combining with different styles of clothes that have been segmented before, as mentioned in Sec.6.6, we can even use skeleton parameters to drive these avatar with different poses, as shown in Fig.13. Notes that we do not use any physical simulation for clothing, the dress floats with only skeleton-driven deformation.

7 Discussion and Conclusion

We present a learning-based scheme for robustly and accurately estimating the clothing tightness as well as human geometry on a single clothed 3D human raw mesh. The key contribution of our approach is the usage of geometry image for tightness prediction, and the alignment of human geometry enable the geometry image correspondence from various types of clothing. Moreover, we collect a large 3D Clothing Tightness Dataset (CTD) for the clothed human reconstruction tasks. We propose and train a modified conditional GAN network to automatically predict the clothing tightness and subsequently the underlying human shape. Experiments demonstrate the reliability and accuracy of our method. We also exhibit two interesting virtual try-on applications, i.e. cloth retargeting and clothed avatar. We believe our scheme will benefit various AR/VR research and applications, such as virtual try-on and avatar.

Limitation and Future Work Though our approach is effective for body shape estimation from single clothed mesh, there are several limitations: 1) it can not handle input meshes with very complex poses, such as crossing legs/arms or curling up, or very low-quality scans. 2) Our clothed avatar can not simulate the dynamic movement of clothing as there is no physically-based simulation for our segmented clothing. 3) Our approach is based on geometry images and hence can only handle genus 0 human geometry. In reality, human model can have very complex topology and a much more sophisticated geometry image generation approach is required. Alignment schemes that can handle topologically complex human models are our immediate future work. Our approach relies on raw 3D human scans which are usually difficult to obtain and the quality can not be guaranteed. Hence we plan to explore the possibility of directly taking a single or sparse set of 2D images as input to recover the 3D clothing

and human shape. Also through augmented training under various lighting conditions using the light stage, it is possible to capture the reflection property of the clothing and for better AR/VR or try-on experience.

ACKNOWLEDGMENTS

The authors would like to thank WenGuang Ma, YeCheng Qiu, MingGuang Chen for help with data acquisition; Hongbo Wang, Gao Ya, Shenze Ye, Teng Su for help with data annotation. This work is supported by the National Key Research and Development Program (2018YFB2100500), the programs of NSFC (61976138 and 61977047), STCSM (2015F0203-000-06), and SHMEC (2019-01-07-00-01-E00003).

REFERENCES

[1] T. Alldieck, M. Magnor, B. L. Bhatnagar, C. Theobalt, and G. Pons-Moll. Learning to reconstruct people in clothing from a single rgb camera. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1175–1186, 2019.

[2] T. Alldieck, M. Magnor, W. Xu, C. Theobalt, and G. Pons-Moll. Video based reconstruction of 3d people models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8387–8397, 2018.

[3] R. Alp Güler, N. Neverova, and I. Kokkinos. Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7297–7306, 2018.

[4] 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.

[5] S. Baker, T. Kanade, et al. Shape-from-silhouette across time part ii: Applications to human modeling and markerless motion tracking. *International Journal of Computer Vision*, 63(3):225–245, 2005.

[6] 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.

[7] 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.

[8] 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.

[9] G. K. Cheung, S. Baker, and T. Kanade. Visual hull alignment and refinement across time: A 3d reconstruction algorithm combining shape-from-silhouette with stereo. In *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, volume 2, pages II–375. IEEE, 2003.

[10] K. Cheung, S. Baker, and T. Kanade. Shape-from-silhouette of articulated objects and its use for human body kinematics estimation and motion capture. In *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, volume 1, pages I–I. IEEE, 2003.

[11] P. Cignoni, C. Rocchini, and R. Scopigno. Metro: measuring error on simplified surfaces. In *Computer graphics forum*, volume 17, pages 167–174. Wiley Online Library, 1998.

[12] 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.

[13] S. Corazza, L. Muendermann, A. Chaudhari, T. Demattio, C. Cobelli, and T. P. Andriacchi. A markerless motion capture system to study musculoskeletal biomechanics: visual hull and simulated annealing approach. *Annals of biomedical engineering*, 34(6):1019–1029, 2006.

[14] 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.

[15] 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.

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

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7122–7131, 2018.

[22] 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.

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

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] O. Litany, A. Bronstein, M. Bronstein, and A. Makadia. Deformable shape completion with graph convolutional autoencoders. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[29] 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.

[30] M. Mikhnevich and P. Hebert. Shape from silhouette under varying lighting and multi-viewpoints. In *2011 Canadian Conference on Computer and Robot Vision*, pages 285–292. IEEE, 2011.

[31] A. Neophytou and A. Hilton. A layered model of human body and garment deformation. In *2014 2nd International Conference on 3D Vision*, volume 1, pages 171–178. IEEE, 2014.

[32] 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.

[33] 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.

[34] 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.

[35] 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.

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

[37] 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.

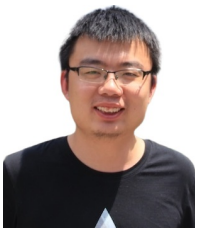
[38] L. Pishchulin, E. Insafutdinov, S. Tang, B. Andres, M. Andriluka, P. V. Gehler, and B. Schiele. Deepcut: Joint subset partition and labeling for multi person pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4929–4937, 2016.

[39] 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.

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

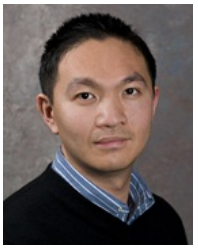
36(4):73, 2017.
[41] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.
[42] J. L. Schönberger, E. Zheng, J.-M. Frahm, and M. Pollefeys. Pixelwise view selection for unstructured multi-view stereo. In *Proceedings of the European Conference on Computer Vision*, pages 501–518. Springer, 2016.
[43] 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.
[44] 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.
[45] 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.
[46] R. W. Sumner, J. Schmid, and M. Pauly. Embedded deformation for shape manipulation. *ACM Transactions on Graphics (TOG)*, 26(3):80, 2007.
[47] J. Tong, J. Zhou, L. Liu, Z. Pan, and H. Yan. Scanning 3d full human bodies using kinects. *IEEE transactions on visualization and computer graphics*, 18(4):643–650, 2012.
[48] 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.
[49] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.
[50] 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.
[51] 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.
[52] 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.
[53] 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.
[54] Y. Xu, S.-C. Zhu, and T. Tung. Denserac: Joint 3d pose and shape estimation by dense render-and-compare. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7760–7770, 2019.
[55] 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.
[56] 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.
[57] 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.
[58] 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.
[59] T. Yu, Z. Zheng, Y. Zhong, J. Zhao, Q. Dai, G. Pons-Moll, and Y. Liu. Simulcap: Single-view human performance capture with cloth simulation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5499–5509. IEEE, 2019.
[60] 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.
[61] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018.
[62] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2017.

Xin Chen is a Ph.D. student at the School of Information Science and Technology (SIST), ShanghaiTech University. He obtained his B.S. from the School of Science at Qingdao University of Technology . His research interests include human performance caption, image-based modeling and deep learning. His homepage is https://chenxin.tech.



Pang Anqi is a Postgraduate Student at the School of Information Science and Technology (SIST), ShanghaiTech University. His research interests include human pose estimation, multi-view human joint estimation and deep learning.

Wei Yang is an engineer at the Advanced Technology and Projects division, Google LLC. He received the BEng and MS degrees from the Huazhong University of Science and Technology and Harbin Institute of Technology respectively, and the PhD degree from the University of Delaware (UDel) in Dec. 2017. He previously worked as a scientist the DGene. Co and joined ATAP at Google in Jan. 2020. His research interests include computer vision and computer graphics, with special focus in computational photography and 3D reconstruction.



Jingyi Yu is the executive dean in the School of Information Science and Technology at ShanghaiTech University. He received B.S. from Caltech in 2000 and Ph.D. from MIT in 2005. He is also affiliated with the University of Delaware. His research interests span a range of topics in computer vision and computer graphics, especially on computational photography and non-conventional optics and camera designs. He is a recipient of the NSF CAREER Award, the AFOSR YIP Award, and the Outstanding Junior Faculty Award at the University of Delaware. He has served as general chair, program chair, and area chair of many international conferences such as CVPR, ICCV, ECCV, ICCP and NIPS. He is currently an Associate Editor of IEEE TPAMI, IEEE TIP and Elsevier CVIU, and will be program chair of ICPR 2020 and IEEE CVPR 2021.