BCNet: Learning Body and Cloth Shape from A Single Image

Boyi Jiang[1], Juyong Zhang[1⋆], Yang Hong[1], Jinhao Luo[1], Ligang Liu[1], and Hujun Bao[2]

[1] University of Science and Technology of China
[2] State Key Lab of CAD&CG, Zhejiang University

Abstract. In this paper, we consider the problem to automatically reconstruct garment and body shapes from a single near-front view RGB image. To this end, we propose a layered garment representation on top of SMPL and novelly make the skinning weight of garment independent of the body mesh, which significantly improves the expression ability of our garment model. Compared with existing methods, our method can support more garment categories and recover more accurate geometry. To train our model, we construct two large scale datasets with ground truth body and garment geometries as well as paired color images. Compared with single mesh or non-parametric representation, our method can achieve more flexible control with separate meshes, makes applications like re-pose, garment transfer, and garment texture mapping possible.

Keywords: clothed body reconstruction, 3D garment shape, 3D body shape, skinning weight

1 Introduction

Applications like virtual try-on, VR/AR, and entertainment need detailed and accurate reconstruction of both body and dressed garments with simple input like color image. However, the variety of body shapes, postures and garment categories, makes it a very challenging problem. A simulation-based method [52] explores this problem, but their solution is dedicated and time-consuming. In this paper, we aim to automatically reconstruct both body and cloth shapes from just a single near-front view image, utilizing the powerful fitting ability of the deep neural network.

In recent years, body shape reconstruction from images has made significant progress [23, 36, 24, 29]. A common way is to infer the shape and pose parameter of a statistical body model, like SMPL [32]. These methods are robust for different posture, but the reconstructed geometry is constrained to be within the model space, which can not capture the complex cloth shape.

To infer detailed geometry beyond body shape, some non-parametric representations have been proposed [46, 57, 35, 43]. These non-parametric representations based on voxel and implicit function can recover arbitrary shapes. However,

⋆ Corresponding author. Email: juyong@ustc.edu.cn.

voxel representation is hard to recover shape details due to their large memory consumption for high resolution. Although implicit representation is more memory efficient, it may generate infeasible results like broken arms. Moreover, the lack of semantic information limits their applications like garment transfer.

Expanding the representation ability of the statistical model of body shape is another solution. Several prior works [7, 38, 3, 4] utilize the vertex displacements of body shape represented by SMPL to represent garment geometry. Under this configuration, tight garments can be reconstructed. However, this representation cannot recover the feature of garment edges. More importantly, binding garments with SMPL points causes the problem that garments have the same skinning weights and connectivity with SMPL. Therefore, large scale displacements of loose garments may cause artifacts because of inappropriate skinning weights. More importantly, garments like skirts which have a different topology with body shape, are beyond the representation range.

Like Bhatnagar *et al.* [7], we train a model to reconstruct body mesh and layered garment meshes separately. The difference in input is that our method only requires a single RGB image and no additional semantic information and body rough A-pose constrain. Another difference is that our garment mesh is not bound with the body mesh, and can reconstruct more garment categories. To this end, we address three major challenges: learning a shared skinning weight network for all garments, garment detail inference, and dataset construction. Our method supports six garment categories, including upper garment, pants, and skirts with short and long templates for each type. For all garment types, we train a network to predict skinning weights related to SMPL's skeleton. For each type, we use graph convolution to recover the details. To train the model, a dataset with various RGB images and their corresponding body and cloth shapes is needed. However, there is no available public datasets that satisfy our demands. Instead, for each type of garment, we design different sizes of clothes dressed on different SMPL neutral bodies and repose these clothes to various postures utilizing a physics engine. Besides, a commercial 3D human dataset with high-definition texture is added to increase the diversity of training data.

Our method can infer both body and garment shapes from a single image with different poses, and also supports loose garment types, like skirts. Based on the reconstruction results by our method, applications like garments and poses transferring between different images can be achieved. In summary, the contributions of this work include the following aspects:

- We present a novel garment representation on top of SMPL and a neural network-based method to reconstruct the shapes of body and garment from a single near-front viewpoint color image.
- Rather than binding the skinning weight of garment with body mesh, we propose a generic skinning weights generating network, which enables our approach to support garments with different topologies.
- We design a complete algorithm pipeline for dressed SMPL body data construction with different types of garments. The constructed dataset, including synthetic images and clothed body shapes, will be publicly available.

2 Related Work

Template-Free Clothed Human Estimation. Some non-parametric methods based on voxel or implicit function have been proposed to address the complex topology of garments. BodyNet [46] directly infers a voxel representation of clothed bodies with a deep network. Due to the large memory cost for high resolution, high-frequency details are often missed. Jackson *et al.* [20] reconstruct the shape of humans via volumetric regression and show the ability to output fine-scale details. Zheng *et al.* [57] infer clothed body volume representation with an initial aligned SMPL body, and combine image features to enhance reconstruction details. Natsume *et al.* [35] propose a reconstruction method based on a multi-view framework using synthesizing new silhouettes from a single image. More recently, [43] proposes a promising clothed body reconstruction network using a memory-efficient implicit representation. Template-free methods do not utilize the human body prior to obtain complex topology modeling ability, at the cost of lacking semantic information and control of reconstructed results.

Template-Based Clothed Human Estimation. Based on human body statistical model [32, 5, 21], many works can estimate naked body shape from image [23, 27, 38, 36, 9, 14, 54]. For better representation ability, a displacement vector is added for each vertex. [3, 1, 2] adopt this strategy to reconstruct clothed body with skin-tight garment. Alldieck *et al.* [4] estimate detailed normal and vector displacement on the UV map, which leads to finer-scale details. Zhu *et al.* [59] model fine-scale details by adding free-form 3D deformation on top of parametric model. Instead of using a single surface to represent both garment and body, [7] separates SMPL mesh to represent upper garment and pant independently, leading to more flexible control. However, binding garment vertices to body model strictly restricts the topology of support garment categories, and it is hard to represent more loose garment types, such as skirts. [39, 55] also use separate body and garment templates to register clothed body motion sequences.

Garment Dataset Construction. BUFF [56] supplies high-quality 4D scans of clothed bodies, but it only has 5 subjects and 2 suits for each subject. Lahner *et al.* [26] collect high-quality 4D scans of garments, but the method leaves out body reconstruction, and their dataset is not publicly available. Recently, [7] constructs a training dataset with garment and body shapes from real scan data, but the training dataset is also unavailable. Moreover, many prior works generate ground truth dataset based on physics-based simulation [29, 28, 44, 49, 12, 17, 40]. [29, 28] dress SMPL bodies and construct more truthful images than SURREAL [47]. [49] simulates three types of garment and dress them on neutral SMPL bodies to learn garments design from sketches. All mentioned datasets do not meet our requirements. Therefore, we build a dataset containing a variety of garments and body types with different sizes and postures.

Garment Deformation Representation. How to represent the deformation of a garment is also related to our work. De Aguiar *et al.* [13] represent the garment dynamic dressed on a specific virtual avatar with a linear combination of pre-computed multiple deformations. DRAPE [17] regresses garment deformation from body shape with a technique derived from SCAPE [5]. Xu *et al.* [50]

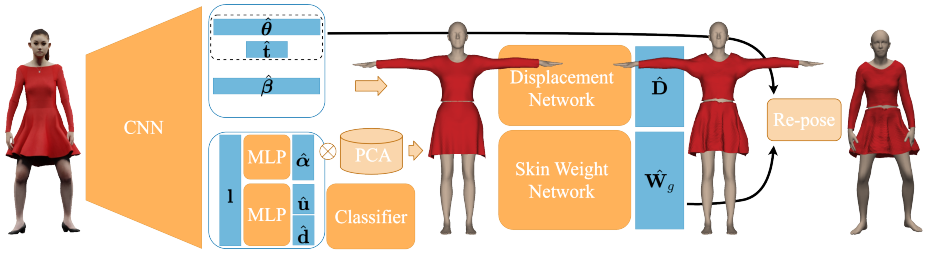

Fig. 1. The architecture of our proposed network. The CNN encodes image into latent feature, then we get reconstructed SMPL parameters $\hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\theta}}$, $\hat{\mathbf{t}}$ and shared garment latent feature $\mathbf{l}$ with respective FC layers. From $\mathbf{l}$, we reconstruct garment shape parameter $\hat{\boldsymbol{\alpha}}$ and garment type scores $\{\hat{\mathbf{u}}, \hat{\mathbf{d}}\}$ for upper and lower garment separately. With the classifier, $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}$, we reconstruct neutral clothed body. Followed a displacement network and skinning weight network, we predict garment vertex displacements and skinning weights separately. Finally, utilizing predicted pose parameters $\hat{\boldsymbol{\theta}}$, $\hat{\mathbf{t}}$ and $\hat{\mathbf{W}}_g$, we re-pose neutral body and garments with displacements to reference posture.

combine rotation and translation weights to approximate the non-local and non-linear clothing deformation and introduce a pose sensitive rigging scheme. Lahner *et al.* [26] recover high-frequency garment details from a normal map created from Generative Adversarial Network. Yang *et al.* [51] model garments with different connectivity based on a body template and use PCA to parameterize garment deformation. Santesteban *et al.* [44] propose to deform base garment conditioned on body parameters and then add high-frequency wrinkles.

3 Algorithm

The target of this work is to automatically reconstruct both body and cloth shapes from a single near-front view image. Our model currently supports six garment categories and can be easily extended to other new types. In the following, we first describe our garment representation model. Then, we introduce our network structure and training loss design.

3.1 Garment Model

We use SMPL [32] as our parametric human body model. SMPL is a function which maps shape parameters $\boldsymbol{\beta} \in \mathbb{R}^{10}$ and pose parameters $\boldsymbol{\theta} \in \mathbb{R}^{72}$ to a body mesh $\mathbf{M}_b(\boldsymbol{\beta}, \boldsymbol{\theta}) \in \mathbb{R}^{3|\mathcal{V}_b|}$, where $\mathcal{V}_b$ is SMPL mesh vertices set. The mapping can be summarized as the following equation:

$$\mathbf{M}_b(\boldsymbol{\beta}, \boldsymbol{\theta}) = W(\mathbf{T}_b(\boldsymbol{\beta}, \boldsymbol{\theta}), \mathbf{J}(\boldsymbol{\beta}), \boldsymbol{\theta}, \mathbf{W}_b), \quad \mathbf{T}_b(\boldsymbol{\beta}, \boldsymbol{\theta}) = \mathbf{B} + \mathbf{B}_s\boldsymbol{\beta} + \mathbf{B}_p\boldsymbol{\theta}, \tag{1}$$

where SMPL applies linear displacement bases $\mathbf{B}_s$ and $\mathbf{B}_p$ on a T-posed template mesh $\mathbf{B}$, and then utilize standard skeleton skinning operation W to get posed

body mesh. $\mathbf{J}(\boldsymbol{\beta}) \in \mathbb{R}^{24\times 3}$ is SMPL body's neutral skeleton and $\mathbf{W}_b \in \mathbb{R}^{|\mathcal{V}_b|\times 24}$ is the skinning weights of each vertex of SMPL.

As most clothes follow the deformations of the body, we compute our garment mesh $\mathbf{M}_g \in \mathbb{R}^{3|\mathcal{V}_g|}$ similarly based on the skin deformation of SMPL:

$$\mathbf{M}_g(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{D}) = W(\mathbf{T}_g(\boldsymbol{\alpha}, \mathbf{D}), \mathbf{J}(\boldsymbol{\beta}), \boldsymbol{\theta}, \mathbf{W}_g(\boldsymbol{\alpha}, \boldsymbol{\beta})), \quad \mathbf{T}_g(\boldsymbol{\alpha}, \mathbf{D}) = \mathbf{G} + \mathbf{B}_g\boldsymbol{\alpha} + \mathbf{D}. \tag{2}$$

For each garment category, a T-posed template mesh $\mathbf{G}$ is defined. On top of the base mesh, we add linear displacement deformation $\mathbf{B}_g$ controlled by PCA coefficients $\boldsymbol{\alpha} \in \mathbb{R}^{64}$. This low dimensional representation is effective in capturing size variations of a specific garment category under T-pose. To deform garments with dressed SMPL body, we share garment pose parameter $\boldsymbol{\theta}$ with SMPL and use SMPL's skeleton $\mathbf{J}(\boldsymbol{\beta})$ as the binding skeleton of the garment. Instead of directly using the skinning weights of SMPL, a neural network is utilized to estimate the skinning weights $\mathbf{W}_g$ of the garment. This design makes garment mesh independent with SMPL mesh and makes our garment model can support more garment topology than SMPL+D methods [7, 38, 3], if providing corresponding garment training data. To capture variations caused by different pose and interaction between clothing and body, we add a high-frequency displacement $\mathbf{D} \in \mathbb{R}^{3|\mathcal{V}_g|}$ for vertices of the clothing. In this paper, for the conciseness of writing symbol, we denote the displacement directly as $\mathbf{D}$ instead of a function of latent dependent variables, such as $\boldsymbol{\alpha}, \boldsymbol{\theta}$.

3.2 Image to Dressed Body

Given a near-front view RGB image depicting a posed subject dressed on specific garments, our model estimates its body shape, pose parameters and global translation with $\hat{\boldsymbol{\beta}} \in \mathbb{R}^{10}, \hat{\boldsymbol{\theta}} \in \mathbb{R}^{72}, \hat{\mathbf{t}} \in \mathbb{R}^3$ and the garment parameters $\hat{\boldsymbol{\alpha}} \in \mathbb{R}^{64}$ and $\hat{\mathbf{D}}$. Our model mainly consists of four modules: image encoder, classification module, skinning weight network, and displacement network. Fig. 1 shows our algorithm pipeline, and we will discuss the details of the last two modules.

Our image encoder uses the feature extraction of ResNet-18 [19] and average pooling the final feature map to 8×8 size. From the map, a fully connected layer is used to get the latent feature. Then, four fully connected layers are used to predict shape parameters $\hat{\boldsymbol{\beta}}$, pose parameters $\hat{\boldsymbol{\theta}}$, translation $\hat{\mathbf{t}}$ and shared garment latent feature $\mathbf{l} \in \mathbb{R}^{256}$. For pose parameters, instead of directly predicting the axis-angle representation parameters $\hat{\boldsymbol{\theta}}$, we predict vectorized rotation matrices $R(\hat{\boldsymbol{\theta}}) \in \mathbb{R}^{24\times 9}$ of all joints, where R is the Rodrigues rotation transformation. This strategy makes training more stable and continuous [27, 37, 38].

From shared garment latent $\mathbf{l}$, two fully connected layers are used to predict upper and lower garment classify scores $\hat{\mathbf{u}} \in \mathbb{R}^2$ and $\hat{\mathbf{d}} \in \mathbb{R}^4$ separately. Then, we concatenate $\hat{\boldsymbol{\beta}}$ and $\mathbf{l}$ as input of a two-layer Multi-layer perceptron(MLP) [41] to predict neutral garment shape parameters $\hat{\boldsymbol{\alpha}}$. After that, utilizing skinning weight and displacement networks, we get garment skinning weights $\hat{\mathbf{W}}_g$ and high-frequency displacements $\hat{\mathbf{D}}$, respectively. Finally, with predicted pose parameters, we can reconstruct the body shapes and dressed garments together.

3.3 Skinning Weight Network

It is an open problem to estimate skinning weights for an arbitrary character given a binding skeleton hierarchy. Recently, Liu *et al.* [30] proposed the first generic network to infer the skinning weights of various characters binding to the mutative skeleton hierarchy. Inspired by [30], we design our skinning weight network to infer weights for neutral garments, and the network makes weights computation fast, differentiable and garment type independent.

Our network predicts the skinning weights of a specific neutral garment $\mathbf{T}_g(\hat{\boldsymbol{\alpha}}, \mathbf{0})$ binding to the skeleton $\mathbf{J}$ of corresponding neutral SMPL body $\mathbf{T}_b(\hat{\boldsymbol{\beta}}, \mathbf{0})$. We compute all distances of each vertex of $\mathbf{T}_g(\hat{\boldsymbol{\alpha}}, \mathbf{0})$ to each joint point of $\mathbf{J}$. Then, the coordinate, normal, and distances of each vertex of $\mathbf{T}_g(\hat{\boldsymbol{\alpha}}, \mathbf{0})$ are concatenated as the input feature for the network, and it computes the weights for all vertices. Our network uses MLP to change the vertex feature dimension and utilizes standard Residual Block [19] to extract features. Besides, we use graph convolution to aggregate the neighborhood information. In order to make our network applicable to different garment categories, we use GAT [48] graph convolution, whose filter weight learning is independent of mesh connectivity, and the weight is determined by the input feature on vertices only. This characteristic makes our network based on GAT suitable for different garment types. The architecture details can be found in the supplementary.

3.4 Displacement Network

The shape structure of the garment can be well reconstructed based on the PCA coefficients $\boldsymbol{\alpha}$. However, high-frequency details, such as folds caused by different pose, are beyond the representation ability of the linear model. We train a displacement network to regress the displacement of each garment vertex on top of the base mesh. For the displacement, we use a similar network structure with the skinning weight network. To improve the regression ability, we train an independent network for each garment category rather than a general network for all types. Moreover, we use spiral graph convolution [10] for each garment category, which has state-of-the-art regression ability for meshes with the same connectivity. To capture high frequency information, we project each vertex of deformed base garment $\mathbf{M}_g(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}}, \mathbf{0})$ on the image, and crop the 32×32 patch centered on the projected vertex. Then, for each vertex, we use a shared MLP to encode its patch into a latent feature, and concatenate the feature with shared garment latent $\mathbf{l}$, predicted SMPL shape parameter $\hat{\boldsymbol{\beta}}$, garment shape parameter $\hat{\boldsymbol{\alpha}}$ as well as its coordinate, normal and skinning transformation together as its input feature for the displacement network. The details of the neural network are given in the supplementary.

3.5 Loss Function

With our constructed dataset, ground truth shape and pose parameters are available for all training data, thus it is natural to adopt supervised training.

In this part, we denote predicted $\mathbf{M}_g(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}}, \hat{\mathbf{D}})$ and $\mathbf{M}_b(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}}))$ as $\hat{\mathbf{M}}_g$ and $\hat{\mathbf{M}}_b$ separately. In the following, we will give the details on how to design the loss terms.

Losses on shape parameters. We directly adopt the MSE between predicted and ground truth shape parameters. The loss for SMPL body parameters and garment parameters are separately defined as:

$$L_{Bp} = \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2^2 + \|R(\hat{\boldsymbol{\theta}}) - R(\boldsymbol{\theta})\|_2^2 + \|\hat{\mathbf{t}} - \mathbf{t}\|_2^2, \quad L_{Gp} = \|\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}\|_2^2. \tag{3}$$

Losses on geometry. We supervise reconstructed geometries and joints with ground truth data. J_B is the mapping to output posed 3D joints of SMPL body $\mathbf{M}_b$.

– Losses on reconstructed garment geometry and reconstructed body joints are separately defined as:

$$L_G = \|\hat{\mathbf{M}}_g - \mathbf{M}_g\|_2^2, \quad L_{J3D} = \|J_B(\hat{\mathbf{M}}_b) - J_B(\mathbf{M}_b)\|_2^2. \tag{4}$$

– Losses on displacements $\mathbf{D}$. To improve detail reconstruction ability, we use ℓ_1 loss for each vertex of $\mathbf{D}$ and ℓ_2 loss on laplacian coordinates of $\mathbf{D}$. $\mathcal{L}$ represents the laplacian coordinates mapping from a 3D mesh.

$$L_{D1} = |\hat{\mathbf{D}} - \mathbf{D}|, \quad L_{D2} = \|\mathcal{L}(\hat{\mathbf{D}}) - \mathcal{L}(\mathbf{D})\|_2^2. \tag{5}$$

Losses of projection. We use Π to represent the camera projection of 3D geometries. All our training data share a common camera intrinsic matrix. The loss of body projections and garment projections are separately defined as:

$$L_{B2D} = \|\Pi(\hat{\mathbf{M}}_b) - \Pi(\mathbf{M}_b)\|_2^2, \quad L_{G2D} = \|\Pi(\hat{\mathbf{M}}_g) - \Pi(\mathbf{M}_g)\|_2^2. \tag{6}$$

Losses of classification. We use standard softmax loss to penalize the classification error of $\hat{\mathbf{u}}$ and $\hat{\mathbf{d}}$ relative to ground truth garment types.

Losses of interpenetration. During training, inferred garments and body are easy to occur interpenetration. We use a simple yet effective interpenetration term inspired by [18] to alleviate this problem:

$$L_{int}(\mathbf{P}, \mathbf{Q}) = \sum_{\{i,j\} \in \mathcal{C}(\mathbf{P}, \mathbf{Q})} ReLU(-\mathbf{n}_{\mathbf{q}_j}^T(\mathbf{p}_i - \mathbf{q}_j))/N, \tag{7}$$

where $\mathbf{P}$,$\mathbf{Q}$ are two interpenetrated meshes. $\mathcal{C}(\mathbf{P}, \mathbf{Q})$ represents the valid corresponding pairs between $\mathbf{P}$ and $\mathbf{Q}$, and these pairs are filtered based on distances and normal angles. This loss penalizes vertex $\mathbf{p}_i$ that is inside the local plane defined by its corresponding point $\mathbf{q}_j$ and its normal $\mathbf{n}_{\mathbf{q}_j}$. We use this loss on reconstructed neutral garments and body as well as posed garments and body separately:

$$L_{inters} = L_{int}(T_g(\hat{\boldsymbol{\alpha}}, \mathbf{0}), T_b(\hat{\boldsymbol{\beta}}, \mathbf{0})) + L_{int}(\hat{\mathbf{M}}_g, \hat{\mathbf{M}}_b) \tag{8}$$

Loss of Skinning Weight Network. As discussed in [30], the weight vector $\{\omega_{ij}|j \in |J(\boldsymbol{\beta})|\}$ of $\mathbf{W}_g$ is a selection of different bones with different probabilities. We use the Kullback-Leibler divergence loss to measure the distance between predicted weights distribution $\hat{\omega_{ij}}$ and ground truth distribution ω_{ij}:

$$L_{ws} = \sum_{i=1}^{|\mathcal{V}_g|} \sum_{j=1}^{24} \hat{\omega}_{ij} (\log \frac{\hat{\omega}_{ij}}{\omega_{ij}}). \tag{9}$$

To train the whole network, we first train the skinning weight network with loss in Eq. (9), and then train other parts together by fixing the skinning weight network.

4 Dataset Construction

4.1 Skinning Weight Dataset

To train our skinning weight network, we need some neutral garments with ground truth skinning weights. Our network training adapts to any weight calculation method. For simpleness, we compute garment weights from the dressed SMPL body.

For vertex $\mathbf{p}_i$ of the garment, we select K vertices from dressed body mesh, based on distance, normal angle, and segmentation prior. Segmentation prior is some information we can utilize, such as corresponding vertices of right trouser leg must belong to the right leg of body mesh. From selected K vertices of body, we average their skinning weights with IDW(inverse distance weighting) as the skinning weight of $\mathbf{p}_i$. After all vertices' weights have been computed, we apply Laplacian smoothing [45] to remove noises and artifacts.

With this method, for all garment types, we construct a skinning weights dataset, which includes 48000 neutral garments for training, and 6467 for test.

4.2 Synthetic Dataset Construction

As there does not exist publicly available dataset containing pairs of the color image and corresponding body and cloth shapes, we construct the dataset with a physics-based simulation method. The dataset construction process can be divided into four steps: sewing pattern design, neutral garment synthesis, posed garment simulation and rendering. [49] proposed a novel method to synthesize neutral garments. We extend their method to support more garment types and posed garments generation.

As shown in Fig. 2, we first design the pant sewing pattern based on body type. Then, around the neutral skeleton, we connect the sewing lines of the front and back pattern and shorten the length gradually. The sewing lines are stitched together after all lengths of the sewing line are less than a threshold. To simulate the realistic result of the garment draped on the neutral body, we inflate the skeleton and add gravity. For posed garment simulation, we deform the neutral

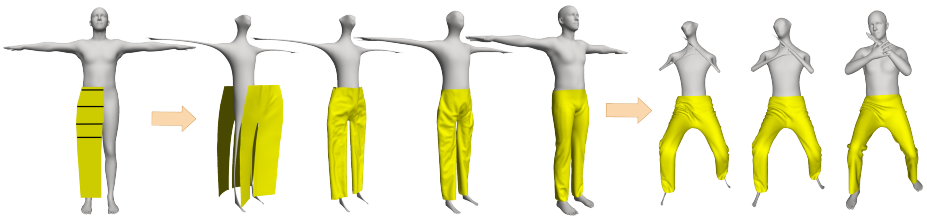

Fig. 2. Our synthesis process of a pant. First, we generate a random sewing pattern based on neutral body type. Then, we stitch the pattern on the skeleton and inflate the skeleton to its original shape to generate the neutral pant. Finally, we skinning deform the skeleton and neutral pant to a posture, and simulate the final pant with gravity by inflating the posed skeleton.

garment to target pose with generated skinning weights and inflate the body and add gravity to simulate the posed garment. In this work, we assume that both the human body and the garment are in a statically stable state. Therefore, we sample discrete pose instead of simulating the whole motion sequence.

After generating the garment shapes, the synthetic images are rendered by following the methods in [29, 28, 47]. By randomly selecting body textures from SURREAL [47], garment textures from Fabrics [22] and DTD [11], background images from Places365-Standard dataset [58] and global illumination from hundreds HDR images, we can render near-front view dressed body images with abundant variations.

Fig. 3. Some examples from our synthetic dataset(left three) and HD texture dataset(right three).

We implement the abovementioned pipeline using the simulator NvFlex[15] and Blender[8]. We utilize 3048 body shape of SPRING dataset [53], and randomly generate neutral clothes dressed on them. For posed garment, we select 55 motion sequences from CMU Mocap [34], whose poses have been converted to SMPL pose parameters with MoSh [31]. For each motion sequence, we randomly select 10 different persons with 4 sets of different clothing separately and sample pose parameters every 30 frames. Finally, we get 168602 dressed bodies as training data and 8874 as test data. The left part of Fig. 3 shows several examples of our synthetic images.

Fig. 4. Two examples of rigged avatar registration. We show the scanned meshes with and without texture, reconstructed geometries and garment with texture in each group.

4.3 HD Texture Dataset

Although synthetic samples are visually realistic, they still have a noticeable domain difference with real images. Therefore, we process another dataset with high-definition (HD) textures. We purchase 104 and 181 rigged avatar from RenderPeople [42] and Axyz [6], respectively. These avatars have high-quality geometry and realistic texture. We use Mixamo [33] to drive avatars and get about 89425 posed meshes as training data and 4386 as test data. The abovementioned rendering pipeline is used to produce high-quality images, and the right part of Fig. 3 shows some examples. Because the body and clothes part of the scanned mesh are not separated, and the connectivities of scanned meshes are not consistent, we need to process these meshes to our representation via the following two steps.

Rigged registration. For a rigged mesh with A- or T-pose, we segment it to garment and skin parts. We optimize garment shape parameters $\boldsymbol{\alpha}$, displacements $\mathbf{D}$, body parameters $\boldsymbol{\beta}$, $\boldsymbol{\theta}$ and translation $\mathbf{t}$ to register our representation to the avatar. We penalize the point-to-plane distance for both reconstructed garment and body. And we use Eq. (7) to reduce the interpenetration among them. To get a size matching garment, we adopt the rendered silhouette loss utilizing [25]. And we add ℓ_2 regular term for garment and body parameters. With this pipeline, we reconstruct all garments and body shapes of rigged avatars, and we extract texture for each garment. Fig. 4 shows two examples.

Posed registration. After we finish the rigged avatar reconstruction, we initialize our posed model optimization with rigged reconstruction parameters and optimize pose parameters $\boldsymbol{\theta}$ and translation $\mathbf{t}$ first. And then, we fine-tune all parameters to get final posed reconstruct results.

5 Experiments

In this part, we first evaluate our BCNet. Then we quantitatively compare with state-of-the-art methods. Finally, we present some qualitative results. More results are supplied in the supplementary.

Fig. 5. The left shows the ablation study for interpenetration loss. The examples demonstrate that the interpenetration term in Eq. (7) alleviates the collision problem. The right shows our predicted displacement results. For each example, We present the base mesh, mesh with displacement and reposed mesh, respectively. The first example captures detail geometry on top of the base mesh, and the second one recovers large scale deformation for the skirt caused by leg movement. For better visualization, we show two viewpoints for the second result.

5.1 Analysis of BCNet

Our Test Set. We test our predicted errors on Synthetic and HD Texture test set, respectively. Table 1 shows the mean Euclidean distance(MED) of reconstructed shapes after Procrustes transformation and ground truth shapes.

Skinning Weight Network. We test the reconstruction ability of the skinning weight network on our test dataset. For each garment, we reconstruct its skinning weights with our network. The average ℓ_1 reconstruction error on the whole test set is 6.5×10^{-4}. Then, we sample 20 poses from the Mocap dataset and deform the neutral clothes to the posture with our predicted weights and ground truth weights separately. The average MED of reposed mesh for all garment types is 0.43mm. These results demonstrate that our skinning weight network can reach very high accuracy. More details are given in supplementary.

Interpenetration. Our network infers human body and layered garments mesh separately, which brings better flexibility but at the cost of introducing more complex interactions between body and garments. Interpenetration is a common unreal phenomenon which is very easy to perceive by a human. Therefore, it is quite necessary to process interpenetration between these meshes. We propose an interpenetration term in Eq. (7) to alleviate this problem, and an ablation study on this term is shown in the left of Fig. 5. We can see that the interpenetration loss is beneficial to alleviate the interpenetration problem.

Displacement Network. Garment PCA shape parameter $\boldsymbol{\alpha}$ can represent the garment structure, while it can not represent the detailed shape of a

Table 1. The MED(cm) between predicted and ground truth shapes on our test dataset. For garments, we report errors with(gray) and without(white) displacement module, respectively.

dataset	shirt	pant	skirt	body
Synthetic	0.91	0.75	0.87	1.57
	1.72	1.59	2.46	
HD Texture	1.71	1.42	1.65	2.93
	1.97	1.72	1.87	

Table 2. The errors(cm) on BUFF rough A-pose dataset(gray) and Digital Wardrobe dataset(white).

Methods	Upper	Lower	Total	Chamfer
MGN-opt-8	1.63	1.91	1.82	1.91
MGN-8	1.78	2.13	1.99	2.08
Octopus-opt-8	1.40	**1.35**	**1.31**	1.41
Octopus-8	1.54	1.74	1.70	1.76
Ours	**1.07**	**1.35**	1.35	**1.34**
PIFu	1.59	**1.37**	1.85	**1.61**
DeepHuman	2.38	2.46	3.15	2.98
Ours	**1.44**	1.78	**1.80**	1.77

specific garment and large scale deformations caused by pose and gravity for loose garments. We train our non-linear displacement network to expand the representation ability. The result of ablation study on displacement network is given in Table 1, we can observe that the displacement network greatly improves the reconstruction accuracy. In the right part of Fig. 5, we show two examples of our displacement results. We present input image, base garment $\mathbf{T}_g(\boldsymbol{\alpha}, \mathbf{0})$, the garment with predicted displacement $\mathbf{T}_g(\boldsymbol{\alpha}, \mathbf{D})$ and final reposed garment $\mathbf{M}_g(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{D})$ for each result. In the first row, we show an example of predicted displacement capturing detailed geometry, such as tie and suit boundary line. In the second row, we show large scale deformations on a skirt caused by bending leg motion, and we use two viewpoints to show the deformation results.

5.2 Quantitative Comparison.

We test our reconstruction accuracy on two public data sets, BUFF [56] and Digit Wardrobe(DW) [7]. We segment the ground truth scan mesh into upper, lower garment and body parts, and compute error for garments and whole clothed bodies separately. Because our model predicts separate body and garment meshes, we extract the outer surface of all meshes as the proxy to do registration and error computing for a fair comparison. We measure the average point-to-surface Euclidean distance(P2S) in cm from the ground truth to predicted surface for upper, lower garments, and the whole surface. We also compute the Chamfer distance [43] between the reconstructed and the ground truth surfaces.

BUFF Dataset. We compare the reconstruction accuracy of our method with SMPL+D based methods octopus [38] and MGN [7]. By default, their methods require multi-view semantic segmentation images and 2D joints of a clothed body under rough A-pose as inputs, and post-optimization is applied to refine the results. Therefore, we select 21 rough A-pose data from BUFF [56] as our test set. Table 2 shows our results, and their results of 8 perspective inputs with and without optimization, respectively. Although the input of our method only needs one image, our method can get better numerical results than theirs without post-optimization, and an equivalent result with Octopus with optimization. The post-optimization is time-consuming and takes several

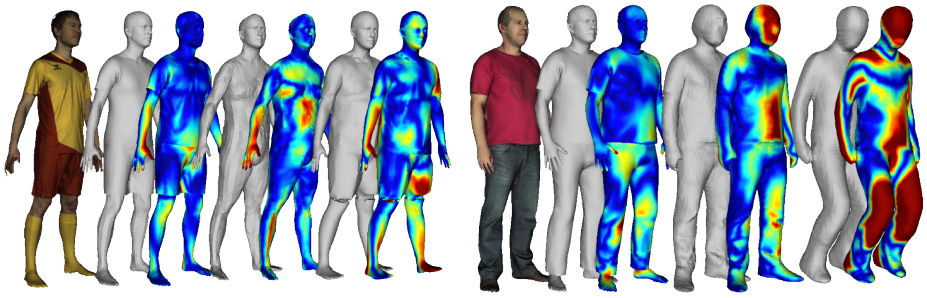

Fig. 6. Error maps on BUFF(left part) and MGN(right part). From left to right, we show the GT mesh, results of ours, Octopus-opt-8, MGN-opt-8 for the BUFF example, and results of ours, PIFu, DeepHuman for the MGN example(red means $\geq$ 4cm).

seconds and several minutes for Octopus and MGN, respectively. For MGN, we manually modified some segmentation error of PGN [16] to refine their results. In the left part of Fig. 6, we show an example of our result and their results with post-optimization. Some unnatural folds are introduced in the post-optimization step of MGN while our method does not have this problem.

DW dataset. Digital Wardrobe [7] includes registered clothed body meshes with real texture under more general posture. We use 94 meshes to compare with non-parametric methods PIFu [43] and DeepHuman [57][3]. Table 2 shows the results. For PIFu with single image input, our method can get similar reconstruction accuracy. However, the reconstructed results by PIFu combine both shapes in one mesh without semantic information, while our method can fully control the predicted separate body and cloth meshes. The results of DeepHuman tend to bend the leg, which introduces large errors for this dataset. The right part of Fig. 6 shows an example of the results.

5.3 Qualitative Results.

In this following, we show some visual results of our method and the comparison with MGN. As our method can reconstruct the body and garments separately, garment transfer between two input images can be achieved. Some garment transfer results are given in the supplementary.

Reconstruction Quality. In the left part of Fig. 7, we present our reconstructed body and garments shapes on several test images. Our method can recover accurate body posture and capture the garment geometry to some extent from a single input image. Thanks to our separated garment representation with adaptive skinning weights, we can reconstruct plausible shape for loose garments with large edges.

Comparison with MGN. As a template-based method, MGN [7] is the most relevant prior method with ours. MGN represents garment by binding the

[3] We did not test [38, 7] on this dataset as most of the samples are not A-pose.

Fig. 7. The left part: reconstructed body and garment shapes by our method on four images of our test set. The right part: qualitative comparison between the results of MGN [7] without post-optimization and ours. In each group, the input image, result of MGN, and ours are displayed respectively.

garment to SMPL vertices and uses a mask to select valid vertices for a specific garment type. MGN needs multi semantic segmentation images as input and constrains the posture to rough A-pose. Besides, MGN needs a time-consuming post-optimization step to refine the predicted result. Differently, Our method only requires one frontal view image with arbitrary posture and directly produces the final results from the network. In the right part of Fig. 7, we show two qualitative comparisons with MGN. Our method can generate more accurate body shape and size of garments, while the results of MGN without post-optimization have similar shapes for different inputs and lack garment details.

6 Conclusion

We introduced BCNet, a novel method to automatically reconstruct both body and garment shapes from a single RGB image. Rather than binding garment with SMPL like prior SMPL+D based representation, our proposed model can produce layered garments with different topology and skinning weights, which makes BCNet a model capable of jointly reconstructing body and loose garment, like skirts. To train BCNet, we designed a complete algorithm pipeline to generate clothed body data. Experiments demonstrated that our method can generate comparable or better reconstruction results compared with state-of-the-art methods, while allowing more flexible controls such as garment transfer. Our constructed dataset and our proposed BCNet would push a step for the research on digitizing human.

References

1. Alldieck, T., Magnor, M., Bhatnagar, B.L., Theobalt, C., Pons-Moll, G.: Learning to reconstruct people in clothing from a single rgb camera. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1175–1186 (2019)
2. Alldieck, T., Magnor, M., Xu, W., Theobalt, C., Pons-Moll, G.: Detailed human avatars from monocular video. In: International Conference on 3D Vision (3DV). pp. 98–109. IEEE (2018)
3. Alldieck, T., Magnor, M., Xu, W., Theobalt, C., Pons-Moll, G.: Video based reconstruction of 3d people models. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8387–8397 (2018)
4. Alldieck, T., Pons-Moll, G., Theobalt, C., Magnor, M.: Tex2shape: Detailed full human body geometry from a single image. In: IEEE International Conference on Computer Vision (ICCV) (2019)
5. Anguelov, D., Srinivasan, P., Koller, D., Thrun, S., Rodgers, J., Davis, J.: Scape: shape completion and animation of people. In: ACM Transactions on Graphics (TOG). vol. 24, pp. 408–416. ACM (2005)
6. axyz: 2019. https://secure.axyz-design.com/
7. Bhatnagar, B.L., Tiwari, G., Theobalt, C., Pons-Moll, G.: Multi-garment net: Learning to dress 3d people from images. In: IEEE International Conference on Computer Vision (ICCV) (2019)
8. Blender: 2019. https://www.blender.org/
9. Bogo, F., Kanazawa, A., Lassner, C., Gehler, P., Romero, J., Black, M.J.: Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In: European Conference on Computer Vision (ECCV). pp. 561–578. Springer (2016)
10. Bouritsas, G., Bokhnyak, S., Ploumpis, S., Bronstein, M., Zafeiriou, S.: Neural 3d morphable models: Spiral convolutional networks for 3d shape representation learning and generation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 7213–7222 (2019)
11. Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., Vedaldi, A.: Describing textures in the wild. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3606–3613 (2014)
12. Daněřek, R., Dibra, E., Öztireli, C., Ziegler, R., Gross, M.: Deepgarment: 3d garment shape estimation from a single image. In: Computer Graphics Forum. vol. 36, pp. 269–280. Wiley Online Library (2017)
13. De Aguiar, E., Sigal, L., Treuille, A., Hodgins, J.K.: Stable spaces for real-time clothing. ACM Transactions on Graphics (TOG) **29**(4), 1–9 (2010)
14. Dibra, E., Jain, H., Oztireli, C., Ziegler, R., Gross, M.: Human shape from silhouettes using generative hks descriptors and cross-modal neural networks. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4826–4836 (2017)
15. Flex, N.: 2019. http://https://developer.nvidia.com/flex/
16. Gong, K., Liang, X., Li, Y., Chen, Y., Yang, M., Lin, L.: Instance-level human parsing via part grouping network. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 770–785 (2018)
17. Guan, P., Reiss, L., Hirshberg, D.A., Weiss, A., Black, M.J.: Drape: Dressing any person. ACM Transactions on Graphics (TOG) **31**(4), 35–1 (2012)
18. Gundogdu, E., Constantin, V., Seifoddini, A., Dang, M., Salzmann, M., Fua, P.: Garnet: A two-stream network for fast and accurate 3d cloth draping. In: IEEE International Conference on Computer Vision (ICCV). pp. 8739–8748 (2019)

19. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778 (2016)
20. Jackson, A.S., Manafas, C., Tzimiropoulos, G.: 3d human body reconstruction from a single image via volumetric regression. In: European Conference on Computer Vision (ECCV). pp. 0–0 (2018)
21. Jiang, B., Zhang, J., Cai, J., Zheng, J.: Learning 3d human body embedding. arXiv preprint arXiv:1905.05622 (2019)
22. Kampouris, C., Zafeiriou, S., Ghosh, A., Malassiotis, S.: Fine-grained material classification using micro-geometry and reflectance. In: European Conference on Computer Vision (ECCV). pp. 778–792. Springer (2016)
23. Kanazawa, A., Black, M.J., Jacobs, D.W., Malik, J.: End-to-end recovery of human shape and pose. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7122–7131 (2018)
24. Kanazawa, A., Zhang, J.Y., Felsen, P., Malik, J.: Learning 3d human dynamics from video. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5614–5623 (2019)
25. Kato, H., Ushiku, Y., Harada, T.: Neural 3d mesh renderer. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3907–3916 (2018)
26. Lahner, Z., Cremers, D., Tung, T.: Deepwrinkles: Accurate and realistic clothing modeling. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 667–684 (2018)
27. Lassner, C., Romero, J., Kiefel, M., Bogo, F., Black, M.J., Gehler, P.V.: Unite the people: Closing the loop between 3d and 2d human representations. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6050–6059 (2017)
28. Liang, J., Lin, M.C.: Shape-aware human pose and shape reconstruction using multi-view images. In: The IEEE International Conference on Computer Vision (ICCV) (2019)
29. Liu, J., Akhtar, N., Mian, A.: Temporally coherent full 3d mesh human pose recovery from monocular video. arXiv preprint arXiv:1906.00161 (2019)
30. Liu, L., Zheng, Y., Tang, D., Yuan, Y., Fan, C., Zhou, K.: Neuroskinning: automatic skin binding for production characters with deep graph networks. ACM Transactions on Graphics (TOG) **38**(4), 114 (2019)
31. Loper, M., Mahmood, N., Black, M.J.: Mosh: Motion and shape capture from sparse markers. ACM Transactions on Graphics (TOG) **33**(6), 220 (2014)
32. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: Smpl: A skinned multi-person linear model. ACM Transactions on Graphics (TOG) **34**(6), 248 (2015)
33. Mixamo: 2019. https://www.mixamo.com/
34. Mocap, C.: 2019. mocap.cs.cmu.edu
35. Natsume, R., Saito, S., Huang, Z., Chen, W., Ma, C., Li, H., Morishima, S.: Siclope: Silhouette-based clothed people. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4480–4490 (2019)
36. Omran, M., Lassner, C., Pons-Moll, G., Gehler, P., Schiele, B.: Neural body fitting: Unifying deep learning and model based human pose and shape estimation. In: International Conference on 3D Vision (3DV). pp. 484–494. IEEE (2018)
37. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A., Tzionas, D., Black, M.J.: Expressive body capture: 3d hands, face, and body from a single image. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10975–10985 (2019)

38. Pavlakos, G., Zhu, L., Zhou, X., Daniilidis, K.: Learning to estimate 3d human pose and shape from a single color image. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 459–468 (2018)
39. Pons-Moll, G., Pujades, S., Hu, S., Black, M.J.: Clothcap: Seamless 4d clothing capture and retargeting. ACM Transactions on Graphics (TOG) **36**(4), 1–15 (2017)
40. Pumarola, A., Sanchez, J., Choi, G., Sanfeliu, A., Moreno-Noguer, F.: 3dpeople: Modeling the geometry of dressed humans. In: IEEE International Conference on Computer Vision (ICCV) (2019)
41. Qi, C.R., Su, H., Mo, K., Guibas, L.J.: Pointnet: Deep learning on point sets for 3d classification and segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 652–660 (2017)
42. Renderpeople: 2019. https://renderpeople.com/3d-people
43. Saito, S., Huang, Z., Natsume, R., Morishima, S., Kanazawa, A., Li, H.: Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In: IEEE International Conference on Computer Vision (ICCV) (2019)
44. Santesteban, I., Otaduy, M.A., Casas, D.: Learning-based animation of clothing for virtual try-on. In: Computer Graphics Forum. vol. 38, pp. 355–366. Wiley Online Library (2019)
45. Sorkine, O.: Laplacian mesh processing. In: Eurographics - State of the Art Reports. pp. 53–70 (2005)
46. Varol, G., Ceylan, D., Russell, B., Yang, J., Yumer, E., Laptev, I., Schmid, C.: Bodynet: Volumetric inference of 3d human body shapes. In: European Conference on Computer Vision (ECCV). pp. 20–36 (2018)
47. Varol, G., Romero, J., Martin, X., Mahmood, N., Black, M.J., Laptev, I., Schmid, C.: Learning from synthetic humans. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 109–117 (2017)
48. Velickovic, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., Bengio, Y.: Graph attention networks. In: International Conference on Learning Representations, ICLR (2018)
49. Wang, T.Y., Ceylan, D., Popovic, J., Mitra, N.J.: Learning a shared shape space for multimodal garment design. ACM Transactions on Graphics (TOG) **37**(6), 203:1–203:13 (2018)
50. Xu, W., Umentani, N., Chao, Q., Mao, J., Jin, X., Tong, X.: Sensitivity-optimized rigging for example-based real-time clothing synthesis. ACM Transactions on Graphics (TOG) **33**(4), 107 (2014)
51. Yang, J., Franco, J.S., Hétroy-Wheeler, F., Wuhrer, S.: Analyzing clothing layer deformation statistics of 3d human motions. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 237–253 (2018)
52. Yang, S., Pan, Z., Amert, T., Wang, K., Yu, L., Berg, T., Lin, M.C.: Physics-inspired garment recovery from a single-view image. ACM Transactions on Graphics (TOG) **37**(5), 170 (2018)
53. Yang, Y., Yu, Y., Zhou, Y., Du, S., Davis, J., Yang, R.: Semantic parametric reshaping of human body models. In: International Conference on 3D Vision (3DV). vol. 2, pp. 41–48. IEEE (2014)
54. Yao, P., Fang, Z., Wu, F., Feng, Y., Li, J.: Densebody: Directly regressing dense 3d human pose and shape from a single color image. arXiv preprint arXiv:1903.10153 (2019)
55. Yu, T., Zheng, Z., Zhong, Y., Zhao, J., Dai, Q., Pons-Moll, G., Liu, Y.: Simulcap: Single-view human performance capture with cloth simulation. arXiv preprint arXiv:1903.06323 (2019)

56. Zhang, C., Pujades, S., Black, M.J., Pons-Moll, G.: Detailed, accurate, human shape estimation from clothed 3d scan sequences. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4191–4200 (2017)
57. Zheng, Z., Yu, T., Wei, Y., Dai, Q., Liu, Y.: Deephuman: 3d human reconstruction from a single image. In: IEEE International Conference on Computer Vision (ICCV) (2019)
58. Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A.: Places: A 10 million image database for scene recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **40**(6), 1452–1464 (2017)
59. Zhu, H., Zuo, X., Wang, S., Cao, X., Yang, R.: Detailed human shape estimation from a single image by hierarchical mesh deformation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4491–4500 (2019)