DeepCloth: Neural Garment Representation for Shape and Style Editing

Zhaoqi Su, Tao Yu, Yangang Wang, *Member, IEEE*, and Yebin Liu, *Member, IEEE*

Abstract—Garment representation, editing and animation are challenging topics in the area of computer vision and graphics. It remains difficult for existing garment representations to achieve smooth and plausible transitions between different shapes and topologies. In this work, we introduce, DeepCloth, a unified framework for garment representation, reconstruction, animation and editing. Our unified framework contains 3 components: First, we represent the garment geometry with a "topology-aware UV-position map", which allows for the unified description of various garments with different shapes and topologies by introducing an additional topology-aware UV-mask for the UV-position map. Second, to further enable garment reconstruction and editing, we contribute a method to embed the UV-based representations into a continuous feature space, which enables garment shape reconstruction and editing by optimization and control in the latent space, respectively. Finally, we propose a garment animation method by unifying our neural garment representation with body shape and pose, which achieves plausible garment animation results leveraging the dynamic information encoded by our shape and style representation, even under drastic garment editing operations. To conclude, with DeepCloth, we move a step forward in establishing a more flexible and general 3D garment digitization framework. Experiments demonstrate that our method can achieve state-of-the-art garment representation performance compared with previous methods.

Index Terms—Garment digitization, garment representation, 3D reconstruction and animation

1 Introduction

3D garment representation, modeling, editing and animation/simulation have numerous applications in clothing design, digital humans, and virtual try-on. Traditional high-fidelity 3D garment modeling and animation often rely on artist design or heavy simulation methods, such as physically based simulation [1], which consume enormous labor costs or computational resources. In recent years, neural garment representations based on deep learning techniques have achieved impressive garment modeling or animation results [2], [3], [4], [5], [6], [7]. However, the majority of these methods either focus on encoding garment dynamics for specific clothing (clothing-specific learning) or aim at 3D clothing recovery from images without any editing capacities. This is because establishing a unified framework for shape/style editable garment representation, reconstruction and animation remains challenging. Although the most recent neural garment modeling/animation work TailorNet [8] achieves impressive detailed clothing dynamics recovery for different human shapes and poses, it still defines garments on top of a predefined fixed template, in which the garment topology is fixed. Such a fixed representation limits its ability to achieve an ideal garment editing framework, e.g., enabling transition from long pants to shorts or from front-opening T-shirts to front-closing shirts. Concurrent work [9] by Corona *et al.* focuses more on establishing garment shape/style representations and less on shape-/style-dependent garment animation based on their styles, while our method further proposes a garment shape-dependent animation module, which shows more dynamics while performing animation with different garment styles.

In this paper, we argue that it is essential to learn a compact and uniform space for garments with different shapes and styles, which will form a unified garment representation framework, and then be further used for garment reconstruction and animation. Such a representation should enable free and smooth style transitions between different garment shapes and styles, even for garments with different topologies, e.g., from front-opening clothes to front-closing clothes. By transferring the garment representation into a neural feature space, and mapping the 3D scanned garment mesh into the same feature space, such a representation can also be used to perform garment animation and 3D garment shape editing using deep neural networks as demonstrated in Fig. 1. However, fulfilling such a representation is challenging due to the large topological changes and nonuniform latent space encoding.

In this paper, we propose DeepCloth, a unified framework for garment representation, reconstruction, animation and editing by assembling different garment shapes and topologies into a unified representation framework. Technically, we propose a "topology-aware UV-position map" representation to encode both the topologies and the

- *Zhaoqi Su, Tao Yu, and Yebin Liu are with Tsinghua University, Beijing 100190, China. E-mail: suzq13@tsinghua.org.cn, {ytrock, liuyebin}@mail.tsinghua.edu.cn.*
- *Yangang Wang is with Southeast University, Nanjing 211189, China. E-mail: yangangwang@seu.edu.cn.*

Manuscript received 7 July 2021; revised 9 Feb. 2022; accepted 10 Apr. 2022. Date of publication 0 . 2022; date of current version 0 . 2022.
This work was supported by NSFC under Grants 62125107 and 62171255, in part by the National Key R&D Program of China under Grant 2021ZD0113503 and in part by China Postdoctoral Science Foundation under Grant 2020M670340.
(Corresponding author: Yebin Liu.)
Recommended for acceptance by H. Li.
Digital Object Identifier no. 10.1109/TPAMI.2022.3168569

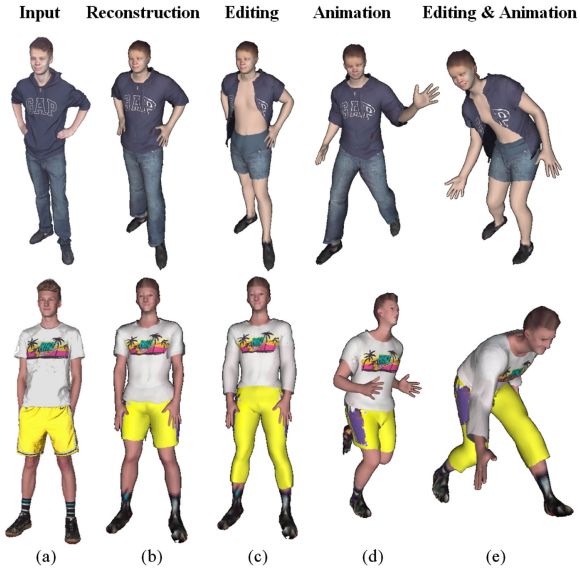


Fig. 1. Our neural garment representation framework, DeepCloth, enables garment reconstruction (b), shape editing (c) (from close to open, from short to long, from loose to tight, etc.) and animation (d,e) (with garment-specific dynamics even after significant topological editing of the garment) given a 3D scan with arbitrary pose (a).

geometric details of garments. To achieve plausible garment transitions under different topologies, we transform the UV binary mask into a continuous distance map. By encoding the UV-position map with a transformed mask into a feature space using our proposed ParamNet, we can represent garments with different shape styles and topologies in a unified network and thus achieve flexible garment editing and continuous shape and style transitions between different garments by feature interpolation and decoding. As the "UV map" is widely used to represent 3D shapes (such as Tex2shape [10] for human modeling), we believe that our introduction of the "topology-aware-mask" for UV-map-representation, as well as the demonstration of neural shape editing capacity, may inspire future 3D mesh/shape/texture editing studies related to the use of UV maps.

With our deep learning network trained on the large-scale synthetic dataset of 3D clothed human sequences with various garment styles, i.e., CLOTH3D [11], we can parameterize clothing shape variations of front-opening T-shirts, T-shirts, shirts, pants, skirts, dresses, and jumpsuits. Additionally, with our proposed animation module AnimNet and 3D-shape inference module 3DInferNet, DeepCloth can generate 4D sequences of garment dynamics (see Figs. 1d and 1e) or extract the clothing shape parameters from a clothed human model under arbitrary poses (see Fig. 1b), which enhances its ability for 3D clothing shape editing (see Fig. 1c). The main contributions of this work are summarized as follows:

- We propose a unified garment modeling framework based on a UV-mask garment representation, which further enables garment reconstruction, animation and editing.
- We propose a topology-aware continuous UV-mask neural garment representation and encode such representation into a unified continuous feature space, which enables joint learning of both the 3D position and the topology of the garments, and neural control of garment shape and topologies. (Sections 4 and 5)
- By mapping the garment 3D information onto the garment feature space, or unifying the proposed garment representation with human shape and pose information, we can perform garment shape reconstruction and animation based on our neural garment representation, which can generate plausible garment dynamics even under drastic garment editing operations. (Sections 6 and 7)

2 Related Work

There are numerous works on garment representation, animation and reconstruction. Here, we mainly review the works that are most related to our approach.

Garment Representation and Animation. There are essentially three approaches for garment animation: physics-based simulation (PBS), data-driven methods, and animation based on capture.

For physics-based simulation (PBS), traditional physics based garment simulation formulates the garment as a mass-spring system with force-based simulation [1], [12], [13] or other physical models based on the finite element method [14], [15], with the explicit Euler method [16] or implicit/semi-implicit Euler method [16], [17], [18]. These methods can generate realistic clothing with vivid dynamics given a designed garment shape and garment template, but mostly incur considerable computational costs for numerous integration iterations for clothing dynamics, and cannot perform a more general shape control of the garment.

Data-driven methods aim to shorten the computational time for garment animation with a more flexible garment representation. Early methods such as [19], [20], [21] use a nearest neighbor search or linear regression to animate clothing on the human body with different poses and shapes. Recent works have mainly adopted deep learning methods to perform garment animation. [22] learns a shared space for garment style variation, and can predict garment shape from a user sketch with a fixed pose. [23], [24] regress the garment shape with various human poses and shapes with MLP or RNN methods. [3], [4], [25] propose garment animation by 3D garment draping or SMPL-based garment deformation, using a graph convolution network to obtain garment shapes worn on a human model with different shapes and poses. [2], [5] propose pixel-based garment 2D representation based on texture mapping on a human model or on a template-based texture space, which is similar to our representation method, but they cannot generate the shape parameters of the garments. [26] leverages the human parsing of the image to mask the UV-map representations of garments and controls the garment shape by editing the masks. However, without compact encoding of the garment representation UV map and masks, it barely performs continuous garment style transition, and the garment animation can only be performed through skinning, without leveraging the masks to infer shape dynamics. Additionally, it cannot represent garments that are not homotopy to human models such as dresses. [7] can interpolate between different garment styles, but it can only interpolate the "sewing patterns", meaning that [7] only interpolates the

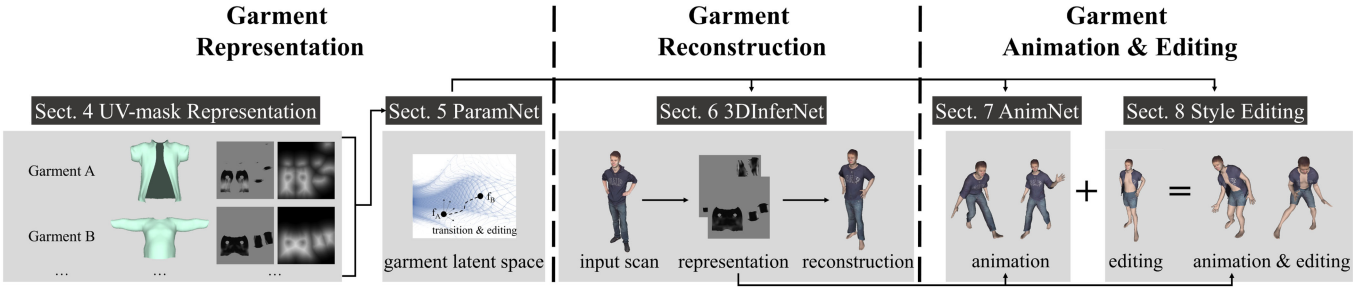


Fig. 2. The demonstration of our DeepCloth framework. From left to right: garment representation, garment reconstruction and garment animation module.

area of the body covered by the garments, without a general shape parametrization framework for generating more garment shapes. These methods use garment vertices, graph-based garment representation or pixel-based garment representation to perform garment animation but lack garment shape parametrization modules for flexibly and conveniently controlling the garment shape. In addition, in regard to computer vision tasks such as relighting on clothed humans, [27] proposes a novel approach for representing the lighting environment and view visibility in UV space, which leverages the flexibility of UV representations for performing realistic and high-quality relighting and view synthesis on real captured data of humans.

For garment animation based on real capture, this kind of method for garment animation focuses on recovering the static or dynamic garment shape from a given picture or video. [28], [29], [30] recover garment shapes from multiview stereo, [2], [31] recover garment dynamics from 4D sequences, [32], [33] propose a system for garment shape recovery from a single RGBD camera, while [34] proposes a method for extracting the garment template shape and recovering garment dynamics from a single RGB input. Recently, [35] propose a method for extracting multiple garments from several input images and dressing them on other human bodies, while [36], [37] propose a CNN-based method for recovering the human shape and pose. Such methods express the garment as a deformation of the subset of the human model, and it is difficult to express more types of garments like dresses and loosened front-opening clothing. Alldieck *et al.* [10] proposes a 2D texture-based human with a garment shape representation method for recovering the whole shape from a single image. [38], [39] propose monocular human performance capture methods that can recover human motion and garment geometry details, given monocular RGB video inputs. These methods mainly focus on recovering garment shapes from input images, without proposing a general garment shape representation framework.

Garment Shape Parametrization. Recently, a few works focus on establishing a garment shape parametrization framework. Shen *et al.* [7] demonstrates the garment style interpolation results, but it only controls the change of the covering area on the human body, without generating a general shape expression. Tiwari *et al.* [6] proposes a framework for parsing the 3D input to extract the garment shape and change the size of the garment, but in view of shape parametrization, it only controls one dimension of the garment shape. TailorNet [8] proposes a garment shape parametrization and animation framework; however, as mentioned in Section 1, with the "offset on template vertices" expression, it is difficult for TailorNet [8] to generalize to more types of garment topology, such as front-opening garments and long dresses. In addition, it shows limited ability to perform large garment shape changes, e.g., from long trousers to shorts or from long dresses to skirts. Additionally, compared to our DeepCloth, it has less capability for performing 3D garment shape inference and flexible 3D shape editing. Therefore, it does not meet the demand for establishing a general framework for garment representation enabling garment shape and style transition. Meanwhile, our DeepCloth proposes a general garment shape representation framework, which enables more general 3D garment reconstruction, animation and editing.

3 Overview

Our goal is to establish a unified framework for garment representation, reconstruction, animation and editing, as shown in Fig. 2. We first introduce the idea behind designing the whole framework. For a unified garment modeling model, traditional methods often rely on a fixed garment mesh template, which can hardly be applied to a general garment style and shape representation. To allow for a flexible and neural editable garment modeling framework, we propose our UV-mask garment representation. Together with our proposed CNN-based ParamNet, a unified and compact garment style space is established. Furthermore, to perform garment shape inference, animation and shape editing supporting various styles of garments, we propose different CNN- and PointNet-based networks, which establish mappings between different data domains, e.g., T-posed and animated garment shapes, or 3D mesh space and 2D garment UV space.

Methodically, we first propose a UV-position map with continuous mask representation, in which the mask denote the topology and covering areas of the garments, while the rendered texels on the UV map denotes the geometry details (see Section 4). Such a representation transfers the garment shape style and topology into a 2D UV-map, which is naturally suitable for the continuous transition between different garment shapes. Then, we perform UV map encoding by introducing ParamNet, which maps both the UV map and its mask information into a feature space by using a CNN-based encoder-decoder structure (see Section 5). By changing and interpolating the features in the feature space, garment shape transition and editing can be performed and can be applied to the following garment shape inference and animation module (see Sections 6 and 7). Specifically, given 3D garment scans, the garment inference module can reconstruct the

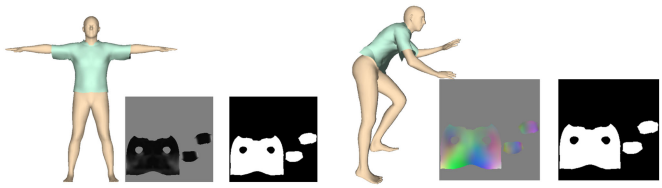


Fig. 3. The demonstration on coupling the UV-map representation of the garment. (a) 2D representation for the T-posed garment for Section 4. (b) 2D representation for the randomly posed garment for Section 7.

garment shape by transforming the point clouds to our neural garment representation (Section 6). With the garment shapes mapped onto the garment feature space, garment animation or shape editing can be applied to the reconstructed garments. (Sections 7 and 8). Note that our garment animation module is able to learn specific garment dynamic information according to different garment topologies, which leads to more plausible 4D garment animation results, even under drastic garment editing operations.

4 T-Posed Garment Representation

For T-posed garment representation, a continuous UV-mask garment representation is proposed to map the 3D garment mesh onto the continuous 2D UV space, as illustrated below. Such a representation naturally encodes the 3D garment geometry distribution on top of human bodies, without relying on a fixed template, therefore supporting style transition and editing among different garment topologies.

The first step of our DeepCloth is to represent garments with different shapes and topologies in a compact space. Therefore, different garment shape styles of a garment type, i.e., front-opening/front-closing T-shirts with long/short sleeves, can be mapped into the same feature space. Note that this section will only deal with the T-posed garment model for garment shape encoding and transition, while garments under arbitrary poses will be handled in Section 7.

In our representation framework, we regard a garment mesh as a geometric structure covering the human surface and then map such clothing to a standard human model UV map [40] which stores the garment topologies and normal distance from the human body. By mapping the 3D garment geometry onto the 2D SMPL UV space, we are able to establish the relationship between the garment vertices and the nearby vertices on the human model, and better represent the geometric features.

Specifically, our goal is to find the correspondences between the garment vertices and the human model UV coordinates. Here, we use the same UV map used in [40]. We emit rays from the T-posed SMPL surface that intersect the garment mesh with garment vertices, thus establishing a one-to-one mapping from garment vertices to SMPL surface points. We denote such an SMPL surface point as the corresponding sub-vertex $\vec{v}_i^T$ of garment vertex $\vec{g}_i^T$. In this way, with the predefined UV coordinates of each SMPL triangle face by [40], we can accordingly find the UV coordinate t_i of the sub-vertex $\vec{v}_i^T$, which serves as the corresponding UV coordinate for $\vec{g}_i^T$. After calculating the length from $\vec{v}_i^T$ to $\vec{g}_i^T$, we set the length value as the rendered texel, which indicates the normal distance from $\vec{v}_i^T$. The rendered T-posed UV-map for a T-shirt is shown in Fig. 3a.

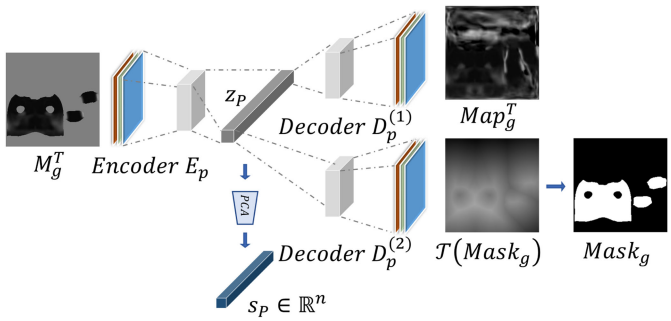


Fig. 4. The basic structure of our proposed ParamNet, which encodes our UV-based garment representation into garment latent space.

In this way, as shown in Fig. 3a, each garment type is represented into one UV space. For each type of garment (upper garments, pants, dresses), the mask of its UV-map denotes its covering area of the human body and thus contains the topology information of different shape styles. The next step is to perform UV map encoding, so that garments with different shape styles and topologies can be then applied to a shape transition framework.

5 Learning the Garment Shape and Style Space

Based on the UV-mask garment representation, we can transfer the complicated 3D shape encoding problem into 2D space by using 2D image auto-encoders. To achieve garment shape and style transition, the UV-position-based garment representation should be mapped into a continuous space, so that the garment shapes and topologies can be smoothly transitioned. By dimensionality reduction and feature extraction, we map the garment representation UV-map to a low-dimensional feature vector, where both the UV-map and its mask information are encoded into a continuous feature space. Therefore, by editing, interpolating and decoding from the feature space, we can achieve our goal of continuous garment shape transition between different garment topologies and shape styles.

We introduce ParamNet, a CNN-based network for garment shape and style space learning. The main idea is to leverage a CNN encoder-decoder structure to encode the given T-posed garment UV representation generated in Section 4. Our UV representation contains two pieces of information: (1) the mask, i.e., the area where the UV map has rendered texels illustrates the area where the garment covers the human body (with T-shirts and pants) or the height range of the T-posed dresses; (2) the rendered texels of the UV map illustrate the vertex positions of the garment. Therefore, when performing the encoding, we also make the decoder generate two maps, one for the mask, and the other for the vertex offsets.

As shown in Fig. 4, the basic structure of ParamNet contains two parts, the encoder $E_P(M_g^T) = z_P \in \mathbb{R}^N$ encodes the T-posed garment 2D representation M_g^T to a high-dimensional hidden space, and decoder $D_P^{1,2}(z_P) = Map_g^T, Mask_g$ decodes vector z_P in a high-dimensional hidden space to the corresponding map and mask.

In practice, we found that the binary masks could barely perform smooth transitions. This is because the discrete binary mask does not have natural continuity in transition. Therefore,

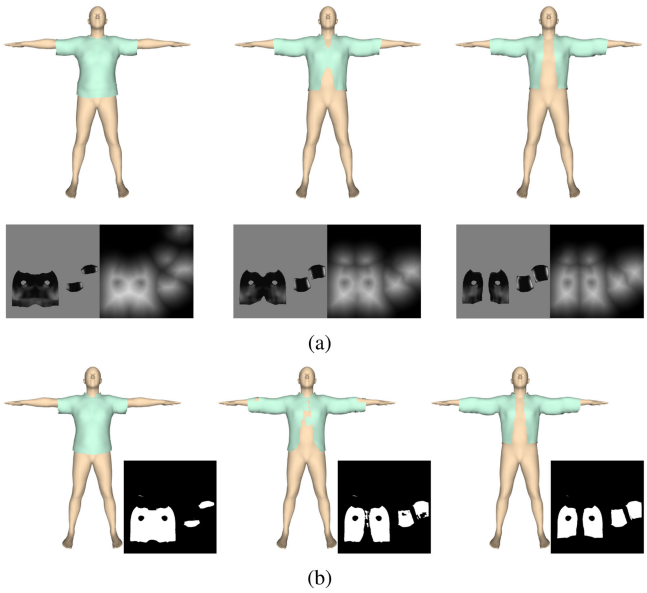


Fig. 5. The demonstration of garment shape transition under UV-mask representations: (a) garment shape transition with the proposed distance-transformed masks, (b) garment shape transition with binary masks. Note that without the distance-transform operation, there will be an unnatural transition between different garment shapes.

to transform the binary masks into a continuous representation, we propose a distance-transform-based method for pre-processing the masks. Specifically, with a given $Mask_g$, we first generate its "bi-distance transform" map as follows:

$$\mathcal{T}(Mask_g) = \mathcal{DT}(Mask_g) - \mathcal{DT}(\mathcal{I} - Mask_g). \quad (1)$$

Here $\mathcal{I}$ is the map with the value 1, and $\mathcal{DT}$ refers to the standard distance transform operation on the mask. The transformed mask, as shown in Fig. 4, demonstrates how far a pixel is from the map boundary, and the continuation of the transformed mask makes it easy to be parameterized and learned from the decoder. Therefore, the decoder becomes $D_P^{1,2}(z_P) = Map_g^T, \mathcal{T}(Mask_g)$, and the loss functions are as follows:

$$\begin{aligned} \mathcal{L}_P^{(map)} &= ||M_g^T - Map_g^T * Mask_g^{(gt)}||_1 \\ \mathcal{L}_P^{(mask)} &= ||\mathcal{T}(Mask_g) - \mathcal{T}(Mask_g^{(gt)})||_1. \end{aligned} \quad (2)$$

After the training phase of ParamNet, the vectors in high-dimensional hidden space $z_P = E_P(M_g^T)$ encrypt the shape variations and characteristics of the T-posed garment shape. To extract the features from the high-dimensional hidden space, we compute the PCA subspace from the hidden space, and sample shape parameters $s_P = E_P(M_g^T) \in \mathbb{R}^n$ from the PCA subspace. To recover the T-posed garment shape from shape parameters, we reversely obtain the vector $z_P = PCA^{-1}(s_P)$, perform the decoder operations to obtain the 2D representation M_g, and finally generate the garment mesh from it.

The demonstration of our T-posed garment shape transition is shown in Figs. 5a and 13, which shows that we can perform a smooth shape transition from short-sleeve T-shirts to long-sleeve shirts, or from skirts to long dresses by interpolating the shape parameters in the feature space. Benefiting from our UV-based representation with mask transformation, we guarantee the continuity in the transition process. The results also demonstrate the function of a general and flexible garment shape encoding and control framework. Although such editing is not strictly semantic (similar to the SMPL [41] model, which cannot control the shape of a specific body part), different PCA basis can still enable garment shape changes on different dimensions of a garment, as demonstrated in our video demo (1:34-1:54). Fig. 5 shows an ablation study using the continuous $\mathcal{DT}$ operations; please refer to Section 10 for more details.

Fig. 6. The demonstration of different garment shape inference methods. From left to right: (a) input 3D scan, (b) segmented garment, (c) garment animation result with garment shape inferred by 3DInferNet, (d) garment animation result with garment shape directly obtained from the scan.

6 Garment Shape Inference

Based on our proposed garment modeling framework and continuous UV-mask representation, a PointNet-based 3DInferNet is proposed to map the 3D garment mesh data domain onto the proposed garment UV space. Note that an alternative method for garment shape inference given a 3D garment scan is directly obtaining the corresponding UV/mask-maps from the scan, similar to our dataset preparation method. However, this method is not feasible. First, for cases when the garment scan is not complete (e.g., Figs. 6a and 6b, the waist of the garment was partially occluded by the hands of the subject), the corresponding UV-maps cannot be generated properly. Second, the garment scans need to be deformed to the standard T-pose for animation, which will produce skinning artifacts especially on the human underarm area. To solve these problems, we propose our garment shape inference module, i.e., 3DInferNet, which maps the scanned point clouds to our garment feature space, and generates garment shapes accordingly. As shown in Fig. 6, our proposed 3DInferNet is necessary for generating complete garment shapes.

To reconstruct the garment shape from any given 3D raw data, and further perform static-to-dynamic garment 3D animation and 3D editing, we introduce the garment shape inference module, which takes a given garment mesh under randomly posed humans as input and extracts the corresponding garment shape parameters. Our method encodes garment shapes with different styles and topologies into a feature space, enabling garment shape extraction from the encoded space. Benefiting from our UV-mask-based garment representation module, our shape inference module can support different garment shapes and styles. In regard to previous works, SIZER [6] can only perform static garment editing. TailorNet [8] has the potential for shape

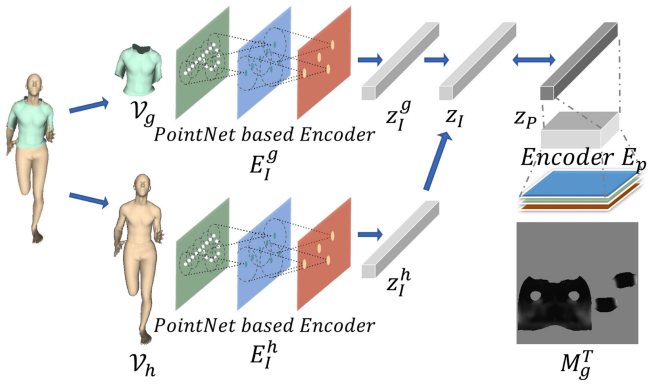


Fig. 7. The basic structure of our 3DInferNet, which infers the garment shape parameters from the input 3D scans.

extraction by un-posing the scan to a standard pose and fitting to its fixed garment template, but it cannot perform large garment shape and topology changes, nor can it deal with garments that do not fit to its templates.

We introduce 3DInferNet to achieve this task. As shown in Fig. 7, in our garment inference module, the input of our model is a pre-segmented garment mesh with an aligned human model. For a given garment mesh $\mathcal{V}_g$ and the corresponding human model mesh $\mathcal{V}_h$, our goal is to generate the corresponding garment parameters s_g. The basic structure of 3DInferNet contains two-branch PointNet-based encoders $E_I^h(\mathcal{V}_h) = z_I^h$ and $E_I^g(\mathcal{V}_g) = z_I^g$, which separately encode the input randomly posed human mesh and garment mesh to hidden space vectors. We implement a fully connected operator $\mathcal{F}$ for extracting features from z_I^h and z_I^g as $\mathcal{F}(z_I^h, z_I^g) = z_I$. The loss is then introduced to constrain the output feature z_I to have less deviation with the vector z_g encoding the shape parameters (see Section 5):

$$\mathcal{L}_I = ||z_I - z_P||_1. \tag{3}$$

As the PointNet [42]-based encoder structure does not rely on the topology or vertex numbers of the input garments, with our style-flexible garment shape representation method, we can extract the garment shape parameters from garment meshes with any 3D inputs. The results are shown in Figs. 1b and 15.

7 Garment Animation

As proposed in Sections 4 and 5, we represent the garment shape with the UV-map and encode it into a feature space. In addition to being applied to the garment shape style transition, the representation and encoding module can also be used for dynamic garment animation to animate the clothed human into arbitrary new poses, which can also be applied to the reconstructed garment shape from Section 6. To better formulate the connection between T-posed garments and corresponding garments under arbitrary poses, we fix the correspondence map defined by the T-posed 1-channel normal distance map, and calculate a 3-channel shift map for each posed garment. Therefore, the static and dynamic garment representations are semantically consistent and suitable for further applications such as garment animation.

Benefiting from our unified UV-mask garment representation, our animation module generates various garment dynamics with different shape styles in a single network, which has not been demonstrated even by the concurrent unified garment representation framework [9]. The garment animation module in our DeepCloth takes the input garment shape parameters from Section 5 with human pose and shape and generates the animated garment mesh. To achieve this goal, we introduce AnimNet, which is a CNN-based network for the garment animation module.

The first step is to represent garments under arbitrary poses. As shown in Fig. 3, we establish a topology-consistent coupling UV-map for a T-posed garment and the same garment under arbitrary poses. As the previous steps determine the UV coordinates of each garment vertex, for a garment animated on a human with other poses, we fix the UV coordinates and set the rendered texels representing the animated shape. Specifically, we calculate the position shift between garment vertex $\vec{g}_i^T$ and corresponding SMPL sub-vertex $\vec{v}_i^T$, and set position shift (dx, dy, dz) as the rendered texels.

There are three main advantages for our coupling UV-map representation. First, by fixing the correspondence between garment vertices and SMPL UV coordinates, a one-to-one mapping can be applied to the T-posed garment vertices and animated garment vertices, and randomly posed garments, e.g., floating front-opening T-shirts or folding skirts, can be represented more easily. Second, with the same UV coordinate for every garment vertex, the mapping between T-posed garments and its randomly posed condition can be learned more easily using a CNN-based network. Third, since the coupling UV-map has the same mask, during animation, we only need to infer the rendered texels of the second map.

We denote M_g^T for the T-posed standard garment UV map, and M_g^A for the animated garment UV map. In AnimNet, we take M_g^T as input and generate M_g^A as output. Meanwhile, to encode the human pose and shape information, we find that the normal information actually guides the position map of the garment; therefore, we use the normal map $N_{\beta,\theta}$ of the human model to represent the human shape β and pose θ information. As shown in Fig. 8, we use a CNN-based encoder E_A and decoder D_A with skip connections to generate the inferred garment map Map_g^A. The main loss function is as follows:

$$\begin{aligned} M_g^A &= Map_g^A * Mask_g^{(gt)} \\ \mathcal{L}_A^{(map)} &= ||M_g^{A(gt)} - M_g^A||_1. \end{aligned} \tag{4}$$

Here $Mask_g^{(gt)}$ is the ground truth mask, and we only need to constraint the generated position map to have the same value as the ground truth map inside the masked area, as the input garment shape parameters contain the mask information.

Note that our UV-mask-based garment representation implicitly encodes the garment shape dynamic information, which helps AnimNet to learn the deformation of different garment styles when performing garment animation(e.g., front-opening/closing garments). As shown in Fig. 9, the network learns garment dynamic shapes with different garment styles, such as the floating bottom part of the front-opening T-shirt. The UV-masks not only mask the garment geometry, but also encode garment styles, which is reflected in animation results. The example garment animation results are shown in Figs. 13 and 14, which show that we

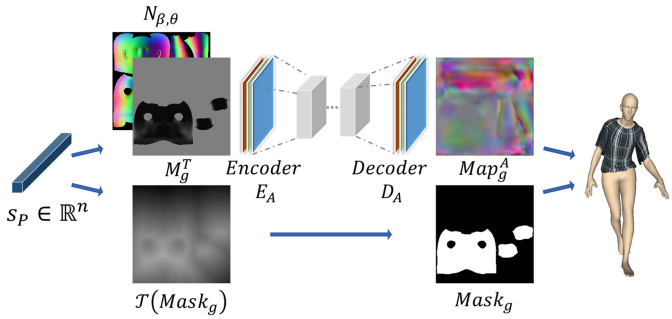


Fig. 8. The basic structure of our AnimNet, which generates garment dynamics under arbitrary human poses and shapes.

can animate different types of garments with various human shapes, poses and garment styles.

8 Garment Shape-Style Editing

As mentioned in Section 5, we determine the PCA subspace from the garment representation latent space, which serves as a more compact and semantic encoding of garment style variation than the original latent space. Therefore, we can perform garment shape and style transition and editing by shifting the PCA space vectors.

Take the garment reconstruction procedure as an example. Given garment 3D scans and human models, with the garment shape parameters solved by 3DInferNet, we can perform shape editing by the following steps: (1) mapping the shape parameters to the same PCA subspace calculated in Section 5 as $s_I = PCA^g(z_I)$, (2) editing some dimensions of shape parameters s_I as s'_I, (3) obtaining the vector $z'_I = PCA^{-1}(s'_I)$, and (4) recovering a new garment shape from z'_I.

After the garment editing step, the new garment can then be fed to our AnimNet for garment animation, so that dynamic garment shape editing can be applied to the given pose sequences. The results are shown in Figs. 1c, 1d, 1e and 15b.

To show the necessity of PCA for garment editing, we perform an ablation study for editing the garment parameters in two ways: (1) editing one dimension each time in PCA space, and (2) editing one dimension each time in the original latent space vector, with the relative parameter shift degree. As shown in Fig. 10, we can perform convenient and semantic garment editing by shifting different dimensions in PCA space, while such editing can hardly be performed by editing in the original latent space. This is because the PCA operation

Fig. 9. The demonstration of different styles garment animation. From left to right: front-opening T-shirt animation, front-closing T-shirt animation, and skinning results with front-opening T-shirt. The results clarify that AnimNet learns garment dynamics from the encoded garment styles.

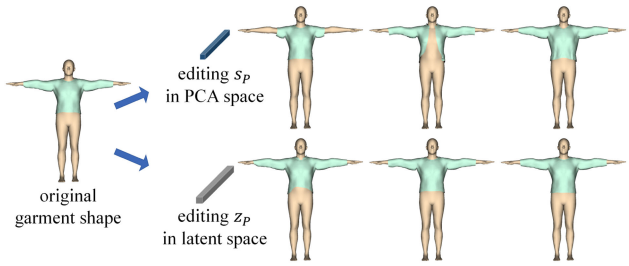


Fig. 10. The ablation study of garment editing in PCA space and the original latent space. Top: editing parameters in PCA space; bottom: editing parameters in original latent space. Different garment shape styles can be edited easily by shifting the PCA vectors. In contrast, directly editing the original latent space vectors can hardly yield meaningful garment shape editing results.

encodes the original garment shape space in a more compact form and extracts the semantic style patterns.

9 Garments Not Homotopy to Human Body

Garments that are not homotopy to the human body, e.g., skirts and dresses, are not suitable for our SMPL-UV-based representation. Therefore, we separately design the UV-mask representation for those garments. We map them to an independent UV coordinate to better reflect the characteristics of the garment geometry. The boundary of the UV map demonstrates the edges and basic shape information (such as the height of dresses), and the rendered texels indicate the 3D positions of the garment vertices.

Specifically, when dealing with clothing that is not homotopy to the human surface (e.g., dresses), we set an independent UV coordinate accordingly. For T-posed dresses and skirts, as their geometry circles around the lower body, we leverage cylindrical coordinates and calculate the UV coordinates as follows: for each garment vertex $\vec{g}_i'^T$, we transfer it into cylindrical coordinates: $\vec{g}_i'^T(x,y,z) \rightarrow \vec{g}_i'^T(r,y,\theta)$ where $r = \sqrt{x^2+z^2}$ and $\theta = \arctan(z,x)$. The UV coordinate t'_i for $\vec{g}_i'^T(r,y,\theta)$ is $t'_i = ((y_0 - y)\cos(\theta) + 0.5, (y_0 - y)\sin(\theta) + 0.5)$, and the rendered texel is just (x,y,z) to indicate the vertex positions. Here, y_0 serves as the height threshold of the skirts; in practice, we set $y_0 = 0.2$ above the root joint of the human model. The rendered T-posed UV-map for a skirt is shown in Fig. 11a. Additionally, continuous style transitions can be performed similar to Section 5, and the results are shown in Fig. 12.

For garments under arbitrary poses, similar to our procedure in Section 7, we fix the garment vertex UV coordinates, and directly use the vertex positions to set the rendered texels, as used in the T-posed scenario. The rendered arbitrarily posed UV-map for a skirt is shown in Fig. 11b.

Apart from the different UV-coordinate layouts of different types of garments, the main pipeline (ParamNet, AnimNet and 3DInferNet) works the same as garments homotopy to human bodies. As shown in Fig. 13, Fig. 14 and our video demo, we can perform skirt style editing with robust and vivid animation results, which shows the robustness of our UV-mask-based representation pipeline for dealing with different types of garments.

10 Experiments

In our experiments, we use CLOTH3D [11], which is a large-scale synthetic dataset with various garment shape styles

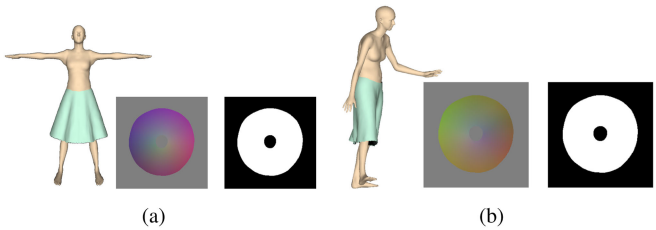


Fig. 11. The demonstration on coupling the UV-map representation of the garment that is not homotopy to the human body. (a) 2D representation for the T-posed skirt. (b) 2D representation for the randomly posed skirt for Section 7.

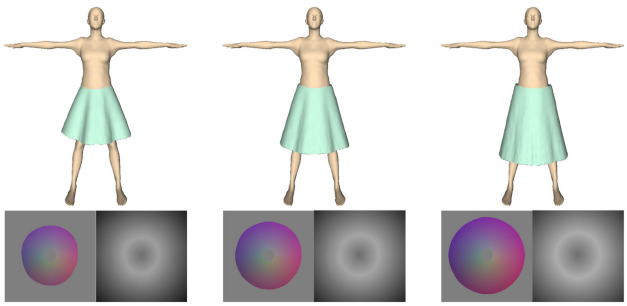


Fig. 12. The demonstration of the skirt shape transition under UV-mask representations.

suitable for our pipeline for our training and testing data. We test our garment encoding module with four garment kinds: upper garments, pants, skirts, and jumpsuits. The animation module is tested with all these kinds of garments. In addition, we take the BUFF Dataset [43] and Twindom Dataset (https://web.twindom.com/) as input to test our garment inference module for 3D garment animation and editing from input 3D data.

The network structures for our CNN-based encoders, e.g., E_P in ParamNet (Section 5) and E_A in AnimNet (Section 6), are based on the ResNet-18 [44] structure. The decoders accordingly are six stacked up-sampling layers with convolution layers. The structure for our PointNet-based [42] encoders E_I^g and E_I^h (Section 7) is based on PointNet structures for extracting features from point clouds.

In addition to the data preparation, with the NVIDIA GeForce GTX TITAN X GPU, the training procedure for ParamNet takes approximately 50 hours, while AnimNet and 3DInferNet take 100 hours separately for each kind of garment. The garment rendering results generated from the network output, together with a standard collision resolving procedure, take approximately 2 seconds per human per frame with one garment.

Garment Shape Representation and Animation. To evaluate our garment shape representation results, Fig. 13 demonstrates the garment shape variations controlled by different parameters, with the PCA parameter s_P varying within the range of 1.0 σ. The results in Fig. 13 show that we can perform plausible garment shape variation from long-sleeve shirts to short-sleeve T-shirts, from front-closing T-shirts to front-opening T-shirts, from long pants to shorts, and from long dresses to short skirts, which clarifies that our method provides a more general garment shape representation model than TailorNet [8]. Additionally, as demonstrated in Fig. 14, given different garment styles and different body shapes, we can generate clothed human animation results, which makes our framework capable of representing garment shapes under various human shapes, poses and garment styles.

To demonstrate the necessity of the $\mathcal{DT}$ operation, we evaluate garment shape transition with and without the "bi-distance transform" operation. To directly use binary masks, we use a network structure for garment shape representation similar to ParamNet. As shown in Fig. 5, the lack of a continuous boundary constraint leads to an unnatural

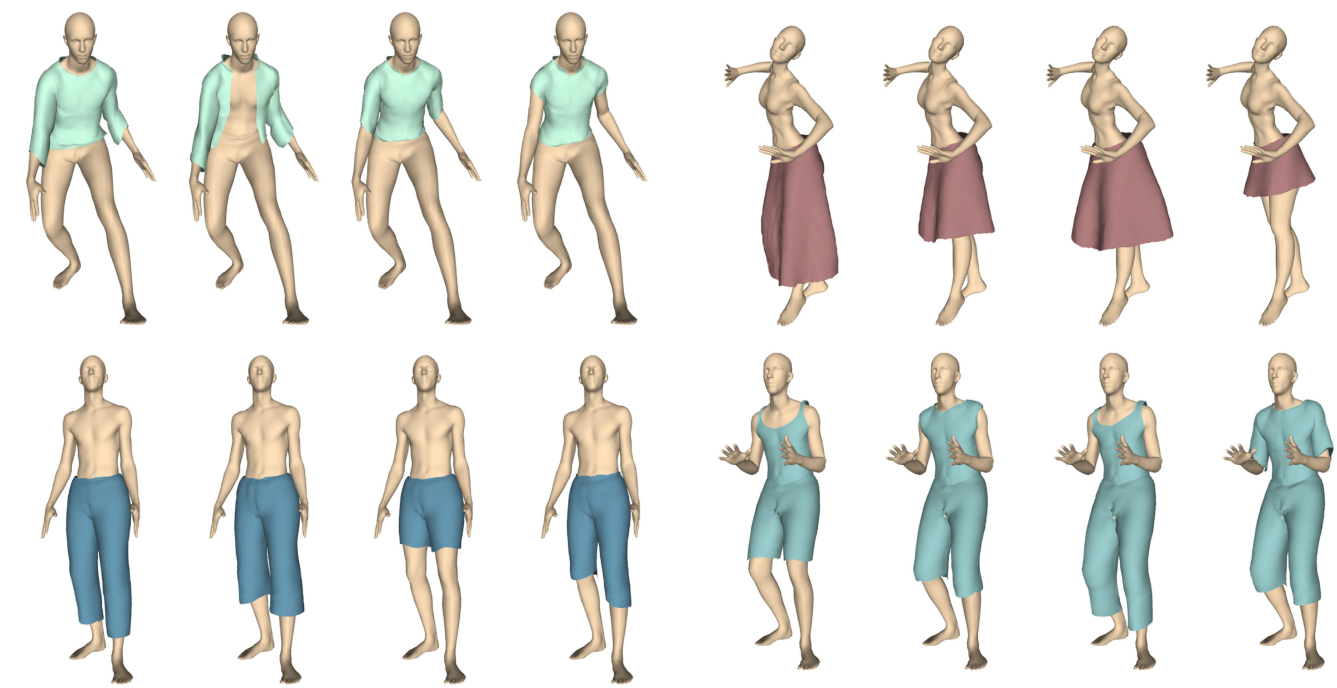


Fig. 13. The demonstration of our shape variation for different kinds of garments. Each block shows the garment shape parameter variation on one type of garments.

Fig. 14. The demonstration of garment animation results. Left: the same garments under different human shapes. Middle: different garment styles on the same person. Right: the same garments under different human poses.

Fig. 15. The experimental results of our garment inference module. (a) Garment inference with a given input, from left to right: original 3D scan, aligned human model, 3D segmented T-shirts with pants, and animation results. (b) Top: garment retargeting results, bottom: garment 3D editing results.

shape transition. In contrast, with our "bi-distance transform", the masks are transformed smoothly, thus generating continuous garment shape transition results.

It should be clarified that although the $\mathcal{DT}$ operations change the binary masks into a continuous form, other continuous $\mathcal{DT}$ transforms, such as the truncated distance transform or cosine encoding based on the distance transform images, can also be used. As shown in Fig. 16, we show that the continuous transforms based on our bi-distance transform can also be applied for a smooth transition.

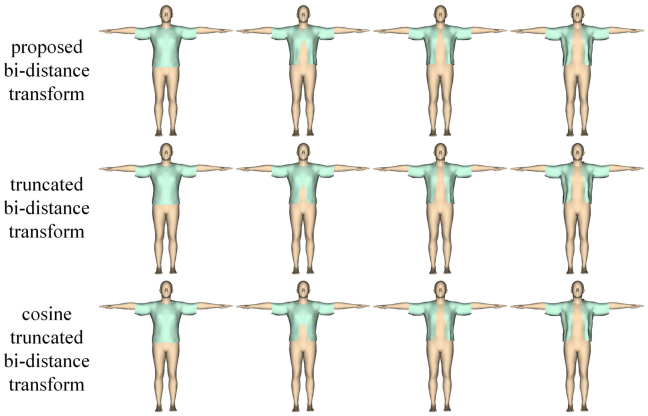


Fig. 16. The ablation study on smooth garment transition using different continuous mask transforms based on the proposed bi-distance transform. From top to bottom: garment shape transition using the proposed method, the truncated bi-distance transform and the cosine encoding on the bi-distance transform.

Here, we provide a qualitative comparison with TailorNet [8]. Benefiting from our UV-position with mask representation, we can represent different garment shape styles and topologies in the same framework, while TailorNet [8] needs separate templates for each kind of garment. As shown in Fig. 17, TailorNet [8] needs two separate templates for representing T-shirts and shirts; thus, it cannot represent tops, front-opening shirts or half-long-sleeve shirts. Meanwhile, our model can represent all these upper garment shapes in the same model and can perform shape parametrization and reasonable transitions between these shapes, as shown in Fig. 14 and our video demo. In addition, TailorNet [8] cannot perform shape control for pants to shorts or deal with long dresses, while our method can address such cases.

Garment Shape Inference and Editing. To evaluate our garment shape inference and editing module, we use the 3D scan of the Twindom dataset to perform garment shape inference and editing. The Twindom dataset is a high-resolution 3D scan dataset with multiple clothed humans under arbitrary poses.

To fit in our module, for the 3D scan clothed model, we first perform a pose alignment procedure with the standard human model to obtain the pose information and the inside posed human mesh $\mathcal{V}_h$, and segment the 3D scan to each garment mesh $\mathcal{V}_g$. We then perform 3DInferNet introduced in Section 6 to extract the shape parameters s_I for the

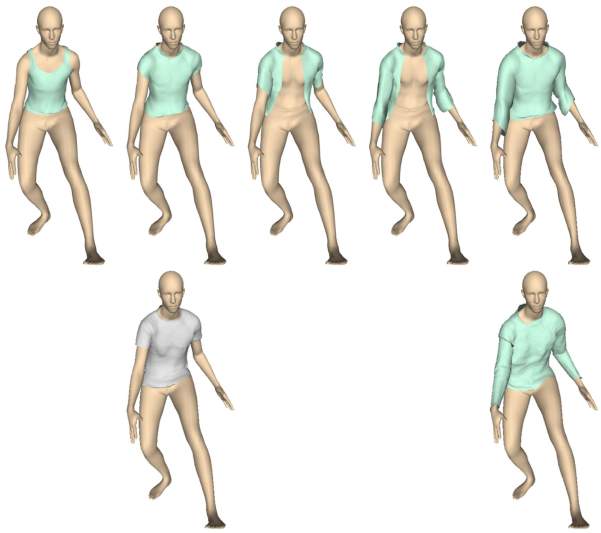


Fig. 17. Comparison between our method and TailorNet [8] on garment representation. Top: our animated garment shapes with different garment styles. Bottom: results generated by TailorNet [8]. Note that TailorNet [8] needs two templates to perform such garment shapes, and cannot represent some of the shapes. Our method can represent all the shapes in the same framework.

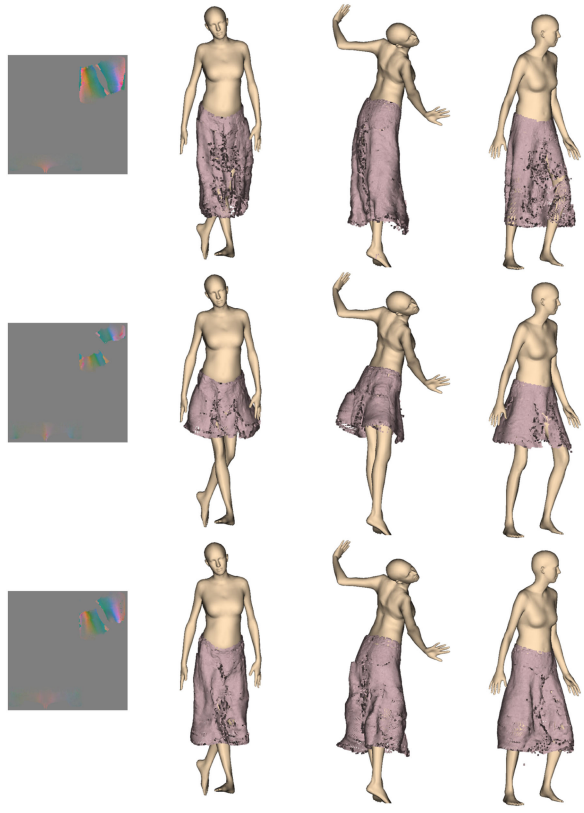


Fig. 18. The demonstration of dress representation and animation using SMPL-UV. From left to right: masked dress representation UV map and three animated poses of the corresponding dress.

garment, and generate posed sequences using AnimNet using garment 2D representation generated by s_I. The results are shown in Fig. 1, which shows that we can correctly recover the shape of the original garments. In addition, we can perform shape editing by shifting the shape parameters s_I as s'_I and perform the animation procedure. The results are shown in Figs. 1 and 15.

Using SMPL-UV in Dress Shape Representation and Animation. In our pipeline, we currently use a different UV layout for dress representation and animation. Actually, it is also feasible to represent dresses with SMPL-UV, but there will be artifacts especially on the between-leg regions, when performing dress animation and transition. Here, we perform ablation studies on using SMPL-UV in dress shape representation and animation.

To represent a T-posed garment that is not homotopy to the human body, the main concern is to determine the correspondence between the garment and the body model. We solve the SMPL-based surface deformation with the SMPL-garment correspondence by minimizing the Chamfer distance energy function between the corresponding human leg areas on a naked human model and the dress mesh. The UV-position map with continuous $\mathcal{DT}$-based masks is then generated accordingly, similar to Section 4. However, as the dress geometry is not homotopy to the human body, it is difficult to describe as a normal distance map, so we adjust it as a 3-channel shift map, as shown in the left column of Fig. 18. Then, we generate the animated dress geometry UV map similar to Section 7 and train our ParamNet and AnimNet accordingly.

As the dresses are not homotopy to the human body, it is difficult to design a mesh layout for completed mesh rendering, so we render the results in a point-based manner. The results are shown in Fig. 18 and our video demo, which show that although dress animation under various styled can also be performed using SMPL-UV, there are still artifacts on the between-leg regions and the boundaries. Meanwhile, the non-homotopy dress design will avoid such problems, which is attributed to a more topology-consistent design for the corresponding garment type.

We also perform a quantitative comparison of animation accuracy between the proposed dress-UV and SMPL-UV, as shown in Table 1. As the SMPL-UV representation better reflects the geometric connection between the SMPL bodies and the dresses, the SMPL-UV performs slightly better quantitatively than the dress-UV. However, the problems addressed above are still difficult to solve. A more general and subtle design may be needed in the future works for representing different garments in the same SMPL-UV.

Experiments on Real-World Data. To illustrate the ability of our model for flexibly representing and animating real-world garments, we also evaluate our model on the real-captured THuman3.0 dataset, which contains 154 human-garment combinations, where each person under 2-3 sets of garments performs 30 to 60 poses. The data samples are shown in Fig. 19. In particular, we evaluate upper garment animation and transition with the given real-captured data. As shown in Fig. 20 and our video demo, by performing fine-tuning with the pre-trained AnimNet and ParamNet on such a dataset, our model can reflect the dynamic 3D patterns of different garments, and perform vivid animation styles given different upper garments. Additionally, our model can perform smooth and natural style transitions between different garments. The experiments show the ability of our model to deal with real-world garment data.

Meanwhile, we also experiment on garment reconstruction from an input RGB image. The network structure is a standard CNN-based encoder-decoder structure, similar to ParamNet

TABLE 1
Mean Vertex-to-Vertex Error (mm) of Our AnimNet Method and PointNet-Based Method for Different Garment Types, With Quantitative Evaluations on SMPL-UV Dresses

garment type	Ours	PointNet-based method
T-shirts & shirts	16.34	20.45
pants & shorts	13.51	18.63
long dresses & skirts	31.32	40.98
long dresses & skirts (SMPL-UV)	27.90	/

in Fig. 4, while the input is replaced by an RGB image. The experiments are also performed on the THuman3.0 dataset. After training, our model can predict the basic garment shapes from an input RGB image in the test set, and the reconstructed garment can be applied to realistic garment animation given garment shape styles, as shown in Fig. 21.

Quantitative Evaluation. For the garment animation module, we compare our UV-based garment animation method with the PointNet-base [42] method, which extracts the point features of the garment mesh and the posed human mesh, to infer the shift of garment vertices. As the CLOTH3D [11] dataset contains various garment styles, e.g., front-opening and front-closing T-shirts with long or short sleeves, the garment styles cannot be fit into a fix garment template. Therefore, traditional MLP or other methods suitable for dealing with meshes with a fixed number of vertices could not be evaluated. The CLOTH3D [11] is split into 95 percent for training and 5 percent for testing. The results applied on the test set are as follows; here, the loss is the mean vertex-to-mesh error.

As shown in Table 1, our method outperforms the PointNet-based method. This is because the garment styles and topologies vary over a wide range in the CLOTH3D [11] dataset, while traditional PointNet-based [42] methods have some limits, especially in long dress cases. Note that as our goal is to establish a general garment representation enabling garment shape and style transition, the PointNet-based method actually does not meet our requirement, while our UV-based representation can handle these problems, as demonstrated in Fig. 13.

We also evaluate our garment animation module using the TailorNet [8] dataset. The dataset we use, i.e., CLOTH3D, does not contain many garment wrinkle details, although it contains numerous human pose sequences with different human shapes, each sequence corresponding to an independent garment mesh, providing multiple garment styles and topologies on both T-pose and animated poses suitable for our framework. In contrast, the TailorNet [8] dataset contains garments with more garment wrinkle details, but due to its fixed garment templates for each type of garment, it cannot be used for training garment representation enabling garment topology and style transition. Thus, we only evaluate our garment animation module here. We train our AnimNet on the TailorNet [8] dataset and compare our performance with TailorNet [8]. As shown in Fig. 22, with our framework trained on the TailorNet [8] dataset, we can also generate vivid garment details when performing garment animation. Additionally, we make a quantitative evaluation on the TailorNet dataset, as shown in Table 2, which shows that by training on a particular garment style, our model can achieve comparable animation qualities with TailorNet. The reason that our model does not achieve better results is that, we focus more on a flexible and general garment representation, while TailorNet focuses more on animation quality and accuracy given fixed garment vertex templates.

For garment shape inference and application, we compare our method with the state-of-the-art garment reconstruction method MulayCap [34], which takes a single-view RGB video as input and dynamically generates a two-layer human with garment mesh. We use a 4D sequence in the BUFF [43] dataset as the input, as demonstrated in Fig. 1. We provide [34] with the aligned SMPL shape and poses for every frame, and compare the vertex-to-mesh error between the generated garments and the ground truth input. For our method, we use only the ground truth garment mesh of the first frame for

Fig. 19. Data samples of the THuman3.0 dataset.

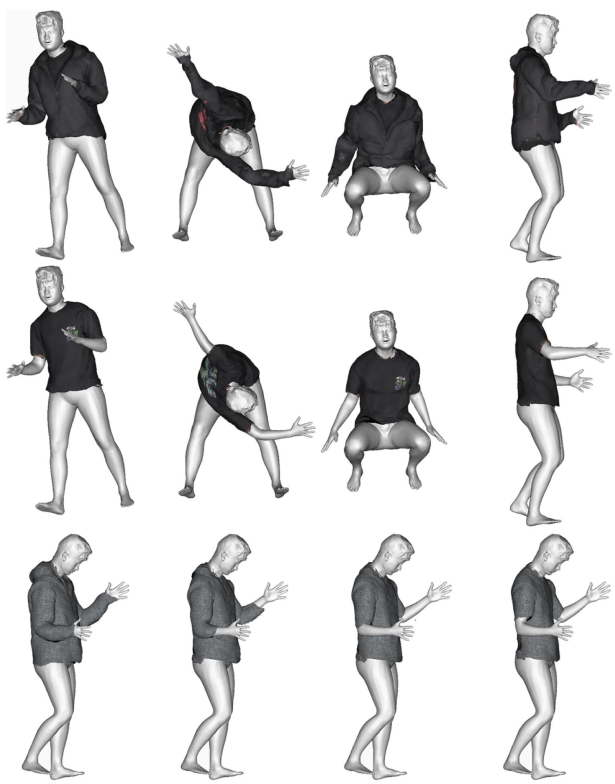


Fig. 20. The demonstration of our DeepCloth models tested on THuman3.0 dataset. From top to bottom: animation results of two kinds of garments (long coat / short T-shirt), and garment shape transition between the two garments. Note that our model can perform vivid animation results given different garment styles.

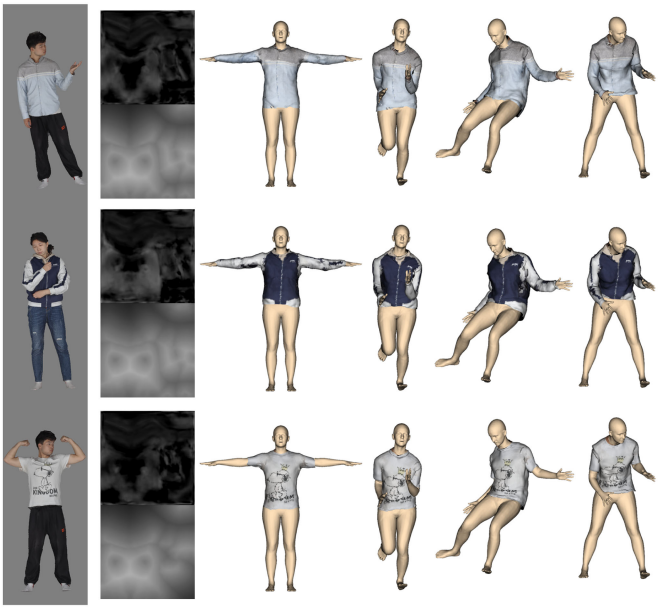


Fig. 21. The demonstration of RGB garment reconstruction. From left to right: input real-world RGB image, predicted UV map and $\mathcal{DT}$-mask, reconstructed garment shape, and three animation results.

garment shape inference, similar to Fig. 1, and provide garment animation with SMPL poses and shape. As demonstrated in Fig. 23, our method performs similarly to MulayCap [34]. From a methodological perspective, MulayCap [34] is a 4D garment reconstruction pipeline that uses RGB and human parsing information in every frame, for per-frame garment geometry optimization and shape-from-shading geometry detail generation. Our method is an animation module, which only takes the garment mesh of the first frame and the SMPL motion sequence as input, without using the input RGB information, which is why our model can hardly outperform MulayCap [34]. However, the comparable results still demonstrate the animation ability of our model.

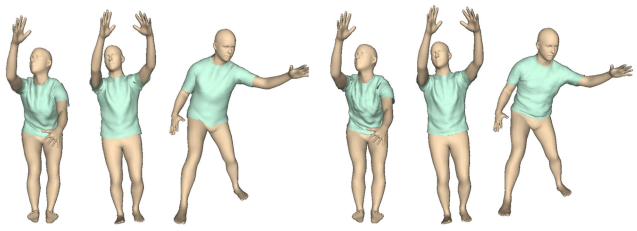


Fig. 22. The demonstration of garment animation evaluated on the TailorNet [8] dataset: (a) our results, (b) results generated by TailorNet [8].

TABLE 2
Mean Vertex-to-Vertex Error (mm) of Our AnimNet Method and TailorNet [8] Method on the TailorNet Dataset for Different Garment Types

garment type	Ours	TailorNet method
male T-shirts	11.58	11.2
male pants	10.39	8.1
female T-shirts	12.97	12.3
female pants	6.10	4.8

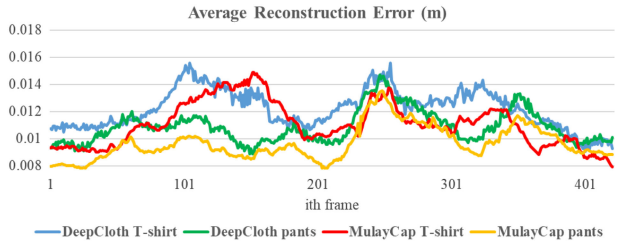


Fig. 23. Quantitative comparison with MulayCap [34] using the rendered 4D model and aligned SMPL poses and shape in the BUFF [43] Dataset as input. The result shows a quantitative comparison between the two methods in one 4D sequence using the per-vertex average error.

11 Conclusion

In this paper, we propose DeepCloth, a unified neural garment representation framework that can perform garment shape and style transitions by learning the shape space of 3D garments. Our method enables modeling garments under different topologies using the "UV-position map with mask" representation, and can perform smooth and free garment transitions by mapping such representations into a continuous feature space. By introducing AnimNet and 3DInferNet, our representation allows the generation of 4D clothed human dynamic sequences or the recovery of garment shapes from 3D scans and performing animation and garment shape editing. We believe that the proposed topology-aware UV-Mask-based representation takes an important step forward in the field of 3D clothing, especially with the introduction of neural masks for controlling the topology and shape of garments.

Limitations and Discussions. Similar to TailorNet [8], we also rely on an explicit collision resolving step. Additionally, at the moment, we cannot handle garments with pockets and collars, which may be resolved by introducing another detailed UV layer. We did not experiment on long dresses that cover the upper body, which can be represented in the future by combining SMPL-UV and dress-UV. Besides above, future works will focus on generating more garment styles based on our work.

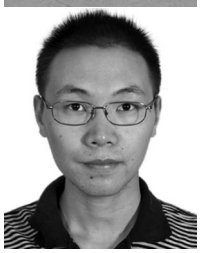
References

[1] X. Provot *et al.*, "Deformation constraints in a mass-spring model to describe rigid cloth behaviour," in *Proc. Graph. Interface*, 1995, pp. 147–147.
[2] Z. Lahner, D. Cremers, and T. Tung, "DeepWrinkles: Accurate and realistic clothing modeling," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 698–715.
[3] E. Gundogdu *et al.*, "GarNet: Improving fast and accurate static 3D cloth draping by curvature loss," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 1, pp. 181–195, Jan. 2022.
[4] Q. Ma *et al.*, "Learning to dress 3D people in generative clothing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 6468–6477.
[5] N. Jin, Y. Zhu, Z. Geng, and R. Fedkiw, "A pixel-based framework for data-driven clothing," in *Proc. ACM SIGGRAPH/Eurographics Symp. Comput. Animation*, 2020, Art. no. 13.
[6] G. Tiwari, B. L. Bhatnagar, T. Tung, and G. Pons-Moll, "SIZER: A dataset and model for parsing 3D clothing and learning size sensitive 3D clothing," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 1–18.
[7] Y. Shen, J. Liang, and M. C. Lin ., "GAN-based garment generation using sewing pattern images," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 225–247.
[8] C. Patel, Z. Liao, and G. Pons-Moll, "TailorNet: Predicting clothing in 3D as a function of human pose, shape and garment style," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 7363–7373.

[9] E. Corona, A. Pumarola, G. Alenya, G. Pons-Moll, and F. Moreno-Noguer, "SMPLicit: Topology-aware generative model for clothed people," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 11870–11880.
[10] T. Alldieck, G. Pons-Moll, C. Theobalt, and M. Magnor, "Tex2Shape: Detailed full human body geometry from a single image," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 2293–2303.
[11] H. Bertiche, M. Madadi, and S. Escalera, "CLOTH3D: Clothed 3D humans," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 344–359.
[12] K.-J. Choi and H.-S. Ko, "Stable but responsive cloth," in *Proc. ACM SIGGRAPH Courses*, 2005, Art. no. 1.
[13] T. Liu, A. W. Bargteil, J. F. O'Brien, and L. Kavan, "Fast simulation of mass-spring systems," *ACM Trans. Graph.*, vol. 32, no. 6, 2013, Art. no. 214.
[14] J. Bonet and R. D. Wood, *Nonlinear Continuum Mechanics for Finite Element Analysis*. Cambridge, U.K.: Cambridge Univ. Press, 1997.
[15] C. Jiang, T. Gast, and J. Teran, "Anisotropic elastoplasticity for cloth, knit and hair frictional contact," *ACM Trans. Graph.*, vol. 36, no. 4, pp. 152:1–152:14, Jul. 2017.
[16] W. H. Press, *Numerical Recipes 3rd Edition: The Art of Scientific Computing*. Cambridge, U.K.: Cambridge Univ. Press, 2007.
[17] D. Terzopoulos, J. Platt, A. Barr, and K. Fleischer, "Elastically deformable models," *ACM SIGGRAPH Comput. Graph.*, vol. 21, pp. 205–214, 1987.
[18] D. Baraff and A. Witkin, "Large steps in cloth simulation," in *Proc. 25th Annu. Conf. Comput. Graph. Interactive Techn.*, 1998, pp. 43–54.
[19] H. Wang, F. Hecht, R. Ramamoorthi, and J. F. O'Brien, "Example-based wrinkle synthesis for clothing animation," *ACM Trans. Graph.*, vol. 29, no. 4, Jul. 2010, Art. no. 107.
[20] E. de Aguiar, L. Sigal, A. Treuille, and J. K. Hodgins, "Stable spaces for real-time clothing," *ACM Trans. Graph.*, vol. 29, no. 4, Jul. 2010, Art. no. 106.
[21] D. Kim, W. Koh, R. Narain, K. Fatahalian, A. Treuille, and J. F. O'Brien, "Near-exhaustive precomputation of secondary cloth effects," *ACM Trans. Graph.*, vol. 32, no. 4, Jul. 2013, Art. no. 87.
[22] T. Y. Wang, D. Ceylan, J. Popovic, and N. J. Mitra, "Learning a shared shape space for multimodal garment design," *ACM Trans. Graph.*, vol. 37, no. 6, pp. 1:1–1:14, 2018.
[23] I. Santesteban, M. A. Otaduy, and D. Casas, "Learning-based animation of clothing for virtual try-on," *Comput. Graph. Forum*, vol. 38, no. 2, pp. 355–366, 2019.
[24] J. Yang, J.-S. Franco, F. Hetroy-Wheeler, and S. Wuhrer, "Analyzing clothing layer deformation statistics of 3D human motions," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 237–253.
[25] E. Gundogdu, V. Constantin, A. Seifoddini, M. Dang, M. Salzmann, and P. Fua, "GarNet: A two-stream network for fast and accurate 3D cloth draping," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 8739–8748.
[26] V. Lazova, E. Insafutdinov, and G. Pons-Moll, "360-degree textures of people in clothing from a single image," 2019, *arXiv: 1908.07117*.
[27] X. Zhang *et al.*, "Neural light transport for relighting and view synthesis," *ACM Trans. Graph.*, vol. 40, no. 1, Feb. 2021, Art. no. 9.
[28] D. Bradley, T. Popa, A. Sheffer, W. Heidrich, and T. Boubekeur, "Markerless garment capture," *ACM Trans. Graph.*, vol. 27, no. 3, 2008, Art. no. 99.
[29] T. Popa *et al.*, "Wrinkling captured garments using space-time data-driven deformation," *Comput. Graph. Forum*, vol. 28, no 2, pp. 427–435, 2009.
[30] C. Stoll, J. Gall, E. de Aguiar, S. Thrun, and C. Theobalt, "Video-based reconstruction of animatable human characters," *ACM Trans. Graph.*, vol. 29, no. 6, pp. 139:1–139:10, 2010.
[31] G. Pons-Moll, S. Pujades, S. Hu, and M. J. Black, "ClothCap: Seamless 4D clothing capture and retargeting," *ACM Trans. Graph.*, vol. 36, no. 4, pp. 73:1–73:15, Jul. 2017.
[32] X. Chen, B. Zhou, F. Lu, L. Wang, L. Bi, and P. Tan, "Garment modeling with a depth camera," *ACM Trans. Graph.*, vol. 34, no. 6, pp. 203:1–203:12, Oct. 2015.
[33] T. Yu *et al.*, "SimulCap : Single-view human performance capture with cloth simulation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 5499–5509.
[34] Z. Su *et al.*, "MulayCap: Multi-layer human performance capture using a monocular video camera," *IEEE Trans. Vis. Comput. Graphics*, vol. 28, no. 4, pp. 1862–1879, Apr. 2022.
[35] B. L. Bhatnagar, G. Tiwari, C. Theobalt, and G. Pons-Moll, "Multi-garment net: Learning to dress 3D people from images," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 5419–5429.
[36] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik, "End-to-end recovery of human shape and pose," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7122–7131.
[37] T. Alldieck, M. Magnor, B. L. Bhatnagar, C. Theobalt, and G. Pons-Moll, "Learning to reconstruct people in clothing from a single RGB camera," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 1175–1186.
[38] M. Habermann, W. Xu, M. Zollhöfer, G. Pons-Moll, and C. Theobalt, "LiveCap: Real-time human performance capture from monocular video," *ACM Trans. Graph.*, vol. 38, no. 2, Apr. 2019, Art. no. 14.
[39] M. Habermann, W. Xu, M. Zollhofer, G. Pons-Moll, and C. Theobalt, "DeepCap: Monocular human performance capture using weak supervision," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 5051–5062.
[40] T. Alldieck, M. Magnor, W. Xu, C. Theobalt, and G. Pons-Moll, "Video based reconstruction of 3D people models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 8387–8397.
[41] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "SMPL: A skinned multi-person linear model," *ACM Trans. Graph.*, vol. 34, no. 6, pp. 248:1–248:16, Oct. 2015.
[42] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "PointNet: Deep learning on point sets for 3D classification and segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 77–85.
[43] C. Zhang, S. Pujades, M. J. Black, and G. Pons-Moll, "Detailed, accurate, human shape estimation from clothed 3D scan sequences," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 5484–5493.
[44] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

Zhaoqi Su received the BS degree from the Department of Physics, Tsinghua University, Beijing, China, in 2017. He is currently working toward the PhD degree in the Department of Automation, Tsinghua University, Beijing, China.

Tao Yu received the BS degree in measurement and control from the Hefei University of Technology, Hefei, China, in 2012, and the PhD degree in instrumental science from Beihang University, Beijing, China. He is a postdoctoral researcher with Tsinghua University. His current research interests include computer vision and computer graphics.



Yangang Wang (Member, IEEE) received the BE degree from Southeast University, Nanjing, China, in 2009, and the PhD degree in control theory and technology from Tsinghua University, Beijing, China, in 2014. He was an associate researcher with Microsoft Research Asia from 2014 to 2017. He is currently an associate professor with Southeast University. His research interests include image processing, computer vision, computer graphics, motion capture, and animation.

Yebin Liu (Member, IEEE) received the BE degree from the Beijing University of Posts and Telecommunications, Beijing, China, in 2002, and the PhD degree from the Automation Department, Tsinghua University, Beijing, China, in 2009. He is currently an associate professor with Tsinghua University. He was a research fellow with the Computer Graphics Group, Max Planck Institute for Informatik, Germany, in 2010. His research areas include computer vision, computer graphics, and computational photography.