The rapid construction method of human body model for virtual try-on on mobile terminal based on MDD-Net

Naiyu Fang¹ · Lemiao Qiu¹ · Shuyou Zhang¹ · Zili Wang¹ · Ye Gu¹ · Kerui Hu¹

Accepted: 16 August 2022 / Published online: 19 September 2022
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2022

Abstract

Traditional anthropometric evaluation needs professional measuring tools and operations, and it is time-consuming, expensive, and not suitable for virtual try-on. As the mobile internet develops, the issue of human body reconstruction toward virtual try-on needs to be tackled. This paper proposes a rapid human body reconstruction method for virtual try-on based on multidimensional dense net (MDD-Net) on mobile terminal. MDD-Net takes fusion features as input and predicts 3D human body model. The acquisition of fusion features and the display of 3D human body are implemented on mobile toward virtual try-on. In the learning fuzzy anthropometric feature module, the example-guided fuzzy anthropometric feature matrix is acquired and default coding elements are interpolated. In the learning multi-perspective silhouette feature module, the fine human body shape features are learned based on DenseNet201. A related fusion feature data set is generated for the training and testing of MDD-Net. Compared to shape pose estimation models, the shape representation spaces of HMR and SMPLify are only 20.34% and 7.59% of our method, and their prediction accuracies are approximately 50% of our method. Compared to accurate shape estimation models, our method is more robust against the pose and perspective noise. The prediction accuracies of our method are improved by 13.34%, 55.77%, 34.6%, and 43.4%, 37.2%, 9.0% on four test sets. Extensive experiments have demonstrated the superiority of our method for human body shape estimation toward virtual try-on.

Keywords Human body shape estimation · Fuzzy anthropometric feature · Multi-perspective silhouette feature · Virtual try-on · MDD-Net

1 Introduction

Three-dimensional human body reconstruction is widely used in virtual try-on (Jiang et al. 2019; Bhatnagar et al. 2019), character animation (Weng et al. 2019; Hornung et al. 2007), and virtual reality (Zhao et al. 2018). The reconstructed human body model could improve the immersion in virtual reality and the realism in online shopping. The previous 3D human body reconstruction relies on the complex coordinate measuring instrument (Wang et al. 2020; Chen et al. 2015) or RGBD camera (Yu et al. 2018; Xu et al. 2019; Rhodin et al. 2016), and mainly focuses on the shape-pose blended estimation. In the virtual try-on on mobile terminal, it is difficult to obtain the point cloud data, has high requirements on real-time performance and body shape representation. Thus, it has to solve how to simplify the human body features acquisition and fully represent the human body shape to apply 3D human body reconstruction into the virtual try-on on the mobile terminal.

The statistical human body model is a crucial basis for 3D human body reconstruction, and it represents different human body objects in terms of shape and pose. SMPL (Loper et al. 2015) is regarded as a representative of the statistical human body model. Based on SMPL, 3D human body reconstruction works toward shape-pose blended estimation or accurate shape estimation. And the acquired human body features and the application scenarios vary greatly with the 3D human body reconstruction intentions. Shape-pose blended estimation has been applied to animation and film, the labeled features (e.g., 2D joint point (Kanazawa et al. 2018), human body parts segmentation (Pavlakos et al. 2018)) are utilized to train networks, and joint angles and low-dimensional shape coefficients are predicted. The main target of shape-pose blended estimation is to display the overall motion state of wild images. Accurate shape estimation has been applied to...
virtual try-on, and the fined human body shape is estimated in the reconstructed human body model. The core content of accurate shape estimation is to predict low-dimensional shape coefficients by only acquiring the human body shape features. In the accurate shape estimation, the human pose is set to be the prior feature (e.g. A-pose or T-pose).

Two-dimensional images are the prime inputs of 3D human body reconstruction; this reconstruction essence is to reconstruct the model or predict the human body shape coefficients from 2D image features. Moreover, some researchers (Baek and Lee 2012; Wuhrer and Shu 2013; Zhang et al. 2015) reconstruct 3D human body models only from human body 1D features. The relationship between the anthropometric features and the 3D human body model coefficients is analyzed. Thus, in the 3D human body reconstruction, some specific human body part features need to be measured.

In the above researches, the shape-pose blended estimation is mainly applied to animation and film while is not suitable for virtual try-on. Its reconstructed human body model lacks detailed muscle representation and human body curve features, and the prior features of 3D human body reconstruction are hard to acquire on the mobile terminal. For the accurate shape estimation, to avoid the pose influence on human body shape representation, the silhouettes are acquired with a similar pose to the standard pose, and the network robustness against the pose noise is improved by learning a large number of silhouettes with various perspectives and segmentation noise (Smith et al. 2019). These increase the cost of acquiring human body shape features with a mechanically repetitive process. For the 3D human body reconstruction only with 1D measured features, measuring anthropometric features requires professional skills and tools (Li et al. 2019), and the 1D coding matrix size is also hard to define. If more human body parts measured values are required, the matrix is easy to defaults. And if fewer human body parts measured values are required, the reconstructed human body model maybe has a large error. Therefore, there is an urgent need for a convenient and fast accurate human body reconstruction method toward virtual try-on on mobile terminal.

This paper takes the fusion features as input and estimates the accurate human body shape with T-Pose. The fusion features consist of multi-perspectives silhouette features and example-guided fuzzy anthropometric features. MDD-Net learns global and local human body shape feature coding from fuzzy anthropometric features and multi-perspective silhouette, respectively. A fusion feature dataset based on SMPL is generated for network training. Extensive experiments demonstrate that our method has a wider shape representation space and a higher prediction accuracy than HMR and SMPLify. And our method is more robust than BiSNet and HS-Net against the pose and perspective noise.

In this paper, the outline of our method is introduced in Sect. 4.1. The training set, the modules, and the feature fusion method of MDD-Net are introduced in Sect. 4.2. The generating of the fusion feature dataset and four test sets based on the SMPL are introduced in Sect. 4.3. The ablation experiments on MDD-Net are introduced in Sect. 5.1. The qualitative and quantitative comparisons between our method and other models are introduced in Sect. 5.2. And the anthropometry evaluation is introduced in Sect. 5.3.

2 Related work

2.1 The statistical human body model

SMPL (Loper et al. 2015) is a shape-pose blended model based on skin vertices, the female and male shape-pose latent spaces of the CAESAR (Robinette et al. 2002) dataset are learned by principal component analysis (PCA), and an end-to-end model is generated. SCAPE (Anguelov et al. 2005) is a shape-pose blended model based on the triangular mesh, which has a worse shape representation ability, poor compatibility, and longer rendering time. In recent works, the fined face and hand parts are integrated into the statistical human body model (Joo et al. 2018; Xiang et al. 2019; Pavlakos et al. 2019), and more real human body model has been released to further expand the shape-pose latent space (Zanfir et al. 2020).

2.2 Shape-pose blended estimation

In the earlier research, the 3D human body reconstruction is reconstructed by aligning the joint points between the image and the template model. This aligning is time-consuming, while the human body model shape representation is neutral (Bogo et al. 2016; Lassner et al. 2017b; Guler and Kokkinos 2019). As CNN develops, aligning the sparse joint points is replaced by optimizing the pixel position, and the end-to-end models are trained to regress shape and pose coefficients directly from the image (Kanazawa et al. 2018; Popa et al. 2017; Tan et al. 2017). The prior features like 2D joint point and human body parts segmentation have been joined into training to further improve the reconstruction accuracy (Omran et al. 2018). In recent studies, inferring motion sequences from videos to achieve the motion simulation (Kocabas et al. 2020; Leroy et al. 2017; Tung et al. 2017) and monocular multi-person reconstruction (Omran et al. 2018) have become new hotspots, but the main content is still about overall pose estimation.
2.3 Accurate shape estimation

In the earlier researches, the single silhouette is utilized for shape coefficient regression with random forest (Dibra et al. 2016b), the 3D feature descriptors are also utilized to improve the shape reconstruction accuracy (Dibra et al. 2017). However, the single silhouette contains limited human body shape features. In recent studies, the multi-perspective silhouette is utilized to train the network and regress the shape coefficients, HS-Net (Dibra et al. 2016a) and BfSNet (Smith et al. 2019) could output the human body model with a more refined shape.

2.4 3D human body reconstruction from anthropometric features

Baek (2012) analyzes the correlation between the shape variation and the human body contour, and proposed a parameterized model that takes the human body parts measured values as inputs and outputs the corresponding human body model. Wuhrer (2013) predicts the human body shape from the encoded anthropometric values by nonlinear optimization, and the shape learning is not restricted by the shape latent space. Zhang (2015) analyzes the relationship between the measured example-guided features, and predicts 3D human body model by the radial basis interpolation and constraint-driven method.

3 Problem description

According to the input form, 3D human body reconstruction can be subdivided into: based on anthropometric feature; based on meshes by 3D scans; based on the image.

Three-dimensional human body reconstruction based on anthropometric features needs professional measuring tools (e.g., contact tapes) and operations. The measuring process usually lasts about 10 min (Li et al. 2019). Smart wearable device (Xu et al. 2018; Uhm et al. 2015) is a new helpful tool to achieve rapid and accurate human body parts measurement. However, these devices have not been widely applied and are inaccessible to the customer.

Three-dimensional human body reconstruction based on meshes needs high-resolution 3D scans (Chen et al. 2019; Bálan and Black 2008) or 4D scans (Pons-Moll et al. 2017; Zhang et al. 2017). The scans of the user rely heavily on the scanner (Zollhöfer et al. 2014) or multi-view vision system (Bogo et al. 2014; Rhodin et al. 2016; Elhayek et al. 2015). Although the reconstructed model has a high precision, manual labeling is usually required, the user must be in a specific measuring environment, which is costly, long waiting, and inconvenient.

In contrast, 3D human body reconstruction based on the images is easy to obtain the user shape feature, and the users can submit their photos anytime, anywhere. As the mobile internet develops, the new requirements for 3D human body reconstruction toward virtual try-on come in:

- The virtual try-on should be available anytime and anywhere, and it should be implemented on the mobile APP.
- The acquisition of human shape features should be fast and convenient. It should be independent of professional tools and complex labels. The ideal way is as Fig. 1a, b.
- The reconstructed 3D human body model should have a fine shape and should be displayed in real-time like Fig. 1c.

Thus, the method of fast convenient human body reconstruction toward virtual try-on needs to be proposed, it will promote more consumers to try on new garments in a relaxed environment and get self-identity with 3D human body model in virtual try-on. In this paper, the acquisition of the human body shape fusion features is analyzed, and the network of the human body shape estimation is proposed. The feature acquisition and model display are implemented on the mobile, and the network can be deployed on the remote server.

4 Human body shape estimation method

4.1 Outline

The user takes photos and submits fuzzy information using the mobile. The photos and fuzzy information are simply processed on mobile terminal to generate the multi-perspective silhouettes and fuzzy anthropometric features. Taking the fusion features as inputs, MDD-Net predicts the shape coefficients, and the corresponding 3D human body model is generated based on SMPL. The reconstructed human body model is downloaded from the cloud server, the user gets the model display in real-time on mobile terminal. The outline of our method is shown in Fig. 2.

4.1.1 Acquiring fuzzy anthropometric feature by mobile

In acquiring human body shape features, the user is not expected to measure precise values. With the interaction convenience of the mobile terminal, fuzzy options are provided to the user. The user merely needs to submit a fuzzy score by comparing with reference examples. The human body part examples of the two ends are corresponding to the extremum in human body shape space, and sufficient example references among extremum by linear interpolation are provided. Fuzzy anthropometric features $f = [c_1, c_2, ...c_i, l_i, l_{i-1}, ...l_1]$
are utilized as one of inputs of the 3D human body reconstruction network. The $c_i$ is the fuzzy circumference and $l_i$ is the fuzzy length, and the range of them is $[0, 100]$.

The selection of fuzzy anthropometric features is crucial. Under high prediction accuracy, it is better to acquire fewer features as much as possible. Therefore, the acquisition rank needs to be sorted, the feature with a large impact on human body shape will be should be acquired firstly. The import factor $I(f_\delta)$ of the coding element $f_\delta$ in the fuzzy matrix is defined as (1) where $\varepsilon_f, \varepsilon_f$ are the reconstruction accuracy with inputting complete feature matrix and partial the fuzzy anthropometric feature matrix without $f_\delta$, respectively.

$$I(f_\delta) = \frac{\varepsilon_f - \varepsilon_\delta}{\sum_{\delta} (\varepsilon_f - \varepsilon_\delta)}$$

When the acquired fuzzy anthropometric feature matrix is sparse, multiple linear regression can estimate default elements. However, the dimension of the input and output matrix for multiple linear regression are indefinite. So, the linear combination of one linear regression is utilized to solve the indefinite problem as (2).

$$f_\delta^\dagger = \sum_{\kappa \neq \delta} w^\delta_{\kappa} f_\kappa + b^\delta_{\kappa}$$

where $N$ is the matrix size; $f_\delta^\dagger$ is the $\delta$-th interpolated coding element; $f_\kappa$ is the $\kappa$-th known element; $w^\delta_{\kappa}, b^\delta_{\kappa}$ are the optimal solution of the least squares estimation of $\{f_\delta, f_\kappa\}$ as (3); $\alpha^\delta_{\kappa}$ is the normalized weight coefficient of Pearson’s correlation coefficient as (4).

$$\begin{align*}
    w^\delta_{\kappa} &= \frac{\sum_{\kappa \neq \delta} f_\delta f_\kappa}{\sum_{\kappa \neq \delta} f_\delta^2 - \frac{1}{N} \left(\sum_{\kappa \neq \delta} f_\kappa\right)^2} \\
    b^\delta_{\kappa} &= \frac{1}{N} \sum_{\kappa \neq \delta} f_\delta w^\delta_{\kappa} f_\kappa \\
    \alpha^\delta_{\kappa} &= \frac{\sum_{\kappa \neq \delta} f_\delta f_\kappa}{\sqrt{\sum_{\kappa \neq \delta} f_\delta^2 \sum_{\kappa \neq \delta} f_\kappa^2}} \\
    \rho^\delta_{\kappa} &= \frac{\sum_{\kappa \neq \delta} (f_\delta - \bar{f}_\delta)(f_\kappa - \bar{f}_\kappa)}{\sqrt{\sum_{\kappa \neq \delta} (f_\delta - \bar{f}_\delta)^2} \sqrt{\sum_{\kappa \neq \delta} (f_\kappa - \bar{f}_\kappa)^2}}
\end{align*}$$

4.1.2 Acquiring silhouettes by mobile

The user is required to take front and side photos. To segment the user images into silhouettes, the end-to-end models of semantic segmentation can be deployed on mobile terminal to predict the pixel-level labels. And the commercial human body segmentation API also can be called. Therefore, the multi-perspective silhouettes are quickly captured on mobile terminal, and the complex preprocessing on PC terminal is avoided.

In acquiring the silhouettes of the user, T-pose is set to the standard pose. Compared to A-pose, T-pose is easier to maintain and is harder to form the self-occlusion. Ultimately, the multi-perspective silhouettes are utilized as the one of inputs of the 3D human body reconstruction network.
4.1.3 Three-Dimensional human body model representation

Three-dimensional human body reconstruction network predicts the shape coefficients instead of the point cloud. The shape coefficients are utilized to generate the corresponding 3D human body model based on SMPL (Loper et al. 2015).

The shape coefficients encode the human body shape in the low dimension by PCA. Three-dimensional human body model with various shape $B_s$ is defined as (5).

$$B_s(\lambda) = \sum_{i=1}^{n} \lambda_i I_i$$

where $I = [I_1, I_2, \ldots, I_n]$ is the first n shape displacement principal components in the PCA; $\lambda = [\lambda_1, \lambda_2, \ldots, \lambda_n] \in \mathbb{R}^n$ are shape coefficients. The mapping relation between the shape coefficient and 3D human body shape is one-to-one correspondence.

Three-dimensional human body model with various pose $B_p$ is defined as (6).

$$B_p(\theta) = \sum_{i=1}^{23} [R_i(\theta) - R_i(\theta^*)] P_i$$

where $R_i(\theta)$ is mapping function converting the pose coefficient $\theta$ to the relative rotation matrix of the joint; $K = 23$ is the joint number; $R_i(\theta^*)$ is rotation matrix in T-pose; and $P = [P_1, P_2, \ldots, P_{23}]$ is mixed pose matrix.

In the 3D human body reconstruction toward virtual try-on, the shape coefficients $\lambda$ are crucial, while the pose coefficients are set to zero. And the generated 3D human body model is displayed on mobile terminal.

4.2 MDD-Net

MDD-Net learns fuzzy anthropometric features by combined FC layers. The channel number of combined FC layers gradually shrinks. MDD-Net learns the multi-perspective silhouette features based on DenseNet201, and the learned silhouette features are merged by the convolutional layers and the FC layers. The global and local shape features are fused at the end of MDD-Net. The structure of MDD-Net is shown in Fig. 3.

4.2.1 Training setting

The objective function of the MDD-Net in the training is defined as (7).

$$\ell = \sqrt{\sum_{i} \zeta_i \cdot \frac{\left|\lambda_i - \lambda'_i\right|}{\sum \left|\lambda_i\right|}}$$

where $\lambda_i, \lambda'_i$ are the predicted human body shape coefficients and the human body shape coefficient labels; $\zeta_i$ is the normalized weight coefficient, it is positively related to the second norm of shape displacement principal components. The objective function is optimized by Adadelta.

The prediction accuracy of network is defined as (8).

$$\epsilon = \sum_{i} \zeta_i \cdot e^{-\left|\lambda_i - \lambda'_i\right|}$$

The difference between the predicted result and the label is inputted into $e^\epsilon$. The prediction accuracy is close to 1 when the difference is close to 0, and the larger the difference, the prediction accuracy is close to 0. The weight coefficients are the same as (7).

4.2.2 The learning fuzzy anthropometric feature module

In the learning fuzzy anthropometric feature module, the fuzzy anthropometric feature matrix is normalized, and the global feature coding of the human body shape is learned by combined FC layers. The channels of the combined FC layers changes from 4096 to 1024. Every FC layer follows by an activation function relu and a dropout layer. The structure of the learning fuzzy anthropometric feature module is shown in half top par of Fig. 3.

To analyze the structure of this module, under the same condition of the training and testing, the training time and the prediction accuracy are the metrics, and the performance of
Fig. 3 is utilized as the benchmark. As shown in Fig. 4a, when the layer number is reduced at the head or tail of this module, the under-fitting will appear. As shown in Fig. 4b, when the layer number is increased at the head of this module, the prediction accuracy improves slightly, and the training time increases exponentially. When the layer number is increased at the tail of this module, the under-fitting will also appear. As shown in Fig. 4c, the variation of channel number either increases the training time exponentially or decreases prediction accuracy. It proves that the structure of this module is optimal.

4.2.3 The learning multi-perspective silhouette feature module

Compared to the single perspective silhouette, the multi-perspective silhouettes contain more human body shape features, while is not time-consuming or expensive for the user. Thus, the multi-perspective silhouettes are taken as the inputs of this module. There are two methods for merging multi-perspective silhouette features. (1) the channels are stacked at the input end; (2) the features of different perspectives are learned separately, and the learned features are merged at the end of the network. And method (2) is chosen for this module.

The human body shape features are learned by DenseNet201, and the learned features are merged by the concatenate layer at the end of DenseNet201. To further fuse the learned features, the channel number is reduced by an FC layer, and the feature map is reduced from $16 \times 16$ to $4 \times 4$ by two convolutional layers, then the feature map is converted into the vector with length 4096 by the flatten layer. The structure of this module is shown in the half bottom part of Fig. 3.

To analyze the structure of this module, under the same condition of training and testing, the training time and the prediction accuracy are the metrics, and the performance of DenseNet201 is utilized as the benchmark. As shown in Fig. 5, DenseNet reuses features by merging features on the channel. Each layer establishes a dense connection with all previous layers. With few parameters, an ultra-deep network is established. It has a higher error backpropagation speed and less training time compared to ResNet (He et al. 2016).

As shown in Fig. 6a, DenseNet201 has the highest prediction accuracy so that more localized features of silhouettes are learned. Therefore, DenseNet201 is optimal to learn the shape features from silhouettes.

4.2.4 Feature fusion

The feature coding 1 and 2 are learned from the fuzzy anthropometric matrix and silhouettes. Their lengths are 4096 and 1024, respectively. The feature vectors are merged by two FC layers to predict the shape coefficients. The structure of feature fusion is shown in the half right part of Fig. 3.

In the feature fusion, the feature ratio can be adjusted by adding an FC layer after the flatten layer. Because the maximum length of feature coding 2 is 4096, if an FC layer with a length greater than 4096 is added, the human body shape features will not increase. Thus, to verify the influence of the merging ratio, the FC layers with the length of 3072, 2048, 512, 256 are added. As shown in Fig. 6b, the merge ratio has little effect on the prediction accuracy. And the ratio of 4:1 is maintained.

MDD-Net can realize the feature fusion of the human body shape in multi-dimension. When a 3D human body is projected into a 2D image, the circumference feature is flattened...
4.3 Generating human body shape fusion features dataset

As introduced in Sect. 4.1, MDD-Net takes labeled silhouettes and fuzzy anthropometric features as inputs for the shape estimation. If the real people data are utilized to train MDD-Net, the shape coefficients need to calculate. So, a great number of people need to scan, and their 3D point clouds need to align with the SMPL by the rigid and non-rigid method (Groueix et al. 2018). This process is time-consuming, expensive, and tool-dependent. Therefore, a fusion feature data set of human body shape is generated to

Fig. 6 a Structure comparison of the learning multi-perspective silhouette feature module. b The merge ratio of the multi-perspective silhouette and fuzzy anthropometric feature
Table 1 Three-dimensional human body model dataset details

| Group | Quantity | The shape coefficients range |
|-------|----------|-----------------------------|
| 1     | 500      | $(-5.5, 5]$, $\xi_m = 0.20$ |
| 2     | 1250     | $(-5.5, 5]$, $\xi_m = 0.04$ |
| 3     | 1000     | $(-2, 2]$                    |
| 4     | 1100     | $(-5, 5]$                    |

Table 2 New joint point landmarks

| No | Description                                      |
|----|--------------------------------------------------|
| 25 | The point with maximum y value in point clouds   |
| 26 | The middle point between point 2 and 5           |
| 27 | The middle point between point 5 and 8           |
| 28 | The middle point between point 17 and 19         |
| 29 | The middle point between point 19 and 24         |
| 30 | The middle point between point 12 and 15         |

4.3.1 Generating 3D human body model

The 3D human body shape data set is generated based on the SMPL as (5). The details are shown in Table 1.

In group 1, the human body shape varies in all principal components of the shape displacement. It consists of ten subgroups and contains 500 models. Its shape coefficients are set as (9) where $k$ is the group number; $R_{\text{min}}$ is the minimum shape coefficient; and $\xi_m$ is the step. In group 2, the samples size of the first five shape displacement principal components is increased. It consists of five subgroups and contains 1250 models. In group 3, the shape coefficients are randomly picked in the range of $(-2, 2]$ and it ensures that this dataset contains common human body shapes. In group 4, the shape coefficients are randomly picked in the range of $(-5, 5]$ and it contains some uncommon human bodies to expand the human body shape space. Some examples of this dataset are shown in Fig. 8a.

$$\lambda^k = [\lambda^k_1, \lambda^k_2, \ldots, \lambda^k_m]$$

$$\lambda^k_i = \begin{cases} R_{\text{min}} + k \cdot \xi_m, & k = i \\ 0.1, & k \neq i \end{cases}$$

4.3.2 Generating fuzzy anthropometric feature dataset

As shown in Fig. 7a, SMPL has 24 joint landmarks (black points). And six new joint landmarks (red points) are added as Table 2. The anthropometric features are generated as Table 3. In generating the leg and arm features, the human body is set to be symmetrical in shape as a whole.

$$f \sim (f_0 + \alpha \cdot \mu_{f_0} \cdot N(0, \sigma_{f_0}))$$

where $f$, $f_0$ are the fuzzy and accurate anthropometric features; $\alpha$ is the fuzzy coefficient; $\mu_{f_0}$ is the mean of $f_0$ in the human body shape space; $N(0, \sigma_{f_0})$ is the standard normal distribution.

Since the human body shape space conforms to the Gaussian distribution, the fuzzy degree of the acquired
### Table 3 Anthropometric features

| Feature                  | Sign | Note                                      |
|--------------------------|------|-------------------------------------------|
| Bust                     | $C_1$| Benchmark:$l_9(y), \xi_f = 0.015$        |
| Waist                    | $C_2$| Benchmark:$l_3(y), \xi_f = 0.015$        |
| Hipline                  | $C_3$| Benchmark:$l_2(y), \xi_f = 0.015$        |
| Thigh circumference      | $C_4$| Benchmark:$l_{26}(y), \xi_f = 0.015$    |
| Shank circumference      | $C_5$| Benchmark:$l_{27}(y), \xi_f = 0.015$    |
| Neck circumference       | $C_6$| Benchmark:$l_{30}(y), \xi_f = 0.015$    |
| Upper arm circumference  | $C_7$| Benchmark:$l_{28}(x), \xi_f = 0.015$    |
| Lower arm circumference  | $C_8$| Benchmark:$l_{29}(x), \xi_f = 0.015$    |
| Height                   | $h$  | $l_{25}(y) - l_{11}(y)$                  |
| Thigh length             | $L_1$| $\|l_2 - l_5\|$                          |
| Shank length             | $L_2$| $\|l_5 - l_8\|$                          |
| Upper arm length         | $L_3$| $\|l_{17} - l_{19}\|$                    |
| Lower arm length         | $L_4$| $\|l_{19} - l_{24}\|$                    |
| Shoulder                 | $L_5$| $\|l_{16} - l_{17}\|$                    |
| Neck length              | $L_6$| $\|l_{12} - l_{15}\|$                    |
| Upper and lower body proportion | $L_7$| $\|l_0 - l_{15}\|/(L_1 + L_2)$          |

Example-guided features also conforms to the Gaussian distribution. For the probability density of the human body shape space, the boundary value is much smaller than the center value. But the example reference and fuzzy options are linear. Thus, the reference of the extreme example is more accurate. And the Gaussian noise is added to the accurate anthropometric feature to generate the fuzzy anthropometric feature with the range of [0, 100].

#### 4.3.3 Generating multi-perspective silhouette dataset

Based on OpenGL, the 3D human body model is orthogonally projected to generate front and side silhouettes. The silhouette is the binary mode and has a resolution of $512 \times 512$. Some examples of silhouettes are shown in Fig. 8b, c.

#### 4.3.4 Generating test set

The generated human body shape fusion feature dataset is divided into a training set, a verification set, and a test set by the ratio of 6:2:2. And another three test sets: segmentation noise, nonstandard poses, and perspective angle errors are also generated to verify the robustness of the network.

- **Unprocessed test set.** It consists of 2750 samples with multi-perspectives silhouettes and fuzzy anthropometric features. It is divided from generated human body shape fusion feature dataset. It ideally reflects the human body shape feature on the silhouettes.

- **Segmentation noise test set.** Because of the complex background and dress, the semantic segmentation of the user’s image will be noised. Thus, the black and white noise blocks have been shown on the silhouettes of the unprocessed test set. The influence of segmentation noise cannot be eliminated by the traditional filter. As Fig. 9 shows, the experiments demonstrate the boundary of the filtered silhouette is distorted, the shape is overall shrunk, or more noise appears.

- **Nonstandard pose test set.** When acquiring the human body silhouettes, the pose held by the user maybe is not the standard T-pose. To simulate pose noise, the shape coefficients are consistent with the unprocessed test set,
Fig. 10 Nonstandard pose test set. a All parts; b the arm part; c the leg part; d A-pose; e H-pose

each joint is rotated in the range $[-10^\circ, 10^\circ]$, and the nonstandard pose test set is generated.

To further analyze the pose error influence, five sub-test sets are also generated. As shown in Fig. 10a, all human body parts are rotated in various angle ranges. The angle ranges are $[-1^\circ, 1^\circ]$, $[-3^\circ, 3^\circ]$, $[-5^\circ, 5^\circ]$, $[-7^\circ, 7^\circ]$, $[-9^\circ, 9^\circ]$ from left to right. When the user maintains a pose, pose errors primarily appear in the arm and leg parts. The arm part rotates within the aforementioned angle ranges in Fig. 10b. And the leg part rotates in Fig. 10c. To evaluate the influence of extreme pose, the model pose is set as A-pose or H-pose in Fig. 10d, e.

- Perspectives error test set. As shown in Fig. 11a, when acquiring human body silhouettes, if the y-axis is not parallel to the y’-axis (i.e., the up axis of the camera is not vertical during the photographing), the acquired silhouettes present the status of looking up or looking down. If the z-axis is not parallel to the z’-axis (i.e., the direction axis of the camera is not pointing to the human body), the acquired silhouettes are visually tilted, and self-occlusion may appear in the arm part for the side silhouette. To simulate the perspective error, in rendering the silhouettes, the view deviation angle of the y-axis and z-axis are set to $[-10^\circ, 10^\circ]$. An example is shown in Fig. 11b.

5 Experiment

5.1 Ablation experiment

All experiments were performed on a PC device with NVIDIA GeForce GTX 1080 GPU. The training set and test set are introduced in Sect. 4.3.

Fig. 11 Perspective error test set. a Perspective; b Silhouettes with perspective error

Fig. 12 a The import factor of coding elements; b The prediction accuracy with various fuzzy degrees

5.1.1 The learning fuzzy anthropometric feature module

In this section, the balance of coding elements is analyzed. In acquiring the fuzzy anthropometric feature matrix, if some coding elements are default, the interpolation estimation is utilized to ensure density. Thus, the effectiveness of the interpolation estimation is verified. In feature acquiring on mobile terminal, the input fuzzy degree is determined by the user’s subjective consciousness. So, with different fuzzy degree inputs, the predicted accuracies of the learning fuzzy anthropometric feature module are compared.

To verify the coding element importance, a specific coding element $f_d$ is removed from the feature matrix, and the rest of coding elements $\{f_i\}$ are utilized to train the learning fuzzy anthropometric feature module. The import factor introduced as (1) is utilized as the metrics.

The import factors of coding elements are shown in Fig. 12a, the average importance of the global length feature is 208.3% of circumference features. Thus, the global length features have high priority in acquiring the fuzzy anthropometric feature matrix. The mean importance factor of coding elements is 6.25, the standard deviation is 4.74, and the importance factor distribution is relatively balanced. It proves that our selection in fuzzy anthropometric feature elements is rational.

To verify the default interpolation estimation method, the coding elements with high import factors are utilized to estimate the remaining default elements. The interpolation error
is the metrics and is defined as (11) where $f_\alpha, f_\alpha^\dagger$ are the original and interpolated coding element, respectively; $m$ is the number of default elements.

$$e_I = \frac{1}{m} \sum_\delta \left| \frac{f_\delta - f_\delta^\dagger}{f_\delta} \right|$$  \hspace{1cm} (11)

The interpolation error rates under various sparsity degrees of fuzzy anthropometric feature matrices are shown in Table 4. It found that the interpolation error rate is about 30%, which fully meets the accuracy requirements of rough interpolation, but acquiring dense fuzzy anthropometric feature matrices is still recommended in virtual try-on on mobile terminal.

With the different fuzzy degree inputs, the prediction accuracies of this module are shown in Fig. 12b. Taking accurate anthropometric features without Gaussian noise as inputs, the predicted accuracy of this module on the validation set reaches 72.54% after 160 epochs. It proves that 3D human body model reconstruction can be implemented only with the anthropometric feature. In the case of fuzzy inputs, if the fuzzy degree $\alpha$ is set to 0.1, the prediction accuracy of this module on the validation set drops to 63.24%. And if $\alpha$ is set to 0.2, the prediction accuracy drops to 60.03%. Thus, the prediction accuracy of the learning fuzzy anthropometric feature module is inversely related to the input fuzzy degree. The human body shape estimation ability of this module is still restricted by the variety of samples.

### 5.1.2 The learning multi-perspective silhouette feature module

To prove the superiority of this module in learning the human body shape feature from silhouettes, all structures are trained by multi-perspective silhouettes generated in Sect. 4.3.3. The prediction accuracy results are shown in Fig. 13a.

Compared with only inputting a single silhouette, this module has higher predicted precision with inputting multi-perspective silhouettes. And compared to merging RGB channels at the beginning, learning the human shape features of each perspective, merging the learned features at the end has a better and more stable training result.

To prove the superiority of this module in training, the comparisons are shown as in Fig. 13b, c. Compared to BfS-Net, this module has a similar prediction accuracy on the validation set, while its training time is only 42.4% after 80 epochs. Compared to HS-Net, the prediction accuracy of this module is improved significantly by 16.4%. Thus, the learning multi-perspective silhouette feature module maintains a better balance between the prediction accuracy and training time.

### 5.1.3 The human body shape features acquired on mobile terminal

Considering the requirements of virtual try-on on mobile terminal introduced in Sect. 3, the human body shape features to be acquired are divided into following cases. The following cases are corresponding to the input cases of MDD-Net as introduced in Sects. 5.1.1 and 5.1.2. It enables MDD-Net to be applied in various scenes.

- One-dimensional fuzzy anthropometric feature. It adapts to the scene that the user has not the condition of the photograph or is unwilling to upload private photos. The guided examples are provided on the mobile terminal as Fig. 1. Then, the fuzzy anthropometric features are acquired, and the shape coefficients are predicted by the learning fuzzy anthropometric feature module. In acquiring fuzzy anthropometric features, \{L_6, L_5, L_4, L_2, C_2, L_7, L_3, L_1\} are the compulsory terms, while \{C_3, C_5, C_4, C_8, C_9, C_1, C_6, h\} are the optional terms. If some anthropometric features have not been acquired from the user, the default features can be calculated by our interpolation estimation method. The shortcoming of this input is that the prediction accuracy is greatly affected by the user’s description of their shape. And the prediction accuracy is approximate 50%.

- Single perspective silhouette. It adapts to the scene that the user has the condition of the photograph. The prediction accuracy inputting single-perspective silhouette is 80% of the being inputting multi-perspectives silhouette. And the cost of acquiring other silhouettes is not expensive. So, this kind of input is not recommended.

- Multi-perspective silhouette. It adapts to the scene that the user can take the front and side photos under the T-pose. The human body images are acquired as Fig. 1. And the multi-perspective silhouettes are generated by calling the human body segmentation API. The learning multi-perspective silhouette feature module is utilized to predict the shape coefficients. The prediction accuracy can reach up to 90% but significantly drops with the stan-
One-dimensional fuzzy anthropometric feature + multi-perspective silhouette. It adapts to the scene that the user is willing to submit the multi-dimension shape features for estimating the human body shape more accurately. MDD-Net can fulfill this task and maintain robustness against noise.

Our method of human body estimation only requires the user’s photos and fuzzy anthropometric features. To satisfy the input conditions of our method, the user only needs a mobile phone equipped with a camera. We figure that the convenience of our method will promote the application of virtual try-on. But our method explicitly fits fusion features to SMPL, it enables the representation ability of human body shape to rely heavily on SMPL. As the mobile equipped with 3D sensors became more popular, some works (Yang et al. 2021) utilized scans to infer finer human body shape, and their reconstructed models appear more realistic than ours. It is worth noting that the mobile equipped with 3D sensors is still expensive for the user.

5.2 Comparison

To verify the superiority of our method in virtual try-on on mobile terminal, for the human body shape representation, MDD-Net is compared with the models of shape-pose blended estimation with a fixed pose. And for the robustness against pose and perspective, MDD-Net is compared with the models of accurate shape estimation.

5.2.1 The application in virtual try-on

Virtual try-on has a high demand on accurate shape estimation in 3D human body reconstruction. So, 3D human body reconstruction methods are required to satisfy following condition: (a) the predicted human body shape space is wide; (b) the predicted human body shape is accurate. Therefore, $\lambda_{\text{std}}$, $\lambda_{\text{max}}$, $\lambda_{\text{min}}$ are utilized to evaluate the predicted human body shape space, and are defined as (12) where $\lambda_{\text{std}}(i)$, $\lambda_{\text{max}}(i)$, $\lambda_{\text{min}}(i)$ are the standard deviation, maximum, minimum of the i-th shape coefficient over all samples. The prediction accuracy $\varepsilon$ as (8) are utilized to evaluate the shape coefficient precision.

\[
\begin{align*}
\lambda_{\text{std}} &= \sum_{i} \zeta_i [\lambda_{\text{std}}(i)] \\
\lambda_{\text{max}} &= \sum_{i} \zeta_i [\lambda_{\text{max}}(i)] \\
\lambda_{\text{min}} &= \sum_{i} \zeta_i [\lambda_{\text{min}}(i)]
\end{align*}
\] (12)

SMPLify (Bogo et al. 2016) and HMR (Kanazawa et al. 2018) are utilized as the representatives of shape-pose blended estimation. SMPLify reconstruct 3D human body model by optimizing joints online, and HMR is an end-to-end model. Their official code and pretrained model are utilized to test on the unprocessed test set. And the learning multi-perspective silhouette feature module in MDD-Net is also utilized to test. The qualitative examples are shown in Fig. 14, and the quantitative comparison is shown in Table 5.

The predicted human body shape spaces predicted by SMPLify and HMR are narrow. The $\lambda_{\text{std}}$ of SMPLify and HMR are only 20.34% and 7.59% of the label space. And their $\varepsilon$ is only about 50%. For the fat and thin human body as Fig. 14, the
Table 5 Interpolation error rate at various sparsity

|          | Label space | SMPLify | HMR | MDD-Net |
|----------|-------------|---------|-----|---------|
| $\lambda_{sd}$ | 2.90        | 0.59    | 0.22 | 2.81    |
| $\lambda_{max}$ | 5.00        | 2.89    | 1.53 | 6.69    |
| $\lambda_{min}$ | 5.00        | 0.06    | -0.21 | -6.18   |
| $\varepsilon$  | 50.10%      | 49.90%  | 90.41% |        |

Fig. 15 Qualitative comparison on four test sets. a Unprocessed test set; b segmentation noise test set; c nonstandard pose test set; d perspectives error test set

The shape of their 3D human body model is close to the mean shape. In contrast, the human body shape space predicted by our method is approximately 96.90% of the label space, while the predicted accuracy is about 90.41%. It demonstrates that MDD-Net has superior ability in reconstructing the user’s human body shape and is more suitable for virtual try-on.

5.2.2 The robustness of human body shape estimation

In acquiring the silhouettes of the user, the prior pose, perspective, and segmentation are not necessarily standard. It requires the 3D human body reconstruction model to be robust against pose and perspective noise. Hs-Net (Dibra et al. 2016a) and BfSNet (Smith et al. 2019) are utilized as representatives of accurate shape estimation models. Hs-Net and BfSNet have been trained on the multi-perspective silhouettes dataset, and MDD-Net has been trained on the human body shape fusion feature dataset. All methods are evaluated on four test sets. In this section, the qualitative and quantitative comparison are demonstrated.

The qualitative comparison on four test sets is shown in Fig. 15. The displacement error is the difference between the original and predicted and 3D human body models.

The quantitative comparison is shown in Fig. 16. The prediction accuracy as (8) is utilized as the metric.

The reconstructed models of Hs-Net have distortions in the belly and legs. BfSNet can predict a perfect result on the unprocessed test set while has higher displacement error on the segmentation noise, the nonstandard pose, and the perspective error test set. Our method can still predict a perfect human body model even with much segmentation noise, the raised arm, and the oblique human body. In the segmentation noise test set, the expression of local shape features is influenced by the segmentation noises. And the large displacement errors appear on the parts with complicated body curves (e.g., head and buttocks). In the perspective error test set, a large displacement error occurs at the junction of the multi-perspectives, which is caused by the self-occlusion of shape features.

The quantitative comparison is shown in Fig. 16. The prediction accuracy as (8) is utilized as the metric.

In the unprocessed test set, the prediction accuracy of MDD-Net is about 89.5%, second only to 92.52% of BfSNet, and far greater than 76.11% of Hs-Net. But in the segmentation noise, nonstandard pose, and perspectives error test sets, compared with BfSNet, MDD-Net improved the prediction accuracy by 43.4%, 37.2%, and 9.0%, respectively. This proves the superiority of our method in the robustness against the pose and perspective noise.

Furthermore, the performance of the MDD-Net is hardly affected by the fuzzy degree of anthropometric feature, and the prediction accuracy is almost the same when $\alpha$ is 0.1 or 0.5. In the contrast, the prediction accuracy of the learning fuzzy anthropometric feature module is inversely related to $\alpha$. Therefore, in the multidimensional feature fusion, MDD-Net intends to merge learned shape features rather than accumulate errors. It means that the network robustness is not implemented by limiting the sample generalization in a certain dimension. So, in acquiring the anthropometric features on the mobile terminal, guided-example qualitative fea-
5.2.3 The influence of nonstandard pose

To further analyze the robustness of MDD-Net to pose noise, MDD-Net has been tested on five sub-test sets in Sect. 4.3.4, and the tendency of the prediction accuracy with the variation of the angle range is shown in Fig. 17.

Experiments demonstrate that the prediction accuracy decreases with the increase in the angle range, but it remains 66.71% within the range of $[-9^\circ, 9^\circ]$. When the user keeps the A-pose, the pose error easily appears in the arms and leg parts. Experiments indicate that the pose noise in the arm part has little effect on $\varepsilon$, and $\varepsilon$ is only reduced by 8.08% even within the angle range of $[-9^\circ, 9^\circ]$. The $\varepsilon$ decline caused by the leg part is much smaller than all parts. It proves that MDD-Net is robust for the common arm and leg error.

To evaluate the robustness of MDD-Net to extreme poses, the fusion features are extracted from models with A-pose or H-pose, and MDD-Net estimates the human shape from the fusion features. As shown in Fig. 18, the reconstructed models are still similar to the truth as a whole. And the $\varepsilon$ with A-pose and H-pose are 76.52% and 74.67%, respectively.

These experiments prove that our method does not require users to maintain an absolute standard T-pose, and our method has good practicability.

5.3 Anthropometry evaluation

The purpose of virtual try-on is not only to show the user the try-on effect of garments, but also to output the user’s human body shape parameters. Then, the human body shape parameters are utilized for personalized clothing customization. This requires that the reconstructed 3d human body model can correctly reflect the user’s shape in some crucial parts. To evaluate the ability of models to estimate the shapes in crucial parts, the anthropomorphic estimation is utilized to measure the human body parts of the reconstructed 3D human models. The human body parts are selected as Fig. 7. In addition, the fusion features are generated from 3D human body model, and MDD-Net takes the fusion features as inputs to estimate the human body shape. So, the anthropomorphic estimation is also the metrics to evaluate the coding and decoding human body feature ability of our method.

The 3D human body models generated in Sect. 4.3.1 are utilized as the label and the average performance of MDD-Net under different $\alpha$ is taken as reference. The average height $h$ of 3D human body models on the unprocessed test set is 1.793, the average height $H$ of Chinese adult males is 169.7 cm [47], and the coefficient $c = H/h$ is utilized to map the point cloud error into real length. The anthropomorphic estimation is evaluated from three metrics: mean value, standard deviation, and truncation score. The truncation score $\rho$ is defined in (13) where $\xi$ is the cutoff coefficient; $e$, $f_0$ are the error and the mean measured value of human body part; and the $\varepsilon(\cdot)$ is the activation function. The $\xi$ is set to 0.01, the $\rho$ equals 1 when the measured value error is less than 1% of $f_0$, otherwise $\rho$ equals 0.

$$\rho = \frac{\varepsilon(\xi \cdot f_0 - e)}{}$$

As shown in Tables 6, 7, 8, it is found that MDD-Net is more accurate and robust in the anthropometry evaluation compared to other networks on four test sets. The mean error, standard deviation is smaller, and the cutoff score is high. The virtual try-on has high requirements on BWH, hand length, leg length and, shoulder breadth, the truncation score of the MDD-Net in the above parts is even twice that of other networks. Due to the fusion feature encodes the global features, MDD-Net is excellent in decoding the global shape features, and is more suitable for virtual try-on.

6 Conclusion

This paper proposes a rapid human body shape estimation method for virtual try-on on mobile terminal based on MDD-Net. MDD-Net is a supervised deep learning network that inputs fuzzy anthropometric feature matrix and the multi-perspective silhouettes and predicts the shape coefficients.
Compared with the previous methods, this paper mainly has three contributions:

1. The fusion feature acquisition method on mobile and the generation method of fusion feature datasets are proposed. The mobile terminal acquires fuzzy anthropometric features and multi-perspective silhouettes and displays the reconstructed 3D human body model. To simulate the real inputs, the related fusion dataset is generated. The fusion dataset consists of the fuzzy anthropometric feature dataset, the silhouettes dataset, and the pose perspective noise test sets.

2. The human body shape estimation network MDD-Net is proposed. MDD-Net consists of the learning fuzzy anthropometric feature module and the learning multi-perspective silhouette feature module. It learns the global
and local shape features from multi-dimensions. And the feature fusion is merged with the ratio of 4:1 to predict the shape coefficients.

(3) The ablation experiments of MDD-Net have been performed and MDD-Net is qualitative and quantitative compared with other models. MDD-Net has a wider human body shape representation space and higher prediction accuracy compared to HMR and SMPLify. MDD-Net is more robust to the pose and perspective noise compared to Hs-Net and BFSNet, and the prediction accuracy is improved by 13.34%, 55.77%, 34.6%, and 43.4%, 37.2% 9.0% on four test sets.

In future work, more input forms like 3d scans will be considered to estimate the finer human body shape, and the robustness against the dressed garments (Lassner et al. 2017a) will also be improved.

Acknowledgements The authors would like to thank anonymous reviewers and the associate editor for their valuable comments and suggestions that greatly improved this paper’s quality.

Author Contributions All authors contributed to the study conception and design. The first draft of the manuscript was written by Naiyu Fang and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding This work was supported by the National Key R&D Program of China [Grant Numbers: 2018YFB1700700], and National Natural Science Foundation of China [Grant Number 51875516].

Data availability Enquiries about data availability should be directed to the authors.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

Informed consent Informed consent was obtained from all individual participants included in the study.

References

Anguelov D, Srinivasan P, Koller D, Thrun S, Rodgers J, Davis J (2005) Scape: shape completion and animation of people. In: ACM SIGGRAPH 2005 Papers, pp 408–416
Baek SY, Lee K (2012) Parametric human body shape modeling framework for human-centered product design. Comput Aided Des 44(1):56–67
Bălan AO, Black MJ (2008) The naked truth: Estimating body shape under clothing. In: European conference on computer vision, Springer, pp 15–29
Bhatnagar BL, Tiwari G, Theobalt C, Pons-Moll G (2019) Multi-garment net: Learning to dress 3d people from images. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 5420–5430
Bogo F, Romero J, Loper M, Black MJ (2014) Faust: Dataset and evaluation for 3d mesh registration. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3794–3801
Bogo F, Kanazawa A, Lassner C, Gehler P, Romero J, Black MJ (2016) Keep it smpl: automatic estimation of 3d human pose and shape from a single image. In: European conference on computer vision, Springer, pp 561–578
Chen X, Pang A, Wei Y, Xui L, Yu J (2019) Tightcap: 3d human shape capture with clothing tightness. arXiv preprint arXiv:1904.02601
Chen Y, Cheng ZQ, Lai C, Martin RR, Dang G (2015) Realtime reconstruction of an animating human body from a single depth camera. IEEE Trans Visual Comput Graphics 22(8):2000–2011
Dibra E, Jain H, Öztireli C, Ziegler R, Gross M (2016a) Hs-nets: Estimating human body shape from silhouettes with convolutional neural networks. In: 2016 fourth international conference on 3D vision (3DV), IEEE, pp 108–117
Dibra E, Öztireli C, Ziegler R, Gross M (2016b) Shape from selfies: Human body shape estimation using cca regression forests. In: European conference on computer vision, Springer, pp 88–104
Dibra E, Jain H, Öztireli C, Ziegler R, Gross M (2017) Human shape from silhouettes using generative hks descriptors and cross-modal neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4826–4836
Elhayek A, de Aguiar E, Jain A, Tompson J, Pishchulin L, Andriluka M, Bregler C, Schiele B, Theobalt C (2015) Efficient convnet-based marker-less motion capture in general scenes with a low number of cameras. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3810–3818
Groueix T, Fisher M, Kim VG, Russell BC, Aubry M (2018) 3d-coded: 3d correspondences by deep deformation. In: Proceedings of the European conference on computer vision (ECCV), pp 230–246
Guler RA, Kokkinos I (2019) Holopose: Holistic 3d human reconstruction in-the-wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 10884–10894
He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778
Hormung A, Dekkers E, Kobelt L (2007) Character animation from 2d pictures and 3d motion data. ACM Trans Graph 26(1):1–es
Jiang L, Ye J, Sun L, Li J (2019) Transferring and fitting fixed-sized garments onto bodies of various dimensions and postures. Comput Aided Des 106:30–42
Joo H, Simon T, Sheikh Y (2018) Total capture: A 3d deformation model for tracking faces, hands, and bodies. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 8320–8329
Kanazawa A, Black MJ, Jacobs DW, Malik J (2018) End-to-end recovery of human shape and pose. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7122–7131
Kocabas M, Athanasiou N, Black MJ (2020) Vibe: Video inference for human body pose and shape estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 5253–5263
Lassner C, Pons-Moll G, Gehler PV (2017a) A generative model of people in clothing. In: Proceedings of the IEEE international conference on computer vision, pp 853–862
Lassner C, Romero J, Kiefel M, Bogo F, Black MJ, Gehler PV (2017b) Unite the people: Closing the loop between 3d and 2d human representations. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 6050–6059

© Springer
The rapid construction method of human body model for virtual try-on on mobile terminal based... 12039

Xu H, Li J, Lu G, Deng H, Zhang D, Ye J (2018) Modeling 3d human body with a smart vest. Comput Graph 75:44–58

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.