

TightCap: 3D Human Shape Capture with Clothing Tightness Field

XIN CHEN and ANQI PANG, ShanghaiTech University, University of Chinese Academy of Sciences, China
 WEI YANG, School of Computer Science & Technology, Huazhong University of Science & Technology, China
 PEIHAO WANG, LAN XU, and JINGYI YU, ShanghaiTech University, China

In this article, we present TightCap, a data-driven scheme to capture both the human shape and dressed garments accurately with only a single three-dimensional (3D) human scan, which enables numerous applications such as virtual try-on, biometrics, and body evaluation. To break the severe variations of the human poses and garments, we propose to model the clothing tightness field—the displacements from the garments to the human shape implicitly in the global UV texturing domain. To this end, we utilize an enhanced statistical human template and an effective multi-stage alignment scheme to map the 3D scan into a hybrid 2D geometry image. Based on this 2D representation, we propose a novel framework to predict clothing tightness field via a novel tightness formulation, as well as an effective optimization scheme to further reconstruct multi-layer human shape and garments under various clothing categories and human postures. We further propose a new clothing tightness dataset of human scans with a large variety of clothing styles, poses, and corresponding ground-truth human shapes to stimulate further research. Extensive experiments demonstrate the effectiveness of our TightCap to achieve the high-quality human shape and dressed garments reconstruction, as well as the further applications for clothing segmentation, retargeting, and animation.

CCS Concepts: • **Computing methodologies** → **Shape inference; Reconstruction; Motion capture; Mesh models;**

Additional Key Words and Phrases: Clothing, human shape capture, try-on, parametric human model

ACM Reference format:

Xin Chen, Anqi Pang, Wei Yang, Peihao Wang, Lan Xu, and Jingyi Yu. 2021. TightCap: 3D Human Shape Capture with Clothing Tightness Field. *ACM Trans. Graph.* 41, 1, Article 9 (November 2021), 17 pages. <https://doi.org/10.1145/3478518>

X. Chen and A. Pang are contributed equally.

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).

Authors' addresses: X. Chen, ShanghaiTech University, Shanghai Institute of Microsystem and Information Technology, Chinese Academy of Sciences, University of Chinese Academy of Sciences, China; email: chenxin2@shanghaitech.edu.cn; A. Pang, ShanghaiTech University, University of Chinese Academy of Sciences, China; email: pangaq@shanghaitech.edu.cn; W. Yang, School of Computer Science & Technology, Huazhong University of Science & Technology, DGene Inc., China; email: wyangcs@udel.edu; P. Wang, ShanghaiTech University, China; email: peihaowang@utexas.edu; L. Xu, ShanghaiTech University, China; email: xulan1@shanghaitech.edu.cn; J. Yu, ShanghaiTech University, China, yujingyi@shanghaitech.edu.cn.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2021 Association for Computing Machinery.

0730-0301/2021/11-ART9 \$15.00

<https://doi.org/10.1145/3478518>

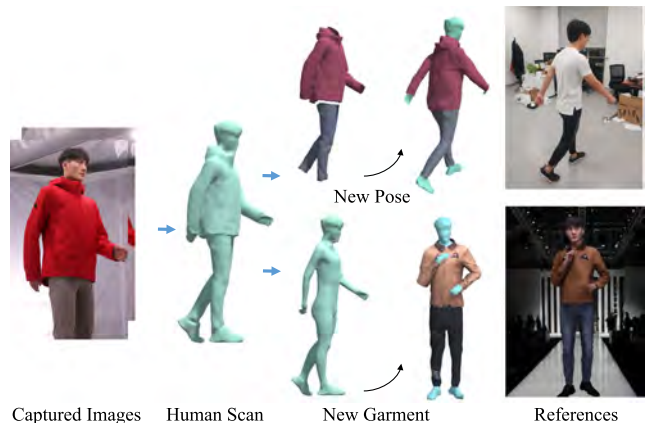


Fig. 1. Our method first reconstructs a clothed 3D scan from images and predicts the underlying human body shape as well as segments the garments. This illustrates how we can support a range of applications related to multi-layer avatar generation with detected pose (right top) and try-on with pre-segmented garment models (right down).

1 INTRODUCTION

With the popularity of commodity three-dimensional (3D) scanners such as Microsoft Kinect or ASUS Xtions, it has become increasingly common to create 3D human models in place of traditional 2D images. How to further reconstruct the human shape as well as dressed garments for challenging human and clothing variations evolves as a cutting-edge technique requiring both refinement and robustness, which has attracted the attention of both the computer graphics and computer vision communities.

Capturing the accurate human shape and garments respectively from only a single complete pre-scanned 3D mesh of a performer is essential for numerous applications (see Figure 1) such as virtual try-on, biometrics, and body evaluation for gymnastics and sports. However, in reality, the human and clothing geometry 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 the extra fabric), and compression (ultra-tight, second-skin fit). Such a variety of clothing categories and looseness makes it very challenging to capture the accurate human shape, let alone the variety of human postures and the further garment reconstruction. Earlier approaches [Yang et al. 2016; Zhang et al. 2017] utilize statistical body models, like **Shape completion and animation of people (SPCAE)** [Angelov et al. 2005] and **A Skinned Multi-Person Linear Model (SMPL)** [Loper et al. 2015], to optimize the human shape parameters in the model space or the vertex displacements from the human template to the captured 3D scans. However, they are restricted by the generalization ability of the naked

human template to handle various clothing geometry. However, the learning-based techniques [Kanazawa et al. 2018; Jiang et al. 2020; Bhatnagar et al. 2019a; Alldieck et al. 2019a] have achieved significant progress recently to infer detailed geometry and body shape from images, but these image-based approaches suffer from scale ambiguity and inferior human shape accuracy.

In this article, we propose TightCap, a data-driven approach to capture both human shape and dressed garments robustly with only a single complete 3D scanned mesh of the performer, which can be generated from multi-view RGB cameras [Schönberger et al. 2016] or a single depth camera [Newcombe et al. 2015]. And, we outperform existing state-of-the-art methods to significantly produce 5.7 mm accuracy. Our novel scheme introduces the clothing tightness field that represents the displacements from various garments to the underlying human body model in a global UV texturing domain, so as to implicitly model the influence of various clothing categories, looseness, and human postures.

More specifically, to estimate the clothing tightness field for various garment categories and human postures in a data-driven manner, we first map the captured human 3D scan into a global geometry image [Gu et al. 2002], called clothed-GI. To this end, we extend the statistical human model SMPL [Loper et al. 2015] by subdividing its geometry features around the garment boundary (e.g., neck, wrist, waist, ankle, and etc.) for the generalization to various clothing categories. Then, an effective multi-stage alignment scheme is adopted to warp the enhanced template to the captured scan, which jointly leverages the skeleton, silhouette, and geometry information in a coarse-to-fine manner to handle human pose and garment variations. Second, we generate a hybrid feature embedding from the generated clothed-GI, including per-pixel texture, position, and normal. We further utilize a conditional **generative adversarial network (GAN)** to regress per-pixel clothing tightness in the UV texturing domain to handle human garment and posture variations, with the aid of a novel per-vertex tightness formulation and a new 3D dataset that consists of a large variety of clothing including T and long shirt, short/long/down coat, hooded jacket, pants, skirt/dress, and the corresponding 3D human shapes. Finally, we propose an effective optimization scheme to reconstruct both the inner human shape and the multi-layer dressed garments accurately from the predicted tightness map in the geometry image domain. Comprehensive experiments on both public and our captured datasets show that, compared with the state of the art, with only a single captured 3D scan, our approach significantly improves the accuracy of human shape prediction, especially under various loose and fitted clothing. We further demonstrate how the recovered multi-layer human geometry can be applied to automatically segment clothing from the human body on 3D meshes as well as cloth retargeting and animation. To summarize, our main contributions include:

- A novel and superior human shape and garment capture scheme with a single captured 3D scan, which models clothing tightness field to handle the garment and posture variations implicitly in the UV texturing domain.
- An effective multi-stage alignment approach to enable clothed-GI generation from the captured scan with the aid of an enhanced statistical model.

- An novel tightness map learning scheme based on a novel per-vertex tightness formulation as well as an effective optimization scheme to recover both the human shape and garments.
- To stimulate further research, we make available our **clothing tightness dataset (CTD)** of totaling 880 human models with 228 different garments under various human postures as well as the ground-truth human shapes.

2 RELATED WORK

Human and Garment Modeling. Most of the early works on human modeling can be categorized as **multi-view stereo (MVS)** vs. depth fusion-based approaches. The former approaches [Furukawa et al. 2013; Strecha et al. 2008; Newcombe et al. 2011b; Collet et al. 2015] employ correspondence matching and triangulation. For example, Collet et al. [2015] use a dense set of RGB and IR video cameras, producing high-quality 3D human results. The latter approaches [Bogo et al. 2015; Newcombe et al. 2015; Yu et al. 2017; Dou et al. 2016; Xu et al. 2019a; Li et al. 2020; Pang et al. 2021; Suo et al. 2021] use active sensors such as structured light and **Time-of-Flight (ToF)** range scanning (e.g., Microsoft Kinect I and II, respectively), which have a much lower cost. For example, DynamicFusion [Newcombe et al. 2015] compensates for geometric changes due to motion captured from a single RGB-D sensor. UnstructuredFusion [Xu et al. 2019a] and RoubustFusion [Su et al. 2020] utilizes several RGBD cameras to capture textured 4D human scans. Recently, many learning-based works utilize statistical body models, like SMPL [Loper et al. 2015], to capture/reconstruct human with clothing [Alldieck et al. 2019a; Zheng et al. 2019; Bhatnagar et al. 2019a, b] or recover human body [Pavlakos et al. 2018; Kanazawa et al. 2018; He et al. 2021; Chen et al. 2021] with the 2D keypoint detectors [Cao et al. 2018; Li et al. 2019]. Also, some notable works [Joo et al. 2018, 2019; Xiang et al. 2019], from Carnegie Mellon University, capture single or multiple 3D humans from 3D pose and body shape based on the multi-view panoptic studio. Similarly to their tasks, Pavlakos et al. [2019] can also capture face, hand, body, and expression with the SMPLX from a single image. For other 3D representations of clothed human, SiCloPe [Natsume et al. 2019] utilizes silhouette-based representation for modeling clothed human bodies using deep generative models. PIFu [Saito et al. 2019] proposes a pixel-aligned implicit representation to digitize detailed clothed humans from images, and PIFuHD [Saito et al. 2020] formulate a multi-level architecture to address the memory limitations of the hardware. PIFusion [Li et al. 2020] combines learning-based 3D recovery with volumetric non-rigid fusion to generate clothed human scans. Bhatnagar et al. [2020] combines implicit functions and parametric representations to reconstruct 3D models of people. Most of these reconstruction works focus on only one of two layers, the top surface layer of human scan or the skin layer of unclothed body shape. In TightCap, we model both the cloth and body layer with tightness field, a specific displacement from the cloth to the body, and build a data-driven method to estimate body shape and build a multi-layer avatar.

Garment modeling could be included in general human modeling, as we introduced, but it usually assumes the clothes and body skin belong to the same surface layer, like Alldieck et al. [2019a, 2018, 2019b], Lazova et al. [2019], Natsume et al. [2019],

Pumarola et al. [2019], and Saito et al. [2019, 2020]. Different from these methods, some other works [Neophytou and Hilton 2014; Pons-Moll et al. 2017; Yu et al. 2019] propose the idea of the multi-layer human model for garment modeling. DoubleFusion [Yu et al. 2018a] presents a 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 treats the inner body shape and outer clothing geometry. ClothCap [Pons-Moll et al. 2017] and SimulCap [Yu et al. 2019] also use SMPL as template model to help model garment from reconstructed human. For learning-based methods, DeepWrinkles [Lahner et al. 2018] proposes a data-driven framework to estimate garment wrinkles from the body motion. Moreover, CAPE [Ma et al. 2020] and TailorNet [Patel et al. 2020] use learning-based methods to generate a 3D mesh model of clothed people or directly proposes a neural garment model with pose and shape. ARCH [Huang et al. 2020] proposes a pose-aware model that produces 3D rigged clothed human avatars from a single image. Tiwari et al. [2020] propose a dataset of people with clothing size variation and model 3D clothing conditioned on body shape and garment size parameters. Most mentioned works of garment modeling, such as Alldieck et al. [2019b], Lazova et al. [2019], and Pumarola et al. [2019], utilize geometric image representation and the parametric body shape as the prior to reconstruct clothed human shape. Contrary to their tasks, we utilize the reconstruction results of clothed human, and focus on predicting the tightness map of different types of clothing. And our approach recovers both the personalized body shape under clothing and the reliable multi-layer avatar.

Shape under Clothing. Estimating body shape under clothing is more challenging, because the clothing occludes the original body shape. Earlier methods [Bálan and Black 2008; Zhang et al. 2017; Pons-Moll et al. 2017] employ a statistical or parametric 3D body model, like SCAPE [Anguelov et al. 2005] and SMPL [Loper et al. 2015]. [Bálan and Black 2008] build on the concept of the visual hull under the assumption that the clothing becomes looser or tighter on different body parts as a person moves. Wuhrer et al. [2014] estimate body shape from static scans or motion sequences by modeling body shape variation with a skeleton-based deformation. Other approaches [Hasler et al. 2009; Zhang et al. 2017; Pons-Moll et al. 2017; Yu et al. 2018a] utilize the parametric body model as the prior of shape, and attempt to optimize the body shape with the boundary constraint of the clothed human scans. ClothCap [Pons-Moll et al. 2017] utilizes a multi-part 3D model to estimate a minimally clothed shape under the clothing and tracks the deformations. DoubleFusion [Yu et al. 2018a] use SMPL to estimate body shape as one layer of their double-layer model. Yang et al. [2018] propose a statistical regression model for the variability of the clothing for capturing underlying shapes. However, they require the subject to wear ultra-tight or fitted clothing and only focus on several types of clothing. Human body shape estimation in wide and puffy clothing is significantly more difficult than in fitted clothing. Therefore, our approach is different from these approaches. We not only utilize a prior of template body mesh, but also exploit a data-driven manner for predicting the inner body shape under hundreds of various clothes. For learning-based methods, Newell et al. [2016], Pishchulin et al. [2016], and Wei et al.

[2016] learn articulated body poses of humans from their occluded body parts via convolutional networks. Lassner et al. [2017b] predict body segments and landmarks from annotated human pose datasets, and conducts body pose estimation with clothing and 3D body fitting. Lassner et al. [2017a] present a generative model of the full body in clothing, but focusing more on appearance generation than body shape estimation. In particular, HMR [Kanazawa et al. 2018] proposes an end-to-end ConvNet to recover the parameters of SMPL, for generating a 3D human body mesh from a single image. Pavlakos et al. [2018] refine this similar generated body mesh by projecting body shape back to the 2D image for full-body pose and shape estimation. Their techniques rely on parameter prediction from the body model and body pose accuracy. Different from these works that only capturing the pose and shape parameters from images, our approach models both clothing and body. The clothing layer produces significant help for more accurate and realistic results.

To contrast our approach with most related methods of shape estimation, we summarize these works in three different technical schemes, including the shape optimization under the geometric constrain of clothing, the learning-based body shape recovery from a single image, the displacement-based human capture in map or mesh (per-vertex). For geometric optimization, these methods, such as Yang et al. [2018] and Zhang et al. [2017], focus on optimizing the variations of the body with geometric constraints of clothing. Different from them, our approach builds a learning-based technique from different subjects in hundreds of different clothes, not only predicting the tightness field between the human body and the cloth layer but also segmenting clothing. With only a static input mesh rather than mesh sequence, our approach can leverage more commercial 3D sensors, such as ToF cameras on mobile phones, to support mobile virtual fitting tasks. For learning-based body shape recovery, like HMR [Kanazawa et al. 2018] and SMPL-X [Pavlakos et al. 2019] (based on learning-based pose estimator), our method can produce more reliable and accurate bodies with the help of the reconstruction of the clothing layer. For these methods based on displacement map [Lazova et al. 2019; Pumarola et al. 2019; Alldieck et al. 2019b] or per-vertex displacement [Alldieck et al. 2019a, 2018], most of these works focus on capturing a clothed human with the template model, rather than focusing on the personalized shape estimation. Their displacement/offset usually points from an impersonalized template body to the cloth, ignoring the influence of various clothing types on body shape. Instead, we consider the tightness field for shape recovery under hundreds of different clothes. Moreover, with the prediction of the tightness map and clothing mask, our approach can easily separate the clothing from the personalized body shape, benefiting many applications like virtual fitting and size measurement.

3 OVERVIEW

Our goal in this article is to reconstruct the human shape and corresponding garments accurately with a single captured 3D scan. To handle this challenging problem under the huge variations of clothing categories and human postures, we utilize the clothing tightness field in a data-driven manner, which relies on a representative dataset with both clothed human scans and corresponding

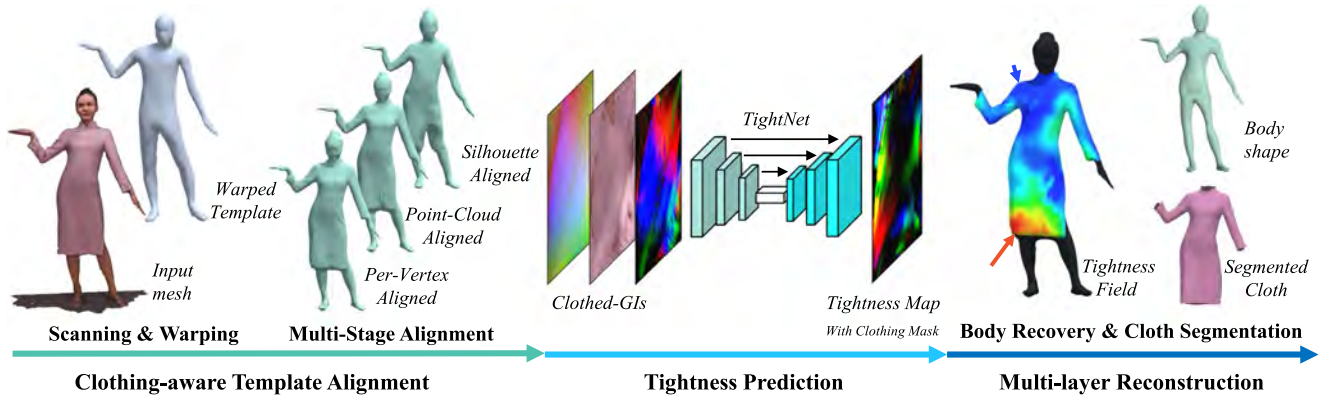


Fig. 2. The pipeline of TIGHTCAP. The first step is to warp our enhanced clothing-adapted SMPL with scanned mesh. Then, we deform warped mesh using *Multi-Stage Alignment*. Next, we estimate the tightness map and the clothing mask from mapped clothed-GI with *Tightness Prediction*. The final step, *Multi-layer Reconstruction*, is to recover body shape from the predicted tightness field on the mesh and segment cloth.

inner body shapes. Thus, we collect a new CTD with dressed human meshes under various clothing and poses, which are reconstructed via a dome system equipped with 80 RGB cameras using the MVS approach [Schönberger et al. 2016] or an RGB-D sensor using DynamicFusion [Newcombe et al. 2015]. The corresponding ground-truth naked human shapes are obtained via the same system and further manually re-posed and annotated by multiple professional artists.

Different from the reconstruction of clothed human, our system focus on learning personalized body shapes and multi-layer avatars from various clothed human scans. To formulate this system around the learning-based module, we build three modules: Template Alignment, Tightness Prediction, and Multi-layer Reconstruction. Figure 2 illustrates these three high-level components of the algorithm pipeline, which achieves considerably more accurate body shape reconstruction results than previous methods.

Human Template Alignment. To model the garment and posture variations implicitly with a learning-based approach, template alignment is an essential process for most network models. Considering the GAN-based CNNs on the image domain can effectively generate more high-frequency details of clothes than GCN with less memory in current hardware, a novel template alignment scheme is adopted to map the input 3D scan into the 2D clothed-GI. To align various clothes with the template model, our scheme relies on a garment-specific human template extended (Section 4.1) from the statistical model SMPL [Loper et al. 2015] with both pose parameters and embedded nodes. And, we utilize a multi-stage alignment with different sensitivities of parameters, including joints, embedded nodes, and vertices, which jointly leverage the skeleton, silhouette, and geometry information in a coarse-to-fine manner (see Section 4.3). After alignment, we utilize geometry image mapping to unwrap the aligned surface into 2D clothed-GI (Section 4.4).

Tightness Prediction. Based on the hybrid feature map from the above clothed-GI, we propose to predict the corresponding 2D clothing tightness map in a data-driven manner, which utilizes a novel per-vertex tightness formulation (see Section 5.1). While

GAN-based networks [Lazova et al. 2019; Pumarola et al. 2019; Alldieck et al. 2019b] achieve reliable results on reconstructed clothed human, we follow this GAN-based framework to predict the tightness pointing from the clothing layer to the body layer for personalized body shape estimation, and build a most effective learning framework based on conditional GAN, named TightNet (see Section 5.2).

Multi-layer Reconstruction. Finally, we utilize the predicted tightness map to reconstruct the multi-layer human shape and dressed garments accurately via an optimization scheme based on Gaussian kernels (see Section 5.3). This optimization can produce more reliable and stable results with the prior of template shape, and help to correct the 3D artifacts for challenging cases. Such multi-layer reconstruction results further enable various applications such as immersive cloth retargeting and avatar animation. The following sections provide the details of the full system.

4 CLOTHED HUMAN TEMPLATE ALIGNMENT

Under the canonical pose, the tightness field between various clothing and body shapes share a similar distribution with the same direction of gravity, e.g., the clothes tend to be loose around the human oter and crotch, and tight on the shoulder and chest, which implies that the tightness field of clothes with the underlying human body can be predicted in a data-driven manner. To this end, we utilize a garment-specific statistical human model on top of SMPL [Loper et al. 2015] (Section 4.1), and adopt an effective multi-stage alignment to warp the enhanced template to the captured scan, which jointly leverages the skeleton, silhouette, and geometry information in a coarse-to-fine manner to handle human pose and garment variations (Section 4.3). We further transfer the input 3D human scan into a global and regular geometric image [Gu et al. 2002], called clothed-GI (Section 4.4), to maintain the continuity of semantic human body distribution and implicitly model the variations of garments and human postures for the clothing tightness training and prediction. We provide the details of each design in the following.

4.1 Clothing Adapted Human Template Model

The most successful parametric human models, i.e., the SMPL [Loper et al. 2015] and SMPL-X [Pavlakos et al. 2019], focus on modeling the naked human body with various poses and shapes, while the displacement-based vertices movement produces the generalization to represent various clothes. To suit most clothed human meshes on shoes and with closed or even hidden hands in our dataset, we modify the human model SMPL [Loper et al. 2015]. We simplify face/hands/feet and subdividing its geometry features around the garment boundary (e.g., neck, wrist, waist, ankle.) to generalize these various clothing categories. We also simplify the template mesh around the ears, nose, and fingers for efficiency and rig the modified model with the skeleton defined by OpenPose [Cao et al. 2017; Simon et al. 2017], i.e., 23 joints for the main body part and 21 joints for each hand, as shown in Figure 3(a). The utilized **clothing-adapted SMPL (CA-SMPL)** model, denoted as \mathcal{M}_T , contains $N_M = 14,985$ vertices, $N_F = 29,966$ facets and $N_J = 65$ joints, which is summarized as follows:

$$\mathcal{M}_T = \{ \mathbf{M} \in \mathbb{R}^{N_M \times 3}, \mathbf{F} \in \mathbb{R}^{N_F \times 3}, \mathbf{J} \in \mathbb{R}^{N_J \times 4+3} \}, \quad (1)$$

where \mathbf{M} , \mathbf{F} , and \mathbf{J} denote the parameter sets of vertices, facets and joints, respectively. Different from the original SMPL model using both pose and shape parameters β and θ , we utilize the scale of bone \mathbf{S} to produce personalized lengths of bone without the shape parameters, which is similar to the scaling factors ϕ_j in Joo et al. [2018]. We only use the pose parameters to drive the template, thus do not need the joint regressor $J(\beta; \mathcal{J}, \bar{\mathbf{T}}, \mathbf{S})$ in SMPL. Hence, the utilized pose parameters in the adopted human template are as follows:

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

including the joint angles Θ with axis-angle representation, the scaling factors \mathbf{S} of each joint along the bone direction, and the global translation \mathbf{m} . Furthermore, let $M(\hat{\mathbf{J}})$ denote the skeletal motion of the human template after applied the joint parameters $\hat{\mathbf{J}}$, and $\hat{\mathbf{M}}$ represents the warped vertices of the template.

To enable robust alignment, we utilize the **embedded deformation (ED)** [Sumner et al. 2007] by sampling the ED graph on the above enhanced human template, which is formulated as follows:

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

where N_G is the number of nodes in ED graph. Then, the warping function G_k of the k th node applied to a vertex \mathbf{v} consists of the rotation $\mathbf{R}_k \in \text{SO}(3)$ and the translate $\mathbf{t}_k \in \mathbb{R}^3$, which is formulated as

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

where $\hat{\mathbf{g}}_k \in \mathbb{R}^3$ indicates the canonical position of the k th node. Thus, the i th vertex $\mathbf{v}_i, i \in [1, N_M]$ on the human template after applied the ED motion \mathcal{G} is formulated as

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

Here $\hat{\mathbf{v}}_i$ is the canonical position of vertex i and $\mathbf{w}_{i,k}^G$ is the skinning weight between the i th vertex and the k th ED node according to the Euclidean distance. Please kindly refer to Sumner et al. [2007] for more details about the setting of skinning weight.

4.2 Human Scan Reconstruction

Our raw 3D human scans can be captured using a multi-view dome system or a single-view depth sensor. For the former setting, we reconstruct using the MVS approach [Schönberger et al. 2016] from multi-view human images captured by a dome system equipped with 80 cameras. We also estimate the 2D human joints as in Cao et al. [2017] and Simon et al. [2017] for the 80 original views and the 30 synthetic views (see Section 4.3) in turn. While the 80 original views help to locate the 30 synthetic views for accurate 2D joint estimation, we obtain the 3D joints through triangulation as in Triggs et al. [1999] for initialization. The other setting is to reconstruct non-rigidly deforming human from commodity sensors, like ToF cameras on mobile phones and Kinect. We implement DynamicFusion [Newcombe et al. 2015] as our reconstruction method to acquire dynamic human scans. We also estimate the 2D human joints for this single view and project it to the depth map as our rough 3D joints.

We reconstruct most of our clothed human mesh sequences in our dataset with the multi-view system, while a small part of them are from the single-view depth system. The main reason is that the multi-view system can capture more realistic clothing movement rather than a single-view system without the details on the back. Meanwhile, we utilize the depth sensor to capture the dark color clothing without a rich image feature. For more generalization, we also utilize many synthetic clothed avatars from Adobe Fuse CC. Although these meshes cannot provide realistic motion of body and clothing, the various color of clothing and skin are convenient for data augmentation. For more detail of our dataset, please refer to Section 6.1.

Our multi-stage deformation scheme can adapt to different application settings. More technical details for each stage are provided as follows.

4.3 Multi-Stage Alignment

Both two popular learning-based approaches, the standard CNNs (image-domain) and the GCN (mesh-domain), need the ground truth of registered meshes. Considering the lack of existing large datasets of registered meshes, especially for the generalization to these various clothes in our dataset, the data-driven mesh alignment approaches are not suitable. Meanwhile, the appearance of clothes also need to be maintained as one important feature for generalization. Thus, we avoid manually mark the points/patterns on the clothes for the correspondence, which is used in the registration of naked human meshes in FAUST [Bogo et al. 2014]. To this end, we adopt a novel multi-stage alignment scheme to transfer the input 3D human scan into the consistent human template, which shares the same topology for various garment categories and human postures. Note that our scheme consists of the silhouette-based, point cloud-based and per-vertex deformation stages to optimize the non-rigid motions from the enhanced human template to the input 3D scan in a coarse-to-fine manner, as illustrated in Figure 3.

Silhouette-based deformation. We first align the enhanced template with the silhouette information from the captured scan using a coarse ED-graph to handle error-prone places due to holes or noise on the raw human scan. To fetch the silhouette

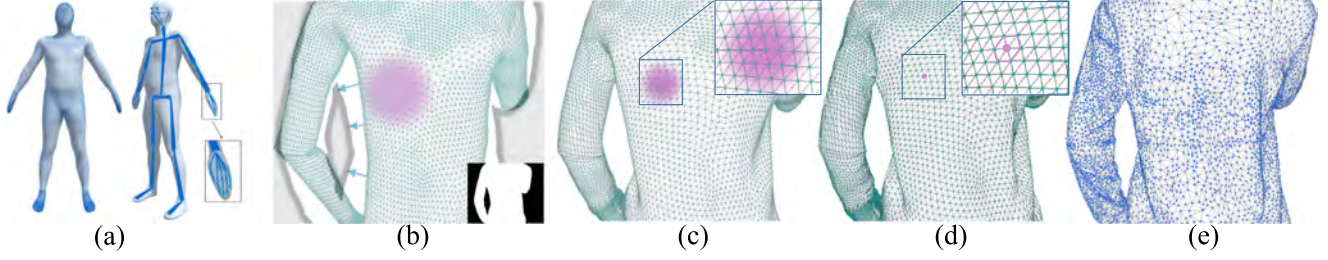


Fig. 3. Each stage of our alignment approach. (a) The template model for alignment. (b) The first stage, silhouette-based deformation. (c) The second stage, point cloud-based deformation. (d) The third stage, per-vertex deformation. (e) The referenced mesh (target mesh). The pink regions indicate the different ranges of ED nodes, while the wires are topological structure.

information, we utilize a virtual capture system with $N_C = 30$ synthetic cameras to view different areas of the captured 3D mesh. Note that for capturing the neck, ankles, and wrists, we set two synthetic cameras orthogonal to each other. Besides, five cameras with different view angles are arranged to capture the upper and lower body torso, respectively. The resulting synthetic camera setting is formulated as follows:

$$C = \left\{ (c_j \in \mathbb{R}^6, w_j^C \in \mathbb{R}^1) \mid j \in [0, N_C) \right\}, \quad (6)$$

where c_j denotes extrinsic parameters of a camera and $w_j^C \in [0.5, 1]$ represents the weighting factor for two different camera positions (0.5 for the torso regions and 1 for capturing limbs). Such a semantic weighting strategy further improves the alignment results, especially for those boundary regions. We first warp the original human template model M_T with the rough 3D joints J_{mv} as our initial mesh for the following alignment. Inspiring by previous silhouette deformation method [Xu et al. 2018], we render the high-resolution silhouette masks of the captured scan for all the virtual views and phrase the coarse-level alignment by solving the following non-linear least-squares optimization problem:

$$E_S(\mathcal{G}) = E_{mv}^S(\mathcal{G}) + \lambda_{reg}^S E_{reg}^S(\mathcal{G}). \quad (7)$$

Similarly to Xu et al. [2018], our multi-view silhouette-based data term E_{mv}^S measures the 2D point-to-plane misalignment:

$$E_{mv}^S(\mathcal{G}) = \sum_{j \in C} \frac{w_j^C}{|\mathbf{v}_j^S|} \sum_{k \in \mathbf{v}_j^S} \|\mathbf{n}_k^T \cdot (P_j(\mathbf{v}_i(\mathcal{G})) - \mathbf{p}_k)\|_2^2, \quad (8)$$

where \mathbf{v}_j^S is the vertex set of virtual silhouettes of the input scan and $P_j(\cdot)$ is the projection function of the j th camera. For each silhouette point $\mathbf{p}_k \in \mathbb{R}^2$ with the 2D normal $\mathbf{n}_k \in \mathbb{R}^2$, we search its corresponding deformed vertex in the utilized human template, denoted as \mathbf{v}_i , found via a projective look-up method in an **Iterative Closest Point (ICP)** manner.

Similarly to Sorkine and Alexa [2007], the regularity term E_{reg}^S produces locally as-rigid-as-possible motions to prevent overfitting to the 3D scan input, which is formulated as

$$E_{reg}^S(\mathcal{G}) = \sum_{k \in \mathcal{G}} \sum_{n \in N_k} w_{k,n}^N \|(\mathbf{g}_k - \mathbf{g}_n) - \mathbf{R}_k(\hat{\mathbf{g}}_k - \hat{\mathbf{g}}_n)\|_2^2, \quad (9)$$

where $N_k \in \mathcal{G}$ is the 1-ring neighborhood of the k th ED node and $w_{k,n}^N$ denotes the KNN weight between the k th and n th nodes.

For each ICP iteration, the resulting optimization problem in Equation (7) is solved effectively using the Conjugate Gradient method. Let M_S denote the vertices of the deformed template after the silhouette-based optimization.

Point cloud-based deformation. After the above silhouette-based alignment, we re-sample a finer ED graph to model the fine-detailed geometry information in the input scan. For clarity and simplification, we reuse \mathcal{G} to represent the ED motion from the previous results M_S to the input 3D mesh. Then, the full energy function for current fine-detailed alignment is formulated as

$$E_D(\mathcal{G}) = E_{data}^D(\mathcal{G}) + \lambda_{reg}^D E_{reg}^D(\mathcal{G}). \quad (10)$$

Here, the data term $E_{data}^D(\mathcal{G})$ measures the fitting from M_S to the input mesh, which is formulated as the sum of point-to-point and point-to-plane distances:

$$E_{data}^D(\mathcal{G}) = \lambda_{point}^D \sum_{i \in M} \|\mathbf{v}_i(\mathcal{G}) - \mathbf{v}_i^c\|^2 + \lambda_{plane}^D \sum_{i \in M} (\mathbf{n}_i^T(\mathcal{G}) \cdot (\mathbf{v}_i(\mathcal{G}) - \mathbf{v}_i^c)), \quad (11)$$

where λ_{point}^D and λ_{plane}^D are the weights to balance two kinds of distances; $\mathbf{n}_i(\mathcal{G})$ represents the normal of the deformed vertex $\mathbf{v}_i(\mathcal{G})$. Note that for each $\mathbf{v}_i(\mathcal{G})$, its corresponding point \mathbf{v}_i^c in the scan is found via the same look-up method in an ICP manner.

The regularity term $E_{reg}^D(\mathcal{G})$ here shares the same formulated as the one in Equation (9) and the full energy is solved using the same conjugate gradient solver. After this point cloud-based alignment on a finer scale, the vertices of the deformed template are denoted as M_D .

Per-vertex deformation. Finally, we refine the deformation from ED graph-based non-rigid result M_D to the input 3D mesh via per-vertex optimization, so as to improve the alignment accuracy, especially for those local regions with fine details like clothing wrinkle and boundary, which is formulated as follows:

$$E_V(M) = E_{data}^V(M) + \lambda_{reg}^V E_{reg}^V(M). \quad (12)$$

Here, similarly to Equation (11), the data term E_{data}^V further measures the per-vertex fitting by minimizing the both the

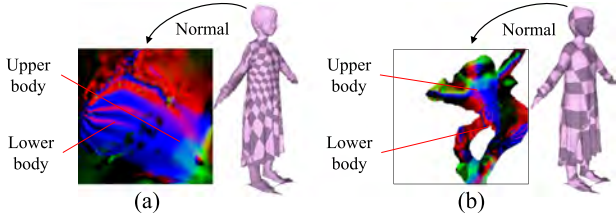


Fig. 4. The comparison of two feature map with different mapping methods. (a) The normal map using geometry image [Gu et al. 2002]. (b) The normal map using OptCuts algorithm [Li et al. 2018].

point-to-point and point-to-plane distances:

$$E_{\text{data}}^V(\mathbf{M}) = \lambda_{\text{point}}^V \sum_{i \in \mathbf{M}} \|\mathbf{v}_i - \mathbf{v}_i^c\|^2 + \lambda_{\text{plane}}^V \sum_{i \in \mathbf{M}} (\mathbf{n}_i^T \cdot (\mathbf{v}_i - \mathbf{v}_i^c)). \quad (13)$$

We utilize the same regularity term E_{reg}^V from Xu et al. [2018] to prevent over-fitting to the 3D input scan. Let \mathbf{M}_V denote the final optimized vertices of the human template. Figure 3 shows the intermediate alignment results of all these stages, which demonstrates the effectiveness of our multi-stage alignment scheme. After the multi-stage alignment, we obtain a deformed human template that is not only fitted to the captured 3D human scan but also owns the global consistent topology.

4.4 Geometry Image Representation

To benefit the tightness prediction with the effective image-to-image translation network, we map the clothed 3D human body mesh with the consistent topology into a regular 2D UV image, which has been proved to be effective in previous works, like Lahner et al. [2018], Xu et al. [2019b], and Alp Güler et al. [2018].

There are many methods to generate a 2D mapping from a 3D mesh. We choose two representative methods for comparison. One is the mapping approach of the geometry image [Gu et al. 2002] with gapless filling but relative large distortions denoted as $M_{GI}(\cdot)$. The other method is OptCuts [Li et al. 2018], denoted as $M_{Opt}(\cdot)$, which automatically seeks the best seam for cutting and generates an image with lower distortion but contains gap area. Figure 4 illustrates the mapping results for both methods, and we utilize the geometry image [Gu et al. 2002] to achieve a more smooth feature representation.

A quantitative comparison of these two methods for our full pipeline is provided in Section 6.2. To generate consistent 2D feature embeddings, we map the positions, normals, and RGB colors of each vertex into its 2D map using the mapping approach [Gu et al. 2002]. Linear interpolation is further conducted to fill the hybrid 2D clothed geometry image, which is denoted as clothed-GI.

5 TIGHTNESS PREDICTION

Previous human reconstruction methods, including methods based on scanned depth map(s) [Newcombe et al. 2011a, 2015; Collet et al. 2015; Dou et al. 2016], and silhouette(s) [Baker et al. 2005; Cheung et al. 2003b, a; Corazza et al. 2006; Mikhnevich and Hebert 2011; Xu et al. 2018], represent the human body as a single layer. Recently,

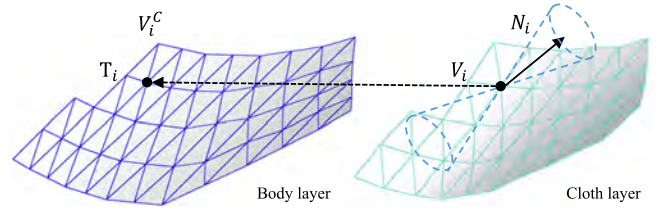


Fig. 5. Illustration of tightness \mathbf{T}_i of vertex i on a mesh, which is the black dotted directed vector pointing from cloth to body.

Neophytou and Hilton [2014], Pons-Moll et al. [2017], Yu et al. [2018b, 2019], and Zhang et al. [2017] proposed the idea of multi-layer body shape recovery. We extend this idea and define a novel clothing tightness formulation, which describes the relationship between the underlying human body shape and the various garment layers (Section 5.1). Subsequently, we propose a conditional GAN to predict the clothing tightness map in a data-driven manner, based on the 2D hybrid clothed-GI input and our novel tightness formulation (Section 5.2). We also introduce an effective optimization scheme to reconstruct both the inner human shape and the multi-layer dressed garments accurately from the predicted tightness map in the geometry image domain (Section 5.3).

5.1 Tightness Measurement

Different to previous human modeling methods [Alldieck et al. 2018, 2019a] that model the displacements from the body template SMPL to the outer garment surface, we introduce the clothing tightness field to measure the displacements from various garments to the personalized underlying human body model. Recall that our CTD contains both the dressed human meshes under various clothing and poses with the ground-truth underlying body models. To model the tightness field from the clothing layer to the real body layer using our CTD dataset, a straightforward formulation is to align our enhanced SMPL model in Section 4.1 to both the dressed mesh and the corresponding inner body mesh simultaneously using our multi-stage alignment method in Section 4.3. Then, the per-vertex tightness is formulated on top of these two non-rigid aligned human templates that share the same topology:

$$\mathbf{T}_i = \mathbf{v}_i - \mathbf{v}_i^c. \quad (14)$$

Here, \mathbf{v}_i and \mathbf{v}_i^c are the i th vertex of the two templates aligned to the clothing layer and the body layer, respectively. Note that the direction of \mathbf{T}_i indicates the corresponding pairs from clothing to the inner body shape, while its magnitude is the euclidean distance between the two corresponding vertices. We can further define the tightness field, denoted as the matrix \mathcal{T} , on top of our enhanced SMPL model as follows:

$$\mathcal{T} = \{\mathbf{T} \in \mathbb{R}^{N_M \times 3}\}, \quad (15)$$

where N_M is the number of the template vertices, same as Equation (1).

However, this straightforward formulation above fails to model the exact correspondences between the clothing layer and the body layer, because the two non-rigid alignments from the template to the dressed mesh and the inner body model are performed totally independently, and the one-to-one correspondences are

fragile to the alignment error. To this end, we formulate the per-vertex tightness as the one-to-many correspondences between the human template aligned to the clothing layer and the ground-truth body model directly, jointly considering the direction and distance information of the clothing layer. For a vertex \mathbf{v}_i on the aligned human template of the dressed mesh, we calculate its approximated tightness $\hat{\mathbf{T}}_i$ as follows:

$$\hat{\mathbf{T}}_i = \frac{\sum_{\mathbf{v}_r^c \in \mathcal{N}_1^c} K_G(\mathbf{v}_i - \mathbf{v}_r^c) + \sum_{\mathbf{v}_d^c \in \mathcal{N}_2^c} K_G(\mathbf{v}_i - \mathbf{v}_d^c)}{\|\mathcal{N}_1^c\| + \|\mathcal{N}_2^c\|}. \quad (16)$$

Here, the one-to-many correspondence set \mathcal{N}_1^c denotes the local vertices set of the ground-truth human body shape found via a ray-tracing operation along the normal direction of \mathbf{v}_i within a double-cone with an aperture of 30° , while \mathcal{N}_2^c is the set of the 20 closest vertices of \mathbf{v}_i in the ground-truth body mesh in terms of Euclidean distance, as shown in Figure 5. Note that $K_G(\cdot)$ is the Gaussian weighting function based on the angle between two vertex normals to enable smooth tightness field estimation.

After the above tightness field estimation from the dressed template to the ground-truth body model, we further apply the same strategy but change the target and source. We calculate this per-vertices tightness from the template for more reliable correspondence and combine such bi-directional estimations to obtain the ground-truth 3D clothing tightness field with the same topology of the enhanced SMPL template. Specifically, we utilize a linear weighting for two tracking results, which is usually 0.8 for the body template to the dressed template and 0.2 for the other. Then, by using the same mapping operation in Section 4.4, we generate the tightness map in the geometry image domain so as to enable end-to-end learning of the clothing tightness field and implicitly to model the influence of various clothing categories, looseness, and human postures.

5.2 TightNet Architecture

Based on the tightness map above and the hybrid 2D feature map from Section 4.4 in the global geometry image domain, we thus propose to train a pix2pix-style [Isola et al. 2017] convolutional neural network, which is the most effective image-to-image translation network structure verified in many previous works [Lazova et al. 2019; Pumarola et al. 2019; Alldieck et al. 2019b]. We denoted this net as TightNet to infer the clothing tightness map and garment masks in an end-to-end manner. In the following, we provide more details about the input/output pair, the used network architectures, losses, and training schemes.

The input to our TightNet is the hybrid feature embedding in the clothed-GI from the raw 3D scan, including the vertex positions, normals, and RGB colors, while the output consists of the predicted tightness map as well as the masks for both the upper and lower garments, so as to enable further multi-layer garment reconstruction. Note that for those clothing categories, we set two main categories, upper garment and lower garment, including shirt, coat, jacket, and dress for upper garment, pant and skirt for the lower garment. We model the upper garment and the lower garment but take the whole dress (not including the skirt) as one upper garment. For the prediction of garment masks, we utilize the same TightNet to predict tightness map with the

five channels of predicted results (three channels for tightness map, two channels for the mask of upper/lower garments), and supervise the mask with the provided segmented garment in our dataset. We use the L1-loss for the garment mask training, which is the same as the tightness map supervision. Thanks to our unique 2D mapping scheme based on geometry image [Gu et al. 2002], both the input and output share the same semantic 2D structure, so as to implicitly handle the huge variations for clothing categories, garment looseness, and human postures.

The network in our TightNet is a conditional Generative Adversarial Network (Pixel2Pixel) [Isola et al. 2017], which learns a mapping from the input hybrid feature map to our tightness map and mask map. More specifically, the generator is U-Net [Ronneberger et al. 2015] encoder-decoder structure with skip connections between convolution-ReLU-batch norm down- and up-sampling layers, which can share information between the input and output. The input is the nine channels of hybrid feature map with 224×224 resolution, including every three channels for vertex positions, normals, and RGB colors, while the output is the five channels of predicted results with the same resolution including three channels for tightness map, two channels for upper/lower clothing mask maps.

In our discriminator, we utilized PatchGAN [Isola et al. 2017] discriminator. However, we take this architecture to predict the full-body tightness field rather than image style transfer. Unlike the original PatchGAN [Isola et al. 2017] to restrict their attention to small local patches, we take the full feature map as the random patches and further normalize these patches in our GAN discriminator. We train the TightNet with the well-established L1-loss instead of L2-loss for fewer blurring artifacts. Benefiting from our tightness predictor, we can extract the hidden information between different clothing appearances and the tightness, while the input positions and normals also help our predictor to consider the effect of the current human pose.

5.3 Shape Recovery Under Clothing

To fine-tune the predicted body from TightNet, especially fixing the noise around the local regions like the oxter and crotch, we utilize both the predicted tightness field and the prior of the warped template. Thus, based on the clothing tightness and mask prediction above, we propose an effective optimization scheme to reconstruct both the inner human shape and the multi-layer dressed garments accurately.

Shape recovery. Recall that our clothing tightness field indicates the displacements from the garment layers to the inner human body layer. To cover the body shape from the tightness prediction, we first utilize the inverse function of the mapping $M_{GI}^{-1}(\cdot)$ in Section 4.4 to generate the per-vertex tightness field $\hat{\mathcal{T}}$ on the final aligned template mesh \mathbf{M}_V , where $\hat{\mathcal{T}} = \{\hat{\mathbf{T}} \in \mathbb{R}^{N_M \times 3}\}$. Then, a straightforward solution to obtain the inner body shape \mathbf{M} based on our tightness field formulation is as follows:

$$\mathbf{M} = \mathbf{M}_V + \hat{\mathbf{T}}. \quad (17)$$

However, such solution above suffers from the tightness field estimation error, especially for those local regions around oxter and crotch under unusual human poses, leading to visually unpleasant

body shape recovery. To this end, we propose a simple and effective optimization scheme to estimate a smoother body shape by solving the following least-squares energy function:

$$E_{\text{body}}(\mathbf{M}) = \lambda_{\text{fit}}(\mathbf{M}_V + \mathbf{T} - \mathbf{M}) + \lambda_{\text{smooth}}(\mathbf{M} - K_G(\mathbf{M})) + \lambda_{\text{reg}}(\mathbf{M} - \mathbf{M}_{\text{warp}}). \quad (18)$$

Here the first data term utilizes our tightness field formulation similar to Equation (17), while the second term enables smooth body shape estimation via the same Gaussian kernel $K_G(\cdot)$ defined in Equation (16). In the final regular term, the warped vertex matrix \mathbf{M}_{warp} denotes the warped body template after the first ICP iteration of the first stage optimization in our multi-stage alignment in Section 4.3. Such regular term forces the optimized body shape to be closed to the utilized human template to penalize unnatural body shapes. All the parameters for these three terms are empirically set to be 1, 0.1, and 0.05, respectively. Finally, by solving the least-squares problem in Equation (18), we reconstruct an accurate and visually pleasant body shape of the input 3D human scan.

Clothing segmentation. Besides the body shape recovery above, we utilize the output of multiple garment masks from our TightNet to automatically segment clothing from the human body on 3D meshes so as to enable further cloth retargeting or animation applications. Since the output masks are not accurate enough to segment the clothing directly in the 3D space, we utilize the following Markov Random Fields in Pons-Moll et al. [2017] to solve the per-vertex clothing label $v_i \in \mathbf{v}$ for each vertex in our final aligned template mesh \mathbf{M}_V :

$$E_{\text{cloth}}(\mathbf{v}) = \sum_{i \in \mathcal{T}} \varphi_i(v_i) + \sum_{(i,j) \in \mathcal{T}} \psi_{ij}(v_i, v_j). \quad (19)$$

To enable fully automatic segmentation, we replace the manually defined garment prior of the original optimization in Pons-Moll et al. [2017] with our predicted garment masks from TightNet. Please kindly refer to Pons-Moll et al. [2017] for more details about how to solve the energy function above.

6 EXPERIMENTAL RESULTS

In this section, we evaluate our method on a variety of challenging scenarios. We first report the implementation of the details of our whole method and two utilized datasets, followed by the evaluation of our main technical contributions. We also include both qualitative and quantitative comparisons with previous state-of-the-art methods. The applications and limitations regarding our approach are provided in the last two subsections.

Implementation details. We run our experiments on a PC with an NVIDIA GeForce GTX 1080Ti GPU, a 4.2-GHz Intel Core i7-7700K CPU, and 16 GB RAM. Our optimized code takes about 12 s per 3D human scan, which divides to 5–8 s for the multi-stage human template alignment (about 1.5, 2, and 3 s for each stage, respectively), 0.5 s for the tightness map prediction, and 0.5 s for the shape recovery from tightness field. All the energy functions of the multi-stage alignment are solved via a GPU-based Conjugate Gradient solver. In all experiments, we use the following empirically determined parameters: $N_{GS} = 1407$, $\lambda_{\text{reg}}^S = 10$; $N_{GD} = 2103$,



Fig. 6. Sample data from our dataset: three real human subjects with scanned body shape meshes, the segmented clothes, and one synthetic model (rightmost) for pre-training.

$\lambda_{\text{reg}}^V = 7$, $\lambda_{\text{point}}^D = 0.5$, $\lambda_{\text{plane}}^D = 1.5$; $\lambda_{\text{reg}}^V = 1$, $\lambda_{\text{point}}^V = 1$, and $\lambda_{\text{plane}}^V = 1.5$. For the tightness prediction, we set the resolution of clothed-GI as 224×224 . Note that the clothing tightness predictor is pre-trained on our dataset, which takes about 3 hours, and the training on CTD needs 4 hours.

6.1 Dataset

For a thorough evaluation of our method, we utilize both the most popular public dataset and a much larger captured dataset for the task of reconstructing both the human body shape and garments.

Bodies under Flowing Fashion Dataset. The BUFF dataset [Zhang et al. 2017] is the most popular public dataset for body shape estimation, which contains three males and three females models wearing two types of clothing (t-shirt/long pants and a soccer uniform). It provides a dynamic sequence for each subject but only with the per-vertex color rather than extra high-quality RGB textures. BUFF also contains the body shapes under the general T pose without garments as the ground truth. Since the data size of BUFF is far from enough to train our tightness prediction network, we only utilize the sampling frames from their scans as input and predict the tightness with the pre-trained model using our dataset.

CTD. To model clothing tightness in a data-driven manner, we propose a new benchmark dataset, which contains 880 dressed human meshes with both the body geometry and segmented individual pieces of garments. 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 dynamic daily actions, including boxing, dancing, playing badminton, keep-fit exercise, and so on. For garment modeling, our CTD contains 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 each dynamic sequence, we capture 400–500 frames under 30 fps and evenly sample 40–50 frames for our dataset. Note that most 3D meshes are reconstructed via a dome system equipped with 80 RGB cameras using the MVS approach [Schönberger et al. 2016], with about 50,000 vertices, 100,000 facets, and a 4K texture, while few 3D meshes are reconstructed via the DynamicFusion approach [Newcombe et al. 2015] with very similar quality. The corresponding



Fig. 7. The gallery of our CTD. The first column from top to bottom: (1) Sampling of various clothed human, including synthetic models (rightmost) and two body shape scans. (2) Various segmented clothes with “A” pose only. (3) Carven body shapes with “A” pose only. The second column is three typical dynamic sequences in our dataset including clothed human, segmented clothes, and carven body shapes.



Fig. 8. The gallery of our results. From bottom to top, the captured meshes, predicted tightness field, recovered body shapes and segmented clothes.

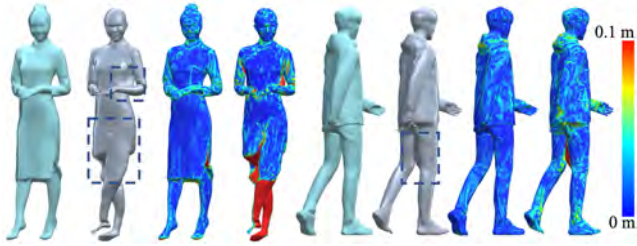


Fig. 9. The comparison between our enhanced CA-SMPL and the high-resolution SMPL [Loper et al. 2015] (27,554 vertices) for the clothing alignment. The green meshes are the results of our alignment algorithm using our enhanced SMPL, while the gray meshes are the results of the same algorithm using SMPL. The alignment error is color-coded from blue to red.

ground-truth naked human shapes are obtained via the same system, and then 5 artists further manually segment each piece of clothing and carve the body shape out from the raw mesh. We then generate the ground-truth tightness field using the novel formulation in Section 5.1. Figure 6 illustrates the high-quality examples from our dataset, while Figure 7 further provides the gallery of our whole dataset. We will make our dataset publicly available. To train our tightness prediction network in Section 5.2, we split the data into 80% vs. 20% with considering the identities. With more dynamic frames for the training set to provide more training instances, we keep half of the identities that do not appear in the training set. We also generate 800 clothed human meshes with synthetic avatars using Adobe Fuse CC for the pre-training of our network. Figure 8 demonstrated the multi-layer results of our approach, where both the human body shape and the garments under various clothing tightness and human postures are faithful reconstructions.

6.2 Evaluation

In this section, we evaluate our individual technique components, i.e., the human template alignment, the clothing tightness prediction, as well as the shape recovery from the tightness map in the following contents, respectively.

Alignment evaluation. We first evaluate the effectiveness of our clothing-adaptive human template (CA-SMPL) in Section 4.1 by comparing it to the original SMPL model [Loper et al. 2015] using the same multi-stage alignment algorithm. As shown in Figure 9, the original SMPL suffers from severe alignment error, especially for those local regions like crotch and neck due to the limited generation ability to handle clothing variations. In contrast, our enhanced template is more robust to both the clothing variations in our dataset, leading to the improvement in the clothing alignment accuracy.

We further evaluate our multi-stage alignment in Section 4.3 by analyzing the influence of each stage, both qualitatively and quantitatively. Let *Silhouette based* and *Point cloud based* denote the variations of our alignment method after the first and the second stages, respectively. Besides, we further compare to the alignment baseline [Tong et al. 2012], which directly aligns a 3D mesh with the input point cloud, denoted as *Non-rigid*. In Figure 10

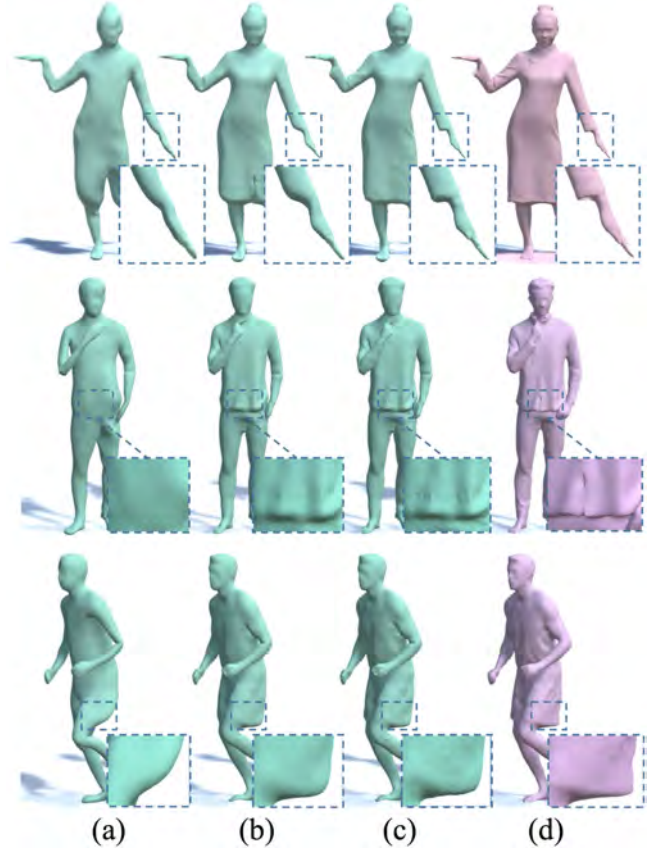


Fig. 10. The qualitative evaluation of our multi-stage alignment method. (a) The results after silhouette-based alignment. (b) The results after point cloud-based alignment. (c) The results after per-vertex alignment. (d) The captured meshes (Target meshes).

we present the qualitative results of each stage for various challenging inputs. Note that our full scheme achieves superior alignment results and can even float the crack on skirt and match the clothing boundary around the wrist. Furthermore, the qualitative and quantitative results in Figure 11 clearly demonstrate the effectiveness of each stage in our alignment scheme. Meanwhile, without the good initial state provided by the silhouette-based deformation, the baseline [Tong et al. 2012] cannot converge to a good result.

For further quantitative evaluation on our dataset, we utilize Metro [Cignoni et al. 1998], which is based on Hausdorff distance for comparing the difference of two meshes, and calculate its normalized *Mean* and **Root-Mean-Square (RMS)** as the metrics with a normalized factor (3 in our setting). We also calculate the per-vertex error as a relative quantitative metric, denoted as *Error (mm)*. Table 1 shows that our full pipeline consistently outperforms the other baseline variations in terms of all these quantitative metrics. We also compare with two SMPL-based registration approaches used in Lazova et al. [2019] and IP-Net [Bhatnagar et al. 2020]. They are based on shape/pose parameters fitting and non-rigid deformation with displacement from original mesh or reconstructed mesh from implicit function. However, we utilize not only

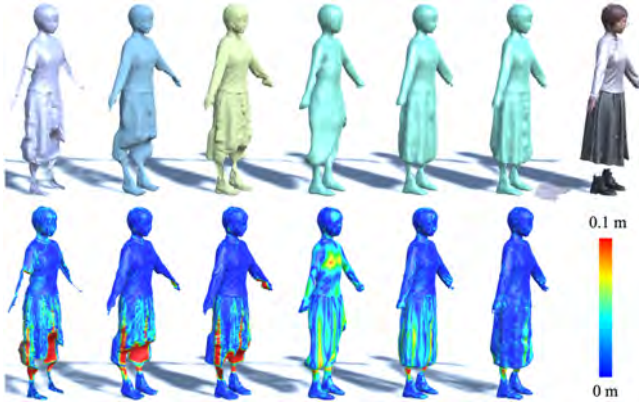


Fig. 11. The qualitative comparison of our multi-stage alignment method on per-vertex error. Non-rigid [Tong et al. 2012] is a baseline to directly align a 3D mesh with the captured mesh. From left to right, aligned mesh using Non-rigid [Tong et al. 2012] (gray), SMPL-based non-rigid method [Lazova et al. 2019] (blue), implicit function-based registration IP-Net [Bhatnagar et al. 2020] (yellow), our aligned mesh with silhouette, our aligned mesh with point cloud, our final aligned mesh, and target clothed mesh. The second row is per-vertex alignment error colored from blue (good) to red (bad).

Table 1. Comparison of Alignment Methods for Clothed Human Mesh

Method	Mean↓	RMS↓	Error (mm)↓
Non-rigid [Tong et al. 2012]	0.448%	0.762%	13.44
Lazova et al. [2019]	0.340%	0.563%	10.02
IP-Net [Bhatnagar et al. 2020]	0.324%	0.528%	9.72
Silhouette based [Xu et al. 2018]	0.494%	0.779%	14.82
Point cloud based	0.286%	0.585%	8.58
Ours	0.263%	0.521%	7.89

Non-rigid [Tong et al. 2012] is a baseline to directly align a 3D scanned mesh with the input point cloud. Lazova et al. [2019] is a SMPL-based registration to align both shape/pose parameters before non-rigid deformation. IP-Net [Bhatnagar et al. 2020] is an implicit function-based registration with SMPL model. *Silhouette based* [Xu et al. 2018] is a baseline to align a 3D mesh from silhouette only, which is also our first stage. *Point cloud based* is our second stage using the results of the first stage as an initial value. ↓ means the smaller is better. *Mean* and *Root-Mean-Square (RMS)* are the metrics of Hausdorff distance [Cignoni et al. 1998] from the sampling of the targets, normalized with the bounding box diagonal of all clothed meshes, which is 3 in our setting. *Error (mm)* represents the per-vertex error with millimeter, using the same Hausdorff distance.

the original 80 cameras but also the 30 synthetic cameras (see Section 4.3) to support the initial 3D poses, and the silhouette and point cloud also provide coarse-to-fine references. Thus, our multi-stage alignment produces 2 to 3 mm improvement on accuracy.

We then evaluate each alignment stage’s correspondences error on FAUST [Bogo et al. 2014] quantitatively. The FAUST [Bogo et al. 2014] dataset provides the ground-truth correspondences although the models are unclothed. As shown in Table 2, each stage of our method gets a 2- to 8-cm decrease for the correspondence error.

This not only highlights the contribution of each alignment stage but also illustrates that our approach can robustly align the enhanced human template to the input 3D scan.

TightNet evaluation. Here, we evaluate our TightNet quantitatively by comparing with two variation baselines using L2 loss or OptCut [Li et al. 2018] for 2D mapping, denoted as *Baseline L2* and

Table 2. Evaluation of our Multi-stage Alignment Method on the FAUST [Bogo et al. 2014] Dataset

Method	Mean↓	RMS↓	Error (mm)↓	Cor. Error (mm) ↓
Silhouette based	0.316%	0.426%	9.48	24.37
Point cloud based	0.083%	0.186%	2.49	16.30
Ours	0.081%	0.157%	2.43	13.93

Silhouette based [Xu et al. 2018] is a baseline to align a 3D mesh from silhouette only, which is also our first stage. *Point cloud based* is our second stage using the results of the first stage as an initial value. ↓ means the smaller is better. *Mean*, *Root-Mean-Square (RMS)*, and *Error (mm)* same as Table 1, and *Cor. Error (mm)* represents the average of the per-vertex error with ground truth of the correspondence vertices in FAUST [Bogo et al. 2014].

Table 3. Evaluation of our TightNet

Method	SSIM↑	L1/L2↓	mask IoU↑
Baseline L2	62.27%	0.281	90.17%
OptCuts L1	43.91%	0.493	88.20%
Ours L1	67.24%	0.222	93.89%

↑ Means the Larger is Better, while ↓ Means the Smaller is Better.

OptCuts L1, respectively. We utilize the L1 norm and the **structural similarity (SSIM)** [Wang et al. 2004] for predicting the perceived quality of images, with window size (11 in our setting) to avoid the unreasonable effectiveness [Zhang et al. 2018]. To evaluate the garment mask prediction, we utilize the mask IoU with a threshold of 0.5, since the mask output of TightNet is 0 to 1 initially. As shown in Figure 4, the GI exhibits more distortions but has a much larger valid area, while OptCuts on the opposite. We find the size of the valid area is more critical during the experiment, and the GI performs better with the same resolution of the feature map. Specifically, as shown in Table 3, our TightNet with L1 loss and geometry image for 2D mapping achieves the highest accuracy, 67.24% for the task of tightness map prediction and 93.89% for the task of garment segmentation. This leads to more robust multi-layer reconstruction from only a single 3D scan as following.

Shape recovery evaluation. We evaluate our optimization-based shape recovery scheme by comparing it with the baseline variation using the straightforward solution in Equation (17). As shown in Figure 12, the variation suffers from inferior reconstruction results, especially in the local regions around the outer and crotch. In contrast, our shape recovery scheme successfully compresses the tightness field error so as to provide accurate and visually pleasant body shape reconstruction results.

6.3 Comparisons of Body Recovery

In this subsection, we demonstrate the overall performance of the proposed approach by comparing it against other state-of-the-art mesh-based and image-based body recovery methods, both qualitatively and quantitatively.

For mesh-based comparison, we compare to the state-of-the-art approaches, including the one proposed by Zhang et al. [2017] and the volumetric optimization stage of DoubleFusion [Yu et al. 2018b]. The former one utilizes a sequence of dressed mesh as input to recover body shape while the latter DoubleFusion [Yu

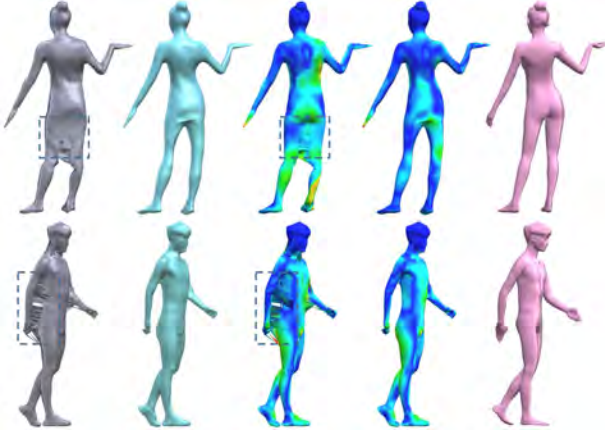


Fig. 12. Evaluation of our shape recovery scheme. From left to right: The recovered body before (gray) and after (green) our shape optimization in Equation (18); the corresponding per-vertex errors that are color-coded from blue to red; the ground-truth bodies (red).

Table 4. Comparison of Recovered Body Shapes in CTD with HMR [Kanazawa et al. 2018], SMPL-X [Pavlakos et al. 2019], CAPE [Ma et al. 2020], and DoubleFusion [Yu et al. 2018b]

Input	Method	Mean↓	RMS↓	Error (mm)↓	Front IoU↑	Side IoU↑
Image	HMR	1.607%	2.195%	48.21	74.70%	67.51%
	SMPL-X	1.478%	2.121%	44.34	78.76%	60.54%
A Mesh	DoubleFusion	0.804%	0.996%	24.12	82.49%	77.29%
	CAPE	0.584%	0.713%	17.52	88.20%	90.50%
	Our	0.191%	0.451%	5.73	90.12%	94.29%

Mean, Root-Mean-Square (RMS), and Error (mm) are same as Tab. 1. *Front IoU* represents the mean IoU of each projected mask pairs (estimated body and ground-truth body) from the view of input image for HMR [Kanazawa et al. 2018] and SMPL-X [Pavlakos et al. 2019]. *Side IoU* uses same metric like *Front IoU* but projected from the sideview.

et al. 2018b] optimizes the body shape from a single volumetric mesh input. For the image-based comparison, we compare to HMR [Kanazawa et al. 2018] and SMPL-X [Pavlakos et al. 2019], which regress the human model directly from only a single RGB image input. As shown in Figure 13, we achieve a comparable result against the sequence-based method of Zhang et al. [2017] on the BUFF dataset. Note that our method only uses a single 3D scan as input rather than a dynamic sequence of human models, which is hard to obtain for daily applications. Besides, our network is only pre-trained using our dataset CTD without fine-tuning on BUFF, which demonstrates the generation ability for our approach to recover both human body shape and garments from only a single 3D scan.

Then, we utilize our dataset CTD with ground-truth annotations for further qualitative and quantitative comparisons.

In Figure 14, we provide a qualitative comparison to DoubleFusion [Yu et al. 2018b] and HMR [Kanazawa et al. 2018] on three challenging cases with various clothing tightness and similar postures to get rid of the posture ambiguity. For a fair comparison, we also fine-tune HMR with the provided images and pose/shape parameters from optimized SMPL with ground-truth body shape. However, DoubleFusion [Yu et al. 2018b] suffers from inferior

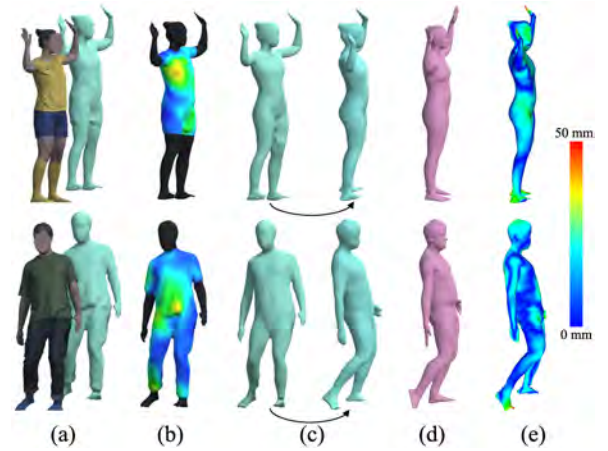


Fig. 13. Comparison with our recovered body shapes and Zhang et al. [2017] in BUFF Dataset. (a) Input Meshes and aligned meshes. (b) Predicted tightness field on meshes. (c) The recovered body shape of our results. (d) The ground truth of Zhang et al. [2017]. (e) Our results with the per-vertex body shape error colored from blue (good) to red (bad).

shape recovery and turns to estimate a fatter human body without considering the clothing tightness, while HMR [Kanazawa et al. 2018] suffers from scale ambiguity to provide only visually pleasant rather than the accurate human body shape, which is the inherent issue of such image-based methods. In contrast, our approach accurately reconstructs the human body shape by modeling various clothing tightness field effectively.

The quantitative comparison on the CTD dataset against both the mesh-based and image-based methods is provided in Table 4, in terms of the *Mean* and *RMS* from Metro [Cignoni et al. 1998] as well as the per-vertex error *Error (mm)*. Besides, we calculate the mean IoU of each projected mask pairs (estimated body and ground-truth body) from the view of rendered image for image-based methods, denoted as *Front IoU* in the CTD dataset. We also utilize *Side IoU* with the same operation as *Front IoU* projected from a side-view for thorough analysis. We also propose the comparison with CAPE [Ma et al. 2020], a representation based on graph convolutional network. We fine-tune the network of CAPE [Ma et al. 2020] with aligned clothed/unclothed mesh as input/output on the CTD dataset, to utilize the GCN for body shape recovery. With the same input of a single mesh, our proposed approach utilizes higher resolution on body shape, thus produces a 1- to 2-cm decrease for the error of body shape estimation, compared with this GCN-based approach. As shown in Table 4, our approach achieves significantly much more accurate body recovery results in terms of all the metrics above, with the aid of modeling the influence of clothing tightness.

For the comparison on the BUFF dataset [Zhang et al. 2017] in Table 5, our approach also achieves accurate body recovery results with only a single mesh input rather than dynamic mesh sequence. Compared to the detailed [Zhang et al. 2017], our approach only loses 1 to 2 mm but can be a more feasible approach on mobile devices. Besides, these comparisons against different inputs (images, single mesh, and mesh sequence) demonstrates that our approach not only produces the highly accurate shape but also

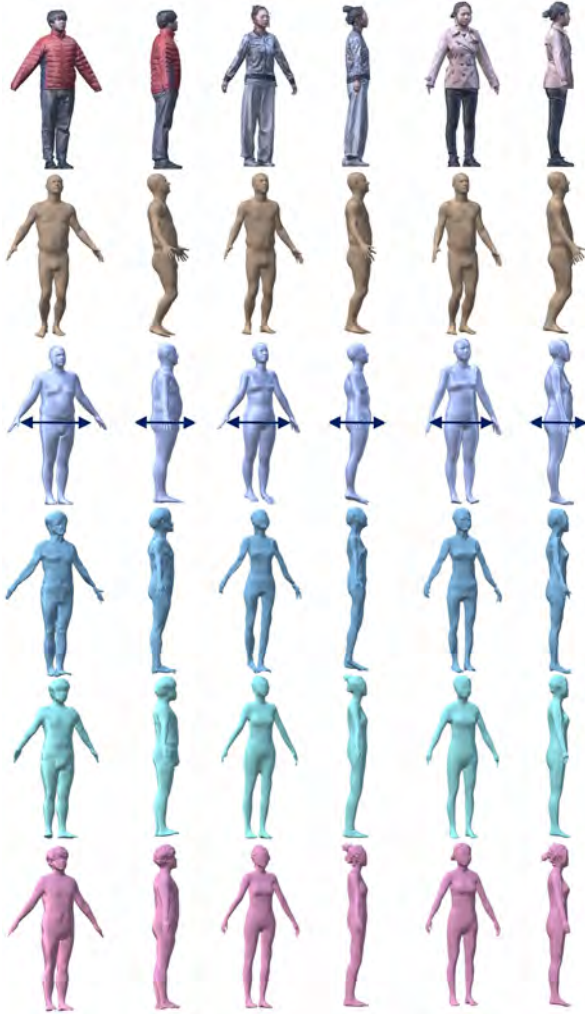


Fig. 14. The comparison of recovered body shapes in the CTD Dataset. From top to down: The input mesh; the recovered bodies using image-based method HMR [Kanazawa et al. 2018], mesh-based method DoubleFusion [Yu et al. 2018b], GCN-based method CAPE [Ma et al. 2020] and our method; the ground-truth bodies. The black arrows indicate the results of DoubleFusion are always fatter with the influence of clothing.

needs a single 3D mesh, which can be more feasible for mobile devices.

6.4 Application

In this subsection, based on our high-quality multi-layer reconstruction, we further demonstrate various interesting applications of our approach, including immersive cloth retargeting and clothed avatar animation.

Cloth retargeting. Recall that in our approach, thanks to our novel clothing tightness fields formulation, both the aligned human template to the dressed scan in Section 4.3 and the recovered body shape in Section 5.3 share the same mesh topology and the rigged skeleton as our enhanced human template in Section 4.1. Note that such displacements between the aligned template and its

Table 5. Comparison of Recovered Body Shapes in BUFF Dataset [Zhang et al. 2017] with HMR [Kanazawa et al. 2018], SMPL-X [Pavlakos et al. 2019], DoubleFusion [Yu et al. 2018b], Yang et al. [2016], and detailed [Zhang et al. 2017]

Input	Method	00005	00114	Avg.
		T-shirt, Pants	Soccer Outfit	
Image	HMR	75.08	44.69	59.89
	SMPL-X	177.29	122.58	149.94
Mesh Seq	Yang et al. [2016]	17.29	16.40	16.85
	detailed	2.52	2.23	2.38
A Mesh	DoubleFusion	32.57	24.68	28.63
	Our	4.73	3.24	3.98

We use *Root-Mean-Square (RMS)*, the same metric in the detailed [Zhang et al. 2017].

recovered body shape are our predicted clothing tightness. Thus, we can directly transfer the clothing to various recovered body shapes in terms of cloth-to-body and body-to-body displacements. As shown in Figure 15, we can achieve highly immersive cloth retargeting, and even fit our enhanced human template back to the original SMPL [Loper et al. 2015] so as to transfer the clothing to shape-variant SMPL model directly.

Clothed avatar animation. Benefiting from our enhanced human template with rigged skeleton and the novel clothing tightness field formulation, we are able to reconstruct a consistent multi-layer avatar from the input 3D scan to infer the body shape and the various segmented garments, such as the more results in Figure 17 from other datasets, CAPE [Ma et al. 2020] and MGN [Bhatnagar et al. 2019a]. Thus, we can not only change the garments of the current human target by using various clothing tightness, but also further animate the dressed avatar naturally by driving its inner body with various postures and maintaining current clothing tightness. As shown in Figure 16 and Figure 17, we can achieve the clothed avatar animation with the rigged skeleton and support the physical simulation engine from Unity3D to generate realistic animation with our multi-layer avatars. Compared with the animated results from ARCH [Huang et al. 2020] in Figure 17, although the effect on body movement is similar, our multi-layer avatars can achieve more realistic movement like the fluttering dress.

6.5 Limitations and Discussion

Though our approach is effective for body and garment reconstruction from only a single 3D scan, it still owns limitations as follows.

First, our scheme still cannot handle extreme human poses with a merged topology such as crossing legs/arms, hidden hands or curling up, or very low-quality scans (see Figure 18(b)). For those severely occluded regions such as a hand in the pocket (see Figure 18(a)), our method cannot accurately deform the template as they are under-observed, which could also exist artifacts for reposing. We plan to address this by incorporating extra data-driven human hand and face key-point priors to provide a good initialization.

Although our main purpose is body shape recovery and cloth modeling, the aligned region around the face and hand still needs improvement, as shown in Figure 18. Without semantic processing on specific body regions, our approach still can recover fine small detail of human body parts. Thus, we plan to implement the mesh



Fig. 15. The application of cloth retargeting. From left to right: the input MVS meshes; the estimated body shape and segmented garments; the retargeted clothing with slim body and fat body, respectively. Note that our enhanced template models are in green while the original SMPL [Loper et al. 2015] models are in gray.



Fig. 16. The application of clothed avatar animation. From left to right: the scans and our recovered inner body shape; our cloth retargeting results; our multi-layer avatar animation results into various human postures.

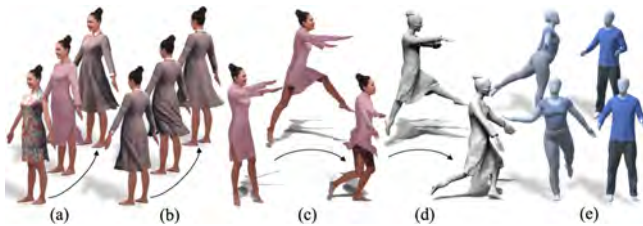


Fig. 17. More results and the simulated clothes. (a) The transferred avatars with various clothes. (b) The fluttering dress with a physical simulation engine. (c) The performance from our multi-layer avatar. (d) The compared performance from the avatar of ARCH Huang et al. [2020]. (e) More results from the CAPE [Ma et al. 2020] and MGN [Bhatnagar et al. 2019b] dataset.

processing strategy with the semantic body part for both alignment and de-noise as one of our future works. Besides, currently utilized geometry images in our approach can only handle genus 0 human geometry. In reality, the human model can have a very complex topology, and a more sophisticated geometry image generation approach is required. Alignment schemes that can handle these topologically complex human models are also our immediate future work.

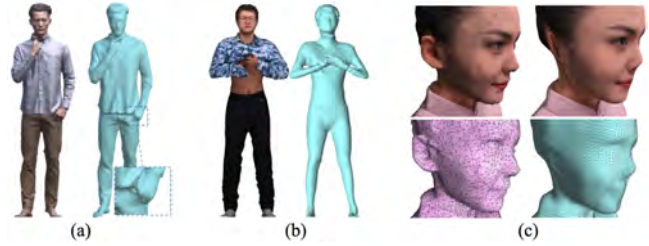


Fig. 18. The failure cases of our approach. (a) The clothed human scan with the left hand in the pocket and the aligned result. (b) The clothed human scan with crossing arms, and the recovered body. (c) Left column: The face of the textured scan and its mesh from the MVS approach. Right column: The face of the baked aligned mesh and its mesh from our approach.

For specific style and of garments, like evening dress, the proposed TightNet is still hard to generate correct results with the influence of the garment style from the dataset. Although the TightNet is based on most effective image-to-image translation architecture, the existing seam during the mapping of geometry image still might cause the inconsistency around the boundary. Thus, we use body shape recovery function to smooth between these regions.

For our applications, although we can achieve both skeleton-driven and physical simulated clothed avatar animation in Figure 16 and Figure 17, our current clothing tightness field formulation still cannot simulate the dynamic movement of clothing in a physically plausible manner for our reconstructed garments. It is a promising direction to further model the clothing tightness field for 4D dynamic sequences with the aid of extra-human motion priors like CAPE [Ma et al. 2020] and Zhou et al. [2020]. Moreover, our approach relies on raw 3D human scans, which are usually difficult to obtain, and the quality cannot be guaranteed. Hence we plan to explore the possibility of directly taking a single or sparse set of 2D images [Chen et al. 2018] as the input of the MVS setting for recovering 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 a better AR/VR or try-on experience.

7 CONCLUSION

We present TightCap, a learning-based scheme for robustly and accurately capturing the clothing tightness field as well as human geometry with 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 enables the geometry image correspondence from various types of clothing. Moreover, we collect a large 3D CTD for the clothed human reconstruction tasks. We propose and train a modified conditional GAN network to automatically predict the clothing tightness map 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 animation. We believe our scheme will benefit various AR/VR research and applications, such as virtual try-on and avatar animation.

ACKNOWLEDGMENTS

The authors thank WenGuang Ma, YeCheng Qiu, and MingGuang Chen for help with data acquisition and Hongbo Wang, Gao Ya, Shenze Ye, and Teng Su for help with data annotation.

REFERENCES

- Thiemo Alldieck, Marcus Magnor, Bharat Lal Bhatnagar, Christian Theobalt, and Gerard Pons-Moll. 2019a. Learning to reconstruct people in clothing from a single RGB camera. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'19)*.
- Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. 2018. Video based reconstruction of 3d people models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 8387–8397.
- Thiemo Alldieck, Gerard Pons-Moll, Christian Theobalt, and Marcus Magnor. 2019b. Tex2Shape: Detailed full human body geometry from a single image. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV'19)*. IEEE.
- Riza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. 2018. Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7297–7306.
- Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. 2005. SCAPE: Shape completion and animation of people. *ACM Trans. Graph.* 24, 3 (2005), 408–416.
- Simon Baker, Takeo Kanade, et al. 2005. Shape-from-silhouette across time part ii: Applications to human modeling and markerless motion tracking. *Int. J. Comput. Vis.* 63, 3 (2005), 225–245.
- Alexandru O. Bălan and Michael J. Black. 2008. The naked truth: Estimating body shape under clothing. In *Proceedings of the European Conference on Computer Vision*. Springer, 15–29.
- Bharat Lal Bhatnagar, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. 2020. Combining implicit function learning and parametric models for 3D human reconstruction. In *Proceedings of the European Conference on Computer Vision (ECCV'20)*. Springer.
- Bharat Lal Bhatnagar, Garvita Tiwari, Christian Theobalt, and Gerard Pons-Moll. 2019a. Multi-garment net: Learning to dress 3d people from images. In *Proceedings of the IEEE International Conference on Computer Vision*. 5420–5430.
- Bharat Lal Bhatnagar, Garvita Tiwari, Christian Theobalt, and Gerard Pons-Moll. 2019b. Multi-garment net: Learning to dress 3D people from images. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV'19)*. IEEE.
- Federica Bogo, Michael J. Black, Matthew Loper, and Javier Romero. 2015. Detailed full-body reconstructions of moving people from monocular RGB-D sequences. In *Proceedings of the IEEE International Conference on Computer Vision*. 2300–2308.
- Federica Bogo, Javier Romero, Matthew Loper, and Michael J. Black. 2014. FAUST: Dataset and evaluation for 3D mesh registration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'14)*. IEEE, Los Alamitos, CA.
- Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2018. OpenPose: Realtime multi-person 2D pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2017. Realtime multi-person 2D pose estimation using part affinity fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
- Xin Chen, Yuwei Li, Xi Luo, Tianjia Shao, Jingyi Yu, Kun Zhou, and Youyi Zheng. 2018. Autosweep: Recovering 3d editable objects from a single photograph. *IEEE Trans. Visual. Comput. Graph.* 26, 3 (2018), 1466–1475.
- Xin Chen, Anqi Pang, Wei Yang, Yuxin Ma, Lan Xu, and Jingyi Yu. 2021. SportsCap: Monocular 3D human motion capture and fine-grained understanding in challenging sports videos. *International Journal of Computer Vision*. DOI: <https://doi.org/10.1007/s11263-021-01486-4>
- German K. M. Cheung, Simon Baker, and Takeo Kanade. 2003a. Visual hull alignment and refinement across time: A 3d reconstruction algorithm combining shape-from-silhouette with stereo. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Vol. 2. IEEE, II–375.
- K. M. G. Cheung, Simon Baker, and Takeo Kanade. 2003b. Shape-from-silhouette of articulated objects and its use for human body kinematics estimation and motion capture. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Vol. 1. IEEE, I–1.
- Paolo Cignoni, Claudio Rocchini, and Roberto Scopigno. 1998. Metro: Measuring error on simplified surfaces. In *Computer Graphics Forum*, Vol. 17. Wiley Online Library, 167–174.
- Alvaro Collet, Ming Chuang, Pat Sweeney, Don Gillett, Dennis Eyseev, David Calabrese, Hugues Hoppe, Adam Kirk, and Steve Sullivan. 2015. High-quality streamable free-viewpoint video. *ACM Trans. Graph.* 34, 4 (2015), 69.
- Stefano Corazza, Lars Muendemann, A. M. Chaudhari, T. Demattio, Claudio Cobelli, and Thomas P. Andriacchi. 2006. A markerless motion capture system to study musculoskeletal biomechanics: Visual hull and simulated annealing approach. *Ann. Biomed. Eng.* 34, 6 (2006), 1019–1029.
- Mingsong Dou, Sameh Khamis, Yury Degtyarev, Philip Davidson, Sean Ryan Fanello, Adarsh Kowdle, Sergio Orts Escolano, Christoph Rhemann, David Kim, Jonathan Taylor, et al. 2016. Fusion4d: Real-time performance capture of challenging scenes. *ACM Trans. Graph.* 35, 4 (2016), 114.
- Yasutaka Furukawa, Carlos Hernández, et al. 2013. Multi-view stereo: A tutorial. *Found. Trends Comput. Graph. Vis.* 9, 1-2 (2013), 1–148.
- Xianfeng Gu, Steven J. Gortler, and Hugues Hoppe. 2002. Geometry images. *ACM Trans. Graph.* 21, 3 (2002), 355–361.
- Nils Hasler, Carsten Stoll, Bodo Rosenhahn, Thorsten Thormählen, and Hans-Peter Seidel. 2009. Estimating body shape of dressed humans. *Comput. Graph.* 33, 3 (2009), 211–216.
- Yannan He, Anqi Pang, Xin Chen, Han Liang, Minye Wu, Yuxin Ma, and Lan Xu. 2021. ChallengeCap: Monocular 3D capture of challenging human performances using multi-modal references. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'21)*.
- Zeng Huang, Yuanlu Xu, Christoph Lassner, Hao Li, and Tony Tung. 2020. ARCH: Animatable reconstruction of clothed humans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3093–3102.
- Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. 2017. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1125–1134.
- Boyi Jiang, Juyong Zhang, Yang Hong, Jinhao Luo, Ligang Liu, and Hujun Bao. 2020. BCNet: Learning body and cloth shape from a single image. In *European Conference on Computer Vision*. Springer.
- Hanbyul Joo, Tomas Simon, Xulong Li, Hao Liu, Lei Tan, Lin Gui, Sean Banerjee, Timothy Godisart, Bart Nabbe, Iain Matthews, et al. 2019. Panoptic studio: A massively multiview system for social interaction capture. *IEEE Trans. Pattern Anal. Mach. Intell.* 41, 1 (2019), 190–204.
- Hanbyul Joo, Tomas Simon, and Yaser Sheikh. 2018. Total capture: A 3D deformation model for tracking faces, hands, and bodies. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 8320–8329.
- Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. 2018. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7122–7131.
- Zorah Lahner, Daniel Cremers, and Tony Tung. 2018. DeepWrinkles: Accurate and realistic clothing modeling. In *Proceedings of the European Conference on Computer Vision*. 667–684.
- Christoph Lassner, Gerard Pons-Moll, and Peter V. Gehler. 2017a. A generative model of people in clothing. In *Proceedings of the IEEE International Conference on Computer Vision*. 853–862.
- Christoph Lassner, Javier Romero, Martin Kiefel, Federica Bogo, Michael J. Black, and Peter V. Gehler. 2017b. Unite the people: Closing the loop between 3D and 2D human representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6050–6059.
- Verica Lazova, Eldar Insafutdinov, and Gerard Pons-Moll. 2019. 360-degree textures of people in clothing from a single image. In *Proceedings of the International Conference on 3D Vision (3DV'19)*. IEEE, 643–653.
- Minchen Li, Danny M. Kaufman, Vladimir G. Kim, Justin Solomon, and Alla Sheffer. 2018. OptCuts: Joint optimization of surface cuts and parameterization. *ACM Trans. Graph.* 37, 6 (2018). <https://doi.org/10.1145/3272127.3275042>
- Zhong Li, Xin Chen, Wangyiteng Zhou, Yingliang Zhang, and Jingyi Yu. 2019. Pose2Body: Pose-guided human parts segmentation. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME'19)*. IEEE, 640–645.
- Zhe Li, Tao Yu, Chuanyu Pan, Zerong Zheng, and Yebin Liu. 2020. Robust 3D self-portraits in seconds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1344–1353.
- Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. 2015. SMPL: A skinned multi-person linear model. *ACM Trans. Graph.* 34, 6 (2015), 248.
- Qianli Ma, Jinlong Yang, Anurag Ranjan, Sergi Pujades, Gerard Pons-Moll, Siyu Tang, and Michael Black. 2020. Learning to dress 3D people in generative clothing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'20)*. IEEE.
- Maxim Mikhnevich and Patrick Hebert. 2011. Shape from silhouette under varying lighting and multi-viewpoints. In *Proceedings of the Canadian Conference on Computer and Robot Vision*. IEEE, 285–292.
- Ryota Natsume, Shunsuke Saito, Zeng Huang, Weikai Chen, Chongyang Ma, Hao Li, and Shigeo Morishima. 2019. Siclope: Silhouette-based clothed people. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4480–4490.
- Alexandros Neophytou and Adrian Hilton. 2014. A layered model of human body and garment deformation. In *Proceedings of the 2nd International Conference on 3D Vision*, Vol. 1. IEEE, 171–178.
- Richard A. Newcombe, Dieter Fox, and Steven M. Seitz. 2015. Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 343–352.
- Richard A. Newcombe, Shahram Izadi, Otmar Hilliges, David Molyneaux, David Kim, Andrew J. Davison, Pushmeet Kohi, Jamie Shotton, Steve Hodges, and Andrew

- Fitzgibbon. 2011a. KinectFusion: Real-time dense surface mapping and tracking. In *Proceedings of the IEEE International Symposium on Mixed and Augmented Reality*. IEEE, 127–136.
- Richard A. Newcombe, Steven J. Lovegrove, and Andrew J. Davison. 2011b. DTAM: Dense tracking and mapping in real-time. In *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, 2320–2327.
- Alejandro Newell, Kaiyu Yang, and Jia Deng. 2016. Stacked hourglass networks for human pose estimation. In *Proceedings of the European Conference on Computer Vision*. Springer, 483–499.
- Anqi Pang, Xin Chen, Haimin Luo, Minye Wu, Jingyi Yu, and Lan Xu. 2021. Few-shot neural human performance rendering from sparse RGBD videos. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI'21)*. International Joint Conferences on Artificial Intelligence Organization, 938–944. DOI: [10.24963/ijcai.2021/130](https://doi.org/10.24963/ijcai.2021/130)
- Chaitanya Patel, Zhouyingcheng Liao, and Gerard Pons-Moll. 2020. TailorNet: Predicting clothing in 3D as a function of human pose, shape and garment style. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'20)*. IEEE.
- Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. 2019. Expressive body capture: 3D hands, face, and body from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'19)*.
- Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. 2018. 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*. 459–468.
- Leonid Pishchulin, Eldar Insafutdinov, Siyu Tang, Bjoern Andres, Mykhaylo Andriluka, Peter V. Gehler, and Bernt Schiele. 2016. Deepcut: Joint subset partition and labeling for multi person pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4929–4937.
- Gerard Pons-Moll, Sergi Pujades, Sonny Hu, and Michael J. Black. 2017. ClothCap: Seamless 4D clothing capture and retargeting. *ACM Trans. Graph.* 36, 4 (2017), 73.
- Albert Pumarola, Jordi Sanchez, Gary Choi, Alberto Sanfeliu, and Francesc Moreno-Noguer. 2019. 3DPeople: Modeling the geometry of dressed humans. In *Proceedings of the International Conference on Computer Vision (ICCV'19)*.
- Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Proceedings of the International Conference on Medical Image Computing and Computer-assisted Intervention*. Springer, 234–241.
- Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. 2019. PIFu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *Proceedings of the IEEE International Conference on Computer Vision*. 2304–2314.
- Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. 2020. PIFuHD: Multi-level pixel-aligned implicit function for high-resolution 3D human digitization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 84–93.
- Johannes L. Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. 2016. Pixelwise view selection for unstructured multi-view stereo. In *Proceedings of the European Conference on Computer Vision*. Springer, 501–518.
- Tomas Simon, Hanbyul Joo, Iain Matthews, and Yaser Sheikh. 2017. Hand keypoint detection in single images using multiview bootstrapping. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
- Olga Sorkine and Marc Alexa. 2007. As-rigid-as-possible surface modeling. In *Proceedings of the Symposium on Geometry Processing*, Vol. 4. 109–116.
- Christoph Strecha, Wolfgang Von Hansen, Luc Van Gool, Pascal Fua, and Ulrich Thoennessen. 2008. On benchmarking camera calibration and multi-view stereo for high resolution imagery. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 1–8.
- Zhuo Su, Lan Xu, Zerong Zheng, Tao Yu, Yebin Liu, and Lu Fang. 2020. Robustfusion: Human volumetric capture with data-driven visual cues using a rgbd camera. In *Proceedings of the 16th European Conference on Computer Vision (ECCV'20)*. Springer, 246–264.
- Robert W. Sumner, Johannes Schmid, and Mark Pauly. 2007. Embedded deformation for shape manipulation. *ACM Trans. Graph.* 26, 3 (2007), 80.
- Xin Suo, Yuheng Jiang, Pei Lin, Yingliang Zhang, Minye Wu, Kaiwen Guo, and Lan Xu. 2021. NeuralHumanFVV: Real-time neural volumetric human performance rendering using RGB cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6226–6237.
- Garvita Tiwari, Bharat Lal Bhatnagar, Tony Tung, and Gerard Pons-Moll. 2020. SIZER: A dataset and model for parsing 3D clothing and learning size sensitive 3D clothing. In *Proceedings of the European Conference on Computer Vision (ECCV'20)*. Springer.
- Jing Tong, Jin Zhou, Ligang Liu, Zhigeng Pan, and Hao Yan. 2012. Scanning 3d full human bodies using kinects. *IEEE Trans. Visual. Comput. Graph.* 18, 4 (2012), 643–650.
- Bill Triggs, Philip F. McLauchlan, Richard I. Hartley, and Andrew W. Fitzgibbon. 1999. Bundle adjustment: A modern synthesis. In *Proceedings of the International Workshop on Vision Algorithms*. Springer, 298–372.
- Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.* 13, 4 (2004), 600–612.
- Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. 2016. Convolutional pose machines. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4724–4732.
- Stefanie Wuhrer, Leonid Pishchulin, Alan Brunton, Chang Shu, and Jochen Lang. 2014. Estimation of human body shape and posture under clothing. *Comput. Vis. Image Understand.* 127 (2014), 31–42.
- Donglai Xiang, Hanbyul Joo, and Yaser Sheikh. 2019. Monocular total capture: Poseing face, body, and hands in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
- Lan Xu, Zhuo Su, Lei Han, Tao Yu, Yebin Liu, and Lu Fang. 2019a. UnstructuredFusion: Realtime 4D geometry and texture reconstruction using commercial RGBD cameras. *IEEE Trans. Pattern Anal. Mach. Intell.* 42, 10 (2019), 2508–2522.
- Weipeng Xu, Avishek Chatterjee, Michael Zollhoefer, Helge Rhodin, Dushyant Mehta, Hans-Peter Seidel, and Christian Theobalt. 2018. Monoperfcap: Human performance capture from monocular video. *ACM Trans. Graph.* 37, 2 (2018), 27.
- Yuanlu Xu, Song-Chun Zhu, and Tony Tung. 2019b. DenseRaC: Joint 3D pose and shape estimation by dense render-and-compare. In *Proceedings of the IEEE International Conference on Computer Vision*. 7760–7770.
- Jinlong Yang, Jean-Sébastien Franco, Franck Hétroy-Wheeler, and Stefanie Wuhrer. 2016. Estimation of human body shape in motion with wide clothing. In *Proceedings of the European Conference on Computer Vision*. Springer, 439–454.
- Jinlong Yang, Jean-Sébastien Franco, Franck Hétroy-Wheeler, and Stefanie Wuhrer. 2018. Analyzing clothing layer deformation statistics of 3D human motions. In *Proceedings of the European Conference on Computer Vision*. 237–253.
- Tao Yu, Kaiwen Guo, Feng Xu, Yuan Dong, Zhaoqi Su, Jianhui Zhao, Jianguo Li, Qionghai Dai, and Yebin Liu. 2017. 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*. 910–919.
- Tao Yu, Zerong Zheng, Kaiwen Guo, Jianhui Zhao, Qionghai Dai, Hao Li, Gerard Pons-Moll, and Yebin Liu. 2018a. 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*. 7287–7296.
- Tao Yu, Zerong Zheng, Kaiwen Guo, Jianhui Zhao, Qionghai Dai, Hao Li, Gerard Pons-Moll, and Yebin Liu. 2018b. 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*. 7287–7296.
- Tao Yu, Zerong Zheng, Yuan Zhong, Jianhui Zhao, Qionghai Dai, Gerard Pons-Moll, and Yebin Liu. 2019. Simulcap: Single-view human performance capture with cloth simulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'19)*. IEEE, 5499–5509.
- Chao Zhang, Sergi Pujades, Michael J. Black, and Gerard Pons-Moll. 2017. Detailed, accurate, human shape estimation from clothed 3D scan sequences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4191–4200.
- Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 586–595.
- Zerong Zheng, Tao Yu, Yixuan Wei, Qionghai Dai, and Yebin Liu. 2019. Deephuman: 3d human reconstruction from a single image. In *Proceedings of the IEEE International Conference on Computer Vision*. 7739–7749.
- Yi Zhou, Chenglei Wu, Zimo Li, Chen Cao, Yuting Ye, Jason M. Saragih, Hao Li, and Yaser Sheikh. 2020. Fully convolutional mesh autoencoder using efficient spatially varying kernels. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS'20)*, Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (Eds.).

Received June 2020; revised June 2021; accepted July 2021