Soft Computing manuscript No.
(will be inserted by the editor)

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

Naiyu Fang · Lemiao Qiu* · Shuyou Zhang · Zili Wang · Ye Gu · Kerui Hu

Received: March 30, 2021/ Accepted:

Abstract Traditional anthropometric evaluation needs professional measuring tools and operations, which 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 solved. 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 the input of fusion features acquired by mobile as input and outputs 3D human body model to mobile supporting for 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 corresponding fusion feature data set based on SMPL also is generated for MDD-Net training. In the experiments, without append fault-tolerant training samples, on the segmentation noise, nonstandard pose, and perspective error test set, the predicted accuracy of MDD-Net is improved by 13.34%, 55.77%, 34.6% and 43.4%, 37.2%, 9.0% respectively compared to Hs-Net and BfSNet proving its robust with the impact of uncertain positions and poses. And MDD-Net has a small error and standard deviation on critical anthropometric features explaining the effectiveness of our method.

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

✉Corresponding author.
E-mail: qiulm@zju.edu.cn

Naiyu Fang, Lemiao Qiu, Shuyou Zhang, Zili Wang, Ye Gu, Kerui Hu: State Key Laboratory of Fluid Power Transmission & Control, Zhejiang University, Hangzhou, 310027, China

1 Introduction

3D human body reconstruction is widely used in virtual try-on (Bhatnagar et al., 2019; Jiang et al., 2019), character animation (Hornung et al., 2007; Weng et al., 2019), 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 (Chen et al., 2015; Wang et al., 2020) or RGBD camera (Rhodin et al., 2016; Xu et al., 2019; Yu et al., 2018), 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, it represents different human body objects in terms of shape and pose. (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 utilizing principal component analysis (PCA), and an end to end model is provided with 10 shape coefficients and 72 pose coefficients to represent the human body difference relative to the template. 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 compared to SMPL. In recent researches, the face and hand features are combined into the statistical human body model (Joo et al., 2018; Xiang et al., 2019), and more human body point clouds are utilized to further expand the shape-pose latent space (Zanfir et al., 2020).

The goal of 3D human body reconstruction can be divided into: shape-pose blended estimation and accurate shape estimation.

Shape-pose blended estimation is usually used for the animation and film, the labeled data (e.g. 2d joint point, human body parts segmentation) are utilized to train the network to realize the estimation joint angles and low-dimensional shape coefficients, and the main target is to display the overall motion state of wild images. In the early research, the human body model is reconstructed by aligning the joint points of the image and the template model, the iteration takes a long time and the human body model shape is neutral (Bogo et al., 2016; Guler and Kokkinos, 2019; Lassner et al., 2017b). As CNN develops, aligning the sparse joint points is replaced by optimizing the pixel level loss, 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 of 2d joint point and human body parts segmentation have also been joined into training to further improve the reconstruction accuracy (Omran et al., 2018). In recent researches, inferring motion sequences from videos to achieve the motion simulation of human body models (Kocabas et al., 2020; Leroy et al., 2017; Tung et al., 2017) and monocular multi-person reconstruction (Zanfir et al., 2018) have become new hotspots, but the main content is still about overall pose estimation.

Accurate shape estimation is usually used for virtual try-on, under the condition of a fixed prior pose (e.g. A pose or T pose), a human body model is reconstructed with fine shape. The core content is to ignore the influence of pose, accurately acquire the human body shape features in the image, and predict low-dimensional shape coefficients. In early researches, the single silhouette sample is used for shape coefficient regression with random forest (Dibra et al., 2016b), the 3D feature descriptors are also utilized to improve the accuracy of shape reconstruction (Dibra et al., 2017). However, the single silhouette contains limited human body shape features. In recent researches, 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) towards could output the human body model with a more refined shape.

The main inputs of 3D human body reconstruction are 2d images, the essence of this reconstruction is to use the 2d features of the human body shape to reconstruct the model or predict the human body shape coefficients. Moreover, some researchers try to achieve the 3D human body reconstruction only by inputting the 1d features of the human body shape. The relationship between the anthropometric features and the coefficients of 3D human body model is analyzed. Thus, the 3D human body could be reconstructed only by measuring the specific human body parts features. Baek (Baek and Lee, 2012) analyzed the correlation between the shape variation and the human body contour, proposed a parameterized model that takes the human body parts measured values as inputs and outputs the corresponding human body model. Wuhrer (Wuhrer and Shu, 2013) predicted the human body shape from the encoded anthropometric values utilizing non-linear optimization, and the shape learning is not restricted by the shape latent space. Zhang (Zhang et al., 2015) analyzed the relationship between the measured example-guided features and could predict the 3D human body model by the radial basis interpolation and constraint-driven method.

In the above researches, the shape-pose blended estimation is usually applied for animation and film and is not suitable for virtual try-on. Its core research content is the pose reconstruction from wild images, the reconstructed human body model lacks detailed muscle representation and human body curve features, and it is difficult to acquire the prior features on mobile terminal. In the accurate shape estimation, to avoid the influence of pose on human body shape representation, it needs to acquire the silhouettes with a similar pose to the standard pose or improve the network robustness against the pose noise by inputting a large number of silhouettes with various perspectives and segmentation noise. This increases 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, and the 1d coding matrix size is also difficult 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 including multi - perspectives silhouette features (Dibra et al., 2016a, 2017; Smith et al., 2019) learned by DenseNet (Huang et al., 2017) and example-guided fuzzy anthropometric feature acquired by the mobile terminal as input and performs accurate human body estimation under the T Pose. MDD-Net learns global and local coding features of human body shape from fuzzy anthropometric features and multi-perspectives silhouette respectively. It also generates a corresponding fusion feature data set based on SMPL model for network training. Experiments show that our method is more robust than BfSNet, HS-Net when taking the samples with segmentation noise, nonstandard pose, and perspective error as input, and has a stronger ability to decode global features which are demanded on virtual try-on.

In this paper, the outline of our method is introduced in Sec 3.1. The training set, the modules, and the feature fusion method of MDD-Net are introduced in Sec.3.2. The

generating of fusion feature data sets for network training and four test sets based on the SMPL model are introduced in Sec.3.3. The ablation experiments on MDD-Net, and the training and predicted accuracy comparison are introduced in Sec.4.1. An example visual comparison and anthropometric evaluation is introduced in Sec.4.2.

2 Problem description

For 3D human body reconstruction by inputting the measured values of human body parts, it needs professional measuring tools (e.g. contact tapes) and operations, and the measuring process usually lasts about 10 minutes (Li et al., 2019). Smart wearable device (Uhm et al., 2015; Xu et al., 2018) 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. For 3D human body reconstruction by inputting point cloud, the scanner (Zollhöfer et al., 2014) or multi-view vision system (Bogo et al., 2014; Elhayek et al., 2015; Rhodin et al., 2016) are utilized to collect the RGBD sequence of consumers, and the point cloud is meshed to generate corresponding 3D human body model. Although the reconstructed model is with high precision, manual labeling is usually required, and consumers must be in a specific measuring environment, which is costly, long waiting, and inconvenient.

As the mobile internet develops, the new requirements for virtual try-on come in:

• The virtual try-on should be available anytime and anywhere, and the whole process should be carried out on the mobile APP.

• The acquisition of human shape features should be fast and convenient. Silhouettes and 1d human body shape features should be collected which should not rely on professional tools and complex label information as shown in Fig. 1 (a)(b).

• The reconstructed 3D human body model should have a certain accuracy and should be displayed in real-time as shown in Fig. 1 (c).

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, we study the acquisition of human body shape fusion feature and a precise human shape estimation network. The feature acquisition and model display could be implemented on the mobile, and the network could be deployed on the remote server.

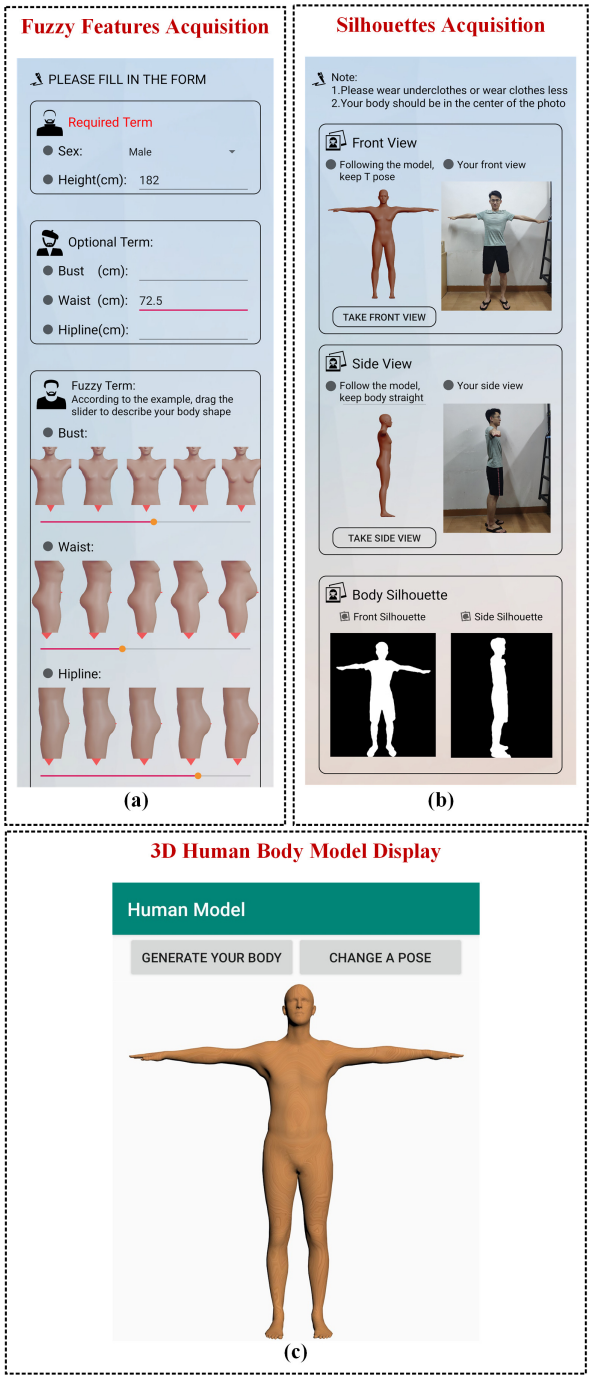

Fig. 1 Features acquisition and model display on APP.

3 Human Body Shape Estimating Method

3.1 Method outline

The consumer 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 the 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 consumer could get the model display in real-time on mobile terminal. The outline of the proposed method is shown in Fig. 2.

3.1.1 Acquire fuzzy anthropometric feature by mobile

The method of fuzzy anthropometric feature acquisition on mobile terminal is to provide people with fuzzy options based on 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 $f = [c_1, c_2, ...c_i, l_1, l_2..., l_i]$ are taken as one of inputs of the 3D human body reconstruction network where c_i is the fuzzy circumference and l_i is the fuzzy length, and the range of it is [0,100].

When acquiring fuzzy anthropometric features, the acquisition rank needs to be sorted, the feature with a large impact on human body shape will be collected first. Thus, the importance of the elements in the fuzzy matrix I_i is defined as (1) where A_{all}, A_i are the reconstruction accuracy of complete feature matrix and partial the fuzzy anthropometric feature matrix without element f_i respectively.

$$I_i = \frac{A_{all} - A_i}{\sum_i (A_{all} - A_i)} \tag{1}$$

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

$$f_\partial = \sum_{\kappa \neq \partial}^{N} \alpha_\kappa^\partial (w_\kappa^\partial f_\kappa + b_\kappa^\partial) \tag{2}$$

where N is the matrix size, f_∂ is the ∂-th default element, f_κ is the κ -th known element, w_κ^∂ , b_κ^∂ are the optimal solution of the least squares parameter estimation of $\{f_\partial, f_\kappa\}$ as (3), α_κ^∂ is the normalized weight coefficient of Pearson's correlation coefficient as (4).

$$\begin{aligned} w_\kappa^\partial &= \frac{\sum_{\kappa \neq \partial}^{N} f_\partial \cdot (f_\kappa - \overline{f_\kappa})}{\sum_{\kappa \neq \partial}^{N} (f_\kappa)^2 - \frac{1}{N} \left(\sum_{\kappa \neq \partial}^{N} f_\kappa \right)^2} \\ b_\kappa^\partial &= \frac{1}{N} \sum_{\kappa \neq \partial}^{N} \left(f_\partial - w_\kappa^\partial f_\kappa \right) \end{aligned} \tag{3}$$

$$\begin{aligned} \alpha_\kappa^\partial &= \frac{|\rho_{x,i}|}{\sum_x |\rho_{x,i}|} \\ \rho_\kappa^\partial &= \frac{\sum_{\kappa \neq \partial}^{N} (f_\kappa - \overline{f_\kappa})(f_\partial - \overline{f_\kappa})}{\sqrt{\sum_{\kappa \neq \partial}^{N} (f_\kappa - \overline{f_\kappa})^2} \sqrt{\sum_{\kappa \neq \partial}^{N} (f_\partial - \overline{f_\kappa})^2}} \end{aligned} \tag{4}$$

3.1.2 Acquire silhouette by mobile

The consumer is asked to take front and side photos, and the photos are simply preprocessed on the mobile terminal to obtain the silhouettes showing the human body shape. The end-to-end model of semantic image segmentation could output pixel-level labels of silhouettes (Chen et al., 2017; Lin et al., 2017). These models could be simply deployed on mobile terminal or the commercial human body segmentation API (e.g. Baidu AI) is utilized. Thus, the mobile could quickly capture the multi-perspective silhouettes, avoiding complex preprocessing on the PC terminal. T pose is chosen as the default pose because that it is easier to maintain and is harder to form the self-occlusion compared to A pose.

To unify the proportion of the human body in the silhouette, multi-perspective silhouettes are scaled based on the height projection ratio $p = \overline{h_r} / \overline{h_s}$ where $\overline{h_r}$, $\overline{h_r}$ are the mean height of the human body in the real world and the silhouette respectively.

Therefore, the multi-perspective human body silhouettes are utilized as the one of inputs of the 3D human body reconstruction network.

3.1.3 Generate 3D human body model

3D human body reconstruction network predicts the shape coefficients instead of 3D human body model, and the shape coefficients are utilized to generate the corresponding 3D human body model based on the SMPL (Loper et al., 2015).

The shape coefficients encode the human body shape space in the low dimension using PCA. 3D human body model with various shape B_s is defined as (5).

$$B_s(\lambda) = \sum_{i=1}^{n} \lambda_i I_i \tag{5}$$

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

3D human body model with various pose is defined as (6).

$$B_p(\theta) = \sum_{i=1}^{9K} (R_i(\theta) - R_i(\theta^*)) P_i \tag{6}$$

where $R_i(\theta)$ is mapping function converting the pose coefficient θ to the relative rotation matrix of the joint, $K = 23$ is

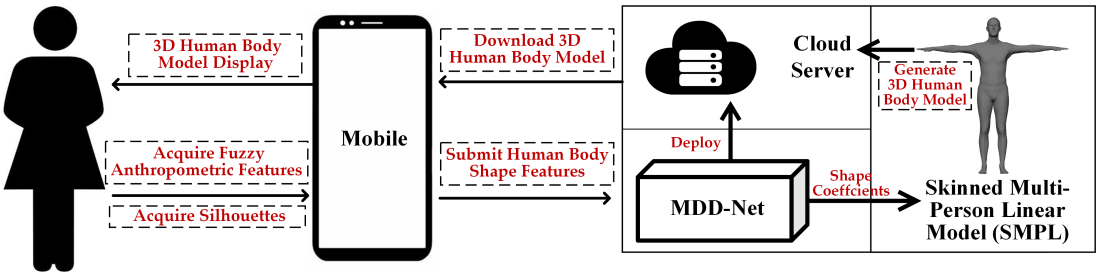

Fig. 2 The method outline.

the joint quantity, $R_i(\theta^*)$ is rotation matrix when the human body keeps the T pose, and $P = [P_1, P_2, ...P_{9K}]$ is mixed pose matrix.

Thus, in generating 3D human body model, the shape coefficient λ is inputted into the temple of SMPL defaulting to zero pose parameter. And the 3D human body model could be displayed on mobile terminal based on OpenGL ES.

3.2 MDD-Net

MDD-Net learns fuzzy anthropometric features combined using the fully connected layers (FC layer) of which channel number gradually shrinks, learns the multi-perspective silhouette features based on the DenseNet201, and merges the learned silhouette features using the convolutional layers and the fully connected layers, and performs feature fusion at the end of the network to merge the global and local feature coding of the human body shape. The network structure is shown in Fig. 3.

3.2.1 Training setting

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

$$\ell = \sqrt{\sum_{i}^{n} \zeta_i \cdot (\lambda_i - \lambda_i')^2}, \zeta_i = \frac{\|I_i\|_2}{\sum_{i}^{n} \|I_i\|_2} \tag{7}$$

where λ_i , λ_i' are the human body shape coefficients of network output and the human body shape coefficient label, ζ_i is the normalized weight coefficient according to the second normal form of shape displacement principal components. The optimizer Adadelta is used for error gradient descent optimization.

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

$$\varepsilon = \sum_{i}^{n} \zeta_i \cdot e^{-|\lambda_i - \lambda_i'|} \tag{8}$$

Taking the difference between the prediction result and the label as the input of the function of , the metrics is close to 1 when the difference is close to 0, and the larger the difference, the evaluation operator is close to 0. The weight coefficient is following (7).

3.2.2 The learning fuzzy anthropometric feature module

The fuzzy anthropometric feature matrix is taken as one of the inputs of MDD-Net. After normalization, the FC layers with the channel gradually narrowing from 4096 to 1024 are utilized to learn the feature coding of global human body shape, and the activation function Relu and Dropout layer also are utilized to increase the nonlinear learning ability of the network. The learning fuzzy anthropometric feature module of MDD-Net is as shown in the half top part of Fig. 3.

To verify the structure rationality of the learning fuzzy anthropometric feature module, under the same condition of training and testing, compare the training time and the predict accuracy of the test set as the reference of Fig. 3 structure. As shown in Fig. 4(a), when reducing the number of layers at the head or tail of the structure, it will appear under-fitting. As shown in Fig. 4(b), when increasing the number of layers at the head of the structure, the prediction effect improves slightly, and the training time increases exponentially while increasing the number of layers at the tail of the structure will cause under-fitting. As shown in Fig. 4(c), changing the number of channels in the structure either cause the training time to increase exponentially or cause the prediction effect to decrease. It explains that the structure of the learning fuzzy anthropometric feature module is optimal.

3.2.3 The learning multi-perspective silhouette feature module

Compared to the single silhouette, the multi-perspective silhouette could contain more human body shape features while is not time-consuming or expensive for the consumer. In the Sec.4.1.2 experiments, taking BfSNet (Smith et al., 2019) as an example, the predicted accuracy of single silhouette on the validation set is 53.06% which is much lower than multi-perspective 78.7% after 80 epochs. Thus, the multi-perspective silhouette is taken as the inputs of MDD-Net.

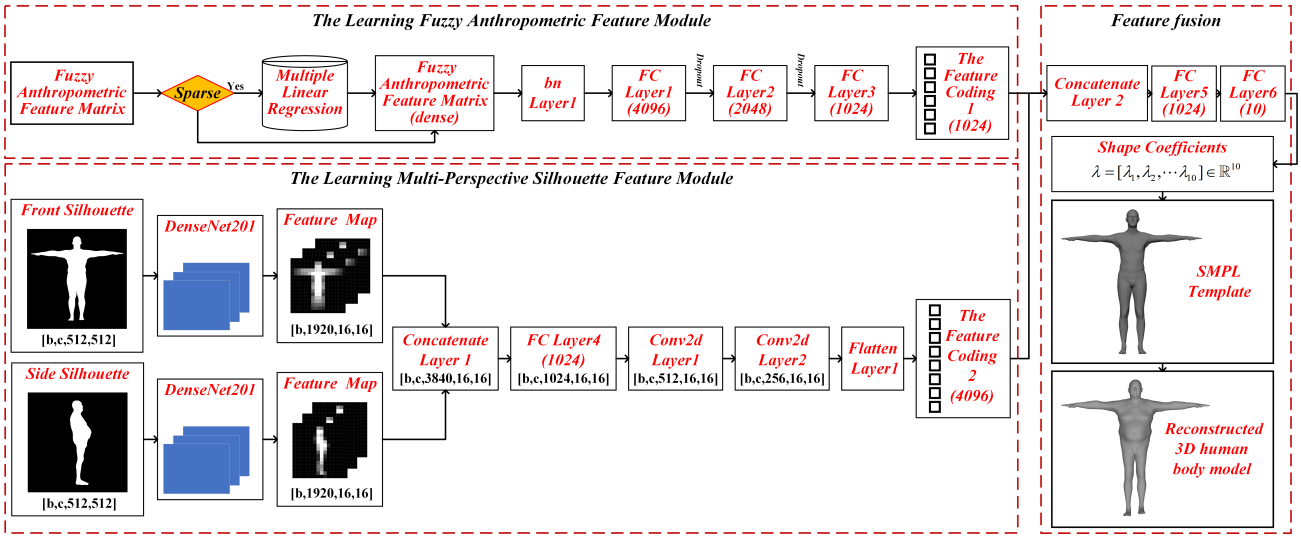

Fig. 3 Multidimensional Dense Net (MDD-Net).

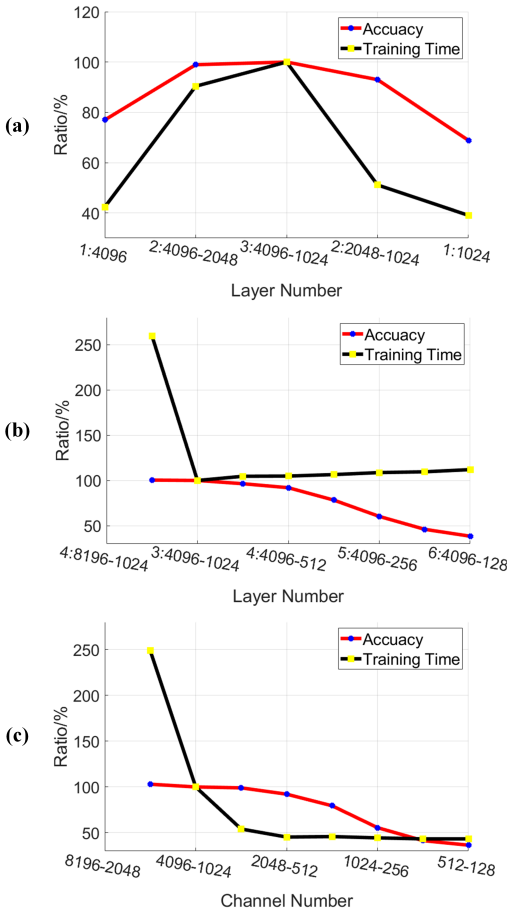

Fig. 4 The learning fuzzy anthropometric feature module structure comparison. (a) (b)different layer number comparison, 4:8196-1024 means that the layer number is 4 and the channel change from 8196 to 1024 with the gradient of 0.5. (c) different channel number comparison.

There are two methods for merging multi-perspective silhouette features. One is to stack the RGB channels of the multi-perspective silhouettes at the network input end and share the weight coefficients during the training. Another is to learn the human body feature from each perspective separately and stack the learned features at the end of the network. In the Sec.4.1.2 experiments, taking BfSNet (Smith et al., 2019) as an example, it is found that the training is unstable when stacking channels at the input end. Thus, merging the learned features of multi-perspective silhouettes at the end of the network is optimal.

The front and side silhouettes are taken as input of the learning multi-perspective silhouette feature module, the human body shape feature is learned through the DenseNet201, and the learned features at the end of DenseNet201 is merged using the concatenate layer. When merging the learned human body features, the FC layers is utilized to reduce the number of channels, the feature map resolution is reduced from 16×16 to 4×4 by two convolutional layers, then the feature map is converted into the vector with length 4096 by the flatten layer, and the shape coefficients are learned through the dense layer. The overall structure of the learning multi-perspective silhouette feature module is shown in the half bottom part of Fig. 3.

To verify the structure rationality of the learning multi-perspective silhouette feature module, under the same condition of training and testing, compare the training time and the predicted accuracy of the test set as the reference of DenseNet201. As shown in Fig.5 , DenseNet realizes feature reuse by merging the features on the channel. Each layer will establish a dense connection with all the previous layers. With a fewer number of parameters, an ultra-deep network will be established that the error back propagation speed is faster and the training time is less compared to ResNet (He et al., 2016). As shown in Fig.6 (a), among all DenseNet models, DenseNet201 has the best performance in predict-

ing the human shape coefficients by which more localized feature coding of silhouettes could be learned.

3.2.4 Feature fusion of fuzzy and multi-perspective silhouette features

The feature coding with lengths of 4096 and 1024 are learned from fuzzy anthropometric features and silhouettes at the end of MDD-Net as shown in Fig.3. Merge features and predict the shape coefficients using two FC layers. The feature fusion structure is shown in the half right part of Fig. 3.

When merging the multidimensional features, the feature ratio can be adjusted by adding an FC layer after the flatten layer. The maximum length of learned multi-perspective feature coding is 4096, and an FC layer with a length greater than 4096 will not contain more human body shape features. Thus, to verify the influence of merging ratio on the predicted accuracy, the FC layers with the length of 3072, 2048, 1024, 512, 256 are added after flatten layer respectively. As shown in Fig. 6 (b), the merge ratio has little effect on the prediction accuracy of MDD-Net, which can be almost ignored. Therefore, the ratio of 4:1 is maintained.

MDD-Net could realize the feature fusion of the human body shape in multidimensional. When a 3D human body is projected into an image, the circumference feature is flattened into a distance feature affected by the pose and position. And on account of the small receptive field of the convolution kernel, it is difficult to learn high-pixel span features at high resolution (e.g. shoulder breadth and upper and lower body proportions). To solve these problems, the global and circumferential fuzzy anthropometric features that are not affected by the pose and position are fused. In the feature fusion, the fuzzy anthropometric feature and silhouettes encode the global and local human body shape features respectively. Compared with only inputting 2d silhouettes, the method of the feature fusion learns human body shape from multi-dimensions to enhance network's robustness replacing mechanically increasing training set size

3.3 Generating human body shape fusion features dataset

As mentioned in Sec. 3.1, MDD-Net takes labeled silhouettes and fuzzy anthropometric features as input for supervised learning. To obtain the real label of shape coefficients, a large sale of real people needs to be scanned, and the 3D point cloud is aligned with the SMPL template using the rigid and non-rigid method (Groueix et al., 2018). The overall process is time-consuming, expensive and tool-dependent. Therefore, a human body shape fusion feature data set is generated to approach the shape space of real person based on the SMPL model, and is utilized to train and evaluate the MDD-Net network.

3.3.1 Generating 3D human body model

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

Table 1 3D human body model dataset details

Group	Quantity	The Shape Coefficients Range
1	500	$(-5,5], \Delta = 0.2$
2	1250	$(-5,5], \Delta = 0.04$
3	1000	$(-2,2]$
4	1100	$(-5,5]$
sum		13750

In group 1, all principal components of the shape displacement are contained to represent the human body shape space, it includes 10 subgroups with 500 human body models, and the shape coefficients are as shown in (9) where k is the group number, $R_{\min}$ is the minimum value of the shape coefficient range,Δ and is the step. In group 2, the samples size of the first 5 shape displacement principal components is increased, and it includes 5 subgroups with 1250 human body models. In group 3, the shape coefficients are randomly selected in range of $(-2,2]$ to increase the sample size of common human body shapes. And in group 4, the shape coefficients are randomly selected in range of $(-5,5]$ to expand the human body shape space, which contains some uncommon human bodies. Some 3D human body models in dataset are shown in Fig. 8(a).

$$\begin{aligned} \lambda^k &= \left[\lambda_1^k, \lambda_2^k, \cdots, \lambda_k^k, \cdots, \lambda_n^k\right] \\ \lambda_i^k &= \begin{cases} R_{\min} + k \cdot \Delta, k = i \\ 0.1, k \neq i \end{cases} \end{aligned} \tag{9}$$

3.3.2 Generating fuzzy anthropometric feature dataset

As shown in Fig.7(a), SMPL has 24 black landmarks of joint points, and 6 new red landmarks of joint points, as Table 2. The anthropometric features are generated as Table 3 to encode the global human body shape feature. In generating the leg and arm feature, the human body is set to be symmetrical. In generating the circumference feature, with the landmark benchmark l_i , the point cloud with the error of $[-\Delta, \Delta]$ is projected onto the orthogonal plane where $\Delta = 0.68\%h$, the 2D convex hull $\{p_1, \cdots, p_m\}$ of the projection point is selected to calculate its perimeter $\|p_m - p_1\| + \sum_{j=1}^{m} \|p_{j+1} - p_j\|$ as the circumference feature. The projected point cloud and convex hull are shown in Fig.7(b)-(i).

To simulate the fuzzy anthropometric feature acquired on mobile terminal, the generated accurate anthropometric feature is mapped into fuzzy anthropometric features as (10).

$$f \propto \left(f_0 + \alpha \cdot \mu_{f_0} \cdot N\left(0, \sigma_{f_0}\right)\right) \tag{10}$$

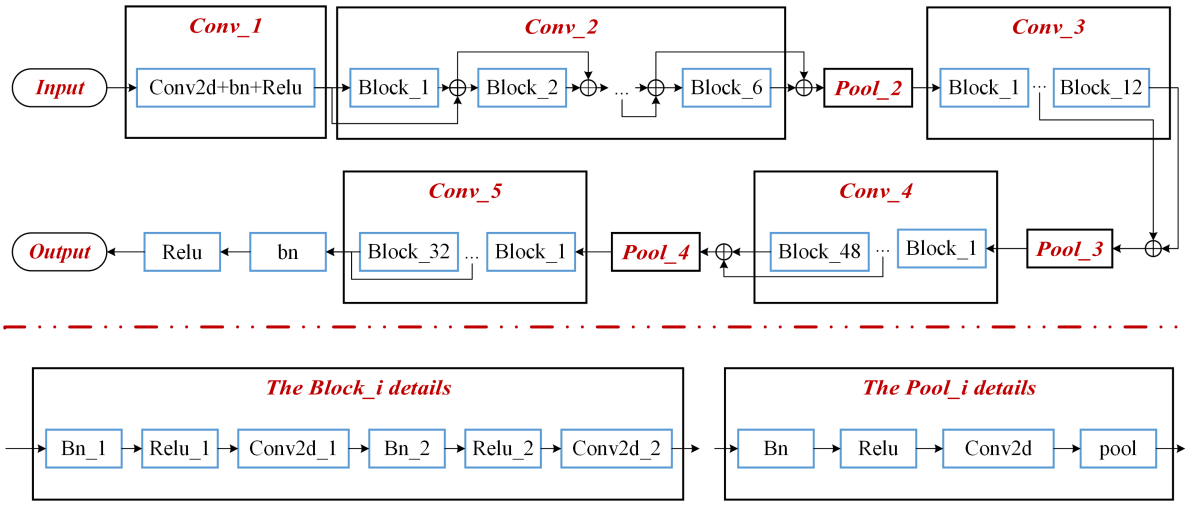

Fig. 5 The structure of DenseNet201.

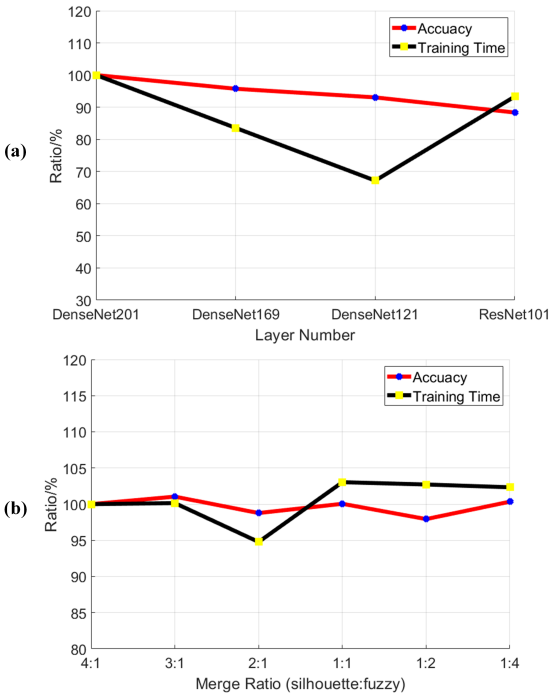

Fig. 6 (a) The learning multi-perspective silhouette feature module structure comparison. (b) The merge ratio of the multi-perspective silhouette and fuzzy anthropometric feature.

Table 2 Additional 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

where f , f_0 are the fuzzy and accurate anthropometric features respectively, α is the fuzzy coefficient, μ_{f_0} is the mean

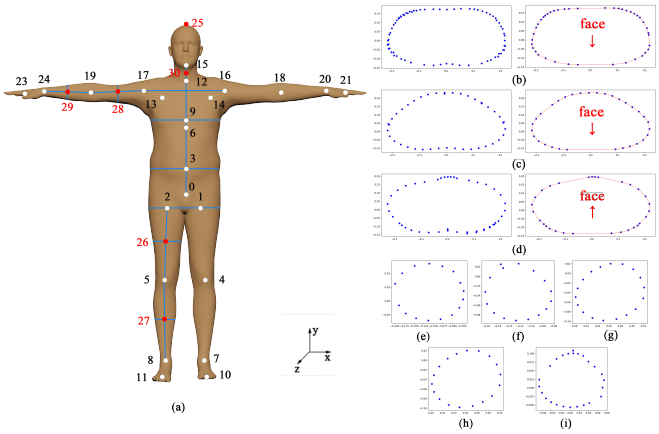

Fig. 7 The anthropometric feature of human body. (a) The representation of human body feature landmark, length and circumference features. (b) The project points and convex hull of bust. (c) waist. (d) hip. (e) thigh. (f) shank. (g) big arm. (h) forearm. (i) neck.

of f_0 in the human body shape space, $N\left(0, \sigma_{f_0}\right)$ is the standard normal distribution.

Since the human body shape space conforms to the gaussian distribution, the fuzzy degree of the acquired example-guided features also conforms to the gaussian distribution. The boundary value of the probability density function of the human body shape is much smaller than the center value, 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 $[0, 100]$.

3.3.3 Generating multi-perspective silhouette dataset

Based on OpenGL, the 3D human body model is orthogonally projected to generate front and side silhouettes with

Table 3 The anthropometric features

Feature	Sign	Note
Bust	C_1	benchmark :$l_9(y), \Delta = 0.015$
Waist	C_2	benchmark :$l_3(y), \Delta = 0.015$
Hipline	C_3	benchmark :$l_2(y), \Delta = 0.015$
Thigh Circumference	C_4	benchmark :$l_{26}(y), \Delta = 0.015$
Shank Circumference	C_5	benchmark :$l_{27}(y), \Delta = 0.015$
Neck Circumference	C_6	benchmark :$l_{30}(y), \Delta = 0.015$
After wrap Circumference	C_7	benchmark :$l_{28}(x), \Delta = 0.015$
Forearm Circumference	C_8	benchmark :$l_{29}(x), \Delta = 0.015$
Height	h	$l_{25}(y) - l_{11}(y)$
Thigh Length	L_1	$\|l_2 - l_5\|$
Shank Length	L_2	$\|l_5 - l_8\|$
After wrap Length	L_3	$\|l_{17} - l_{19}\|$
Forearm 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	$\|l_0 - l_{15}\| / (L_1 + L_2)$

the resolution of 512×512 and with a black background and a white human body foreground as shown in Fig. 8(b) and Fig. 8(c).

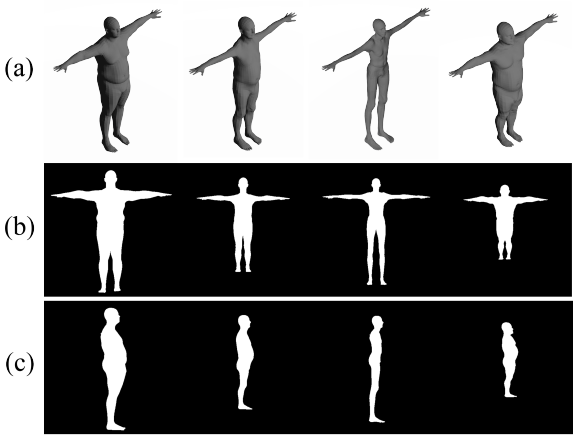

Fig. 8 The example of human body shape fusion features dataset. (a) 3D human body model. (b) The human body silhouette in the front view. (c) The human body silhouette in the side view.

3.3.4 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 four 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 which is divided from generated human body shape fusion feature dataset with 13750 samples. It ideally reflects the human body shape feature on the silhouettes.

• Segmentation noise test set. In the semantic segmentation of human body image, segmentation noise usually occurs due to noisy background environment or complicated dress. Thus, this test set is generated by randomly sowing black and white noise blocks on the silhouettes of the unprocessed test set. To prove that the influence of segmentation noise on the silhouette could not be eliminated by the traditional filter, as shown in Fig. 9, the experiments show the boundary of the filtered silhouette is distorted, the shape is overall shrunk, or more noise appears.

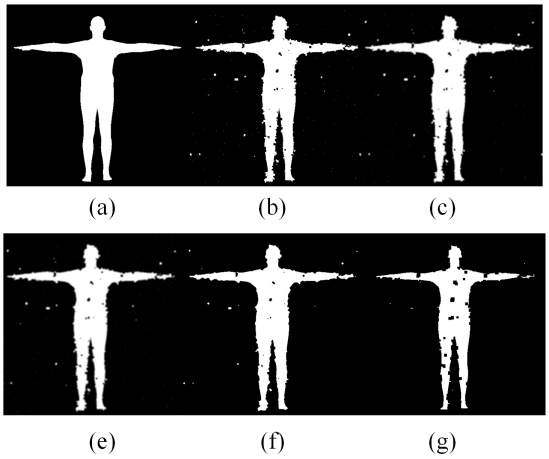

Fig. 9 Segmentation noise test set. (a) Unprocessed silhouette. (b) Silhouette with segmentation noise. (c) Gauss filtering result; (d) Mean filtering result. (e) Median filtering result. (f) Corrosion result.

• Nonstandard pose test set. When acquiring the human body silhouettes, the pose held by the people maybe has an error with the standard T pose. Keeping the shape coefficients unchanged, a rotation matrix is randomly generated for each joint in the model of the unprocessed test set, and the global rotation vector remains unchanged. As shown in Fig. 10, the samples with the pose error are contained in this test set (e.g. the two arms are not raised horizontally, the waist or the head is oblique, and the legs are close together).

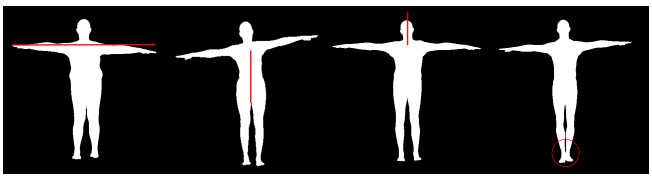

Fig. 10 Nonstandard Pose Test set.

• Perspectives error test set. As shown in Fig. 11(a), 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 kept vertical during the photographing), the acquired silhouettes present the status of looking up or looking down. If 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 of 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 $[-10°, 10°]$. An example is shown in Fig. 11(b).

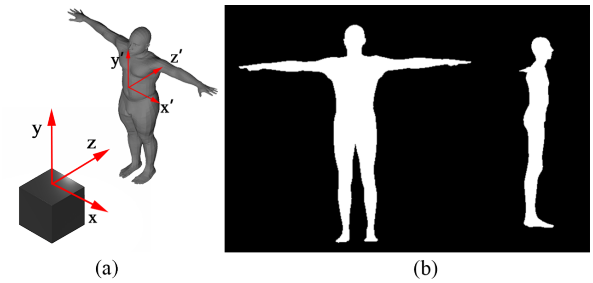

Fig. 11 Perspective error test set. (a) Perspective. (b) Silhouettes with perspective error.

4 Experiment

4.1 Training comparison

All experiments are performed on a single NVIDIA GeForce GTX 1080 GPU. The training set and validation set are introduced in Sec. 3.3, and the loss function and evaluation metrics of network are introduced in Sec. 3.1.1. Ablation experiments are carried out on the two modules of MDD-Net. And the training effects of MDD-Net under different fuzzy degrees and the predicted accuracy of other networks on the four test sets are compared in Sec. 3.3.4.

4.1.1 Ablation experiment on learning fuzzy anthropometric feature module

In this section, the importance of the coding elements in the feature matrix generated in Sec. 3.3.2 is analyzed, the default interpolation estimation method proposed in Sec3.1.1 is verified. And the predicted accuracy of the learning fuzzy anthropometric feature module with input of different fuzzy degree samples is verified.

By removing coding elements f_∂ , the learning fuzzy anthropometric feature module is only trained with the fuzzy anthropometric feature matrix formed by the remaining coding elements $\{f_i\}$. As shown in Fig. 12(b), the average importance of global length feature is 208.3% of circumference features. Thus, in acquiring the fuzzy anthropometric feature matrix, the global length features have high priority. The average importance of the elements is 6.25, the standard deviation is 4.74, and the importance distribution of each element is relatively balanced proving that the selection of fuzzy anthropometric feature elements is rational.

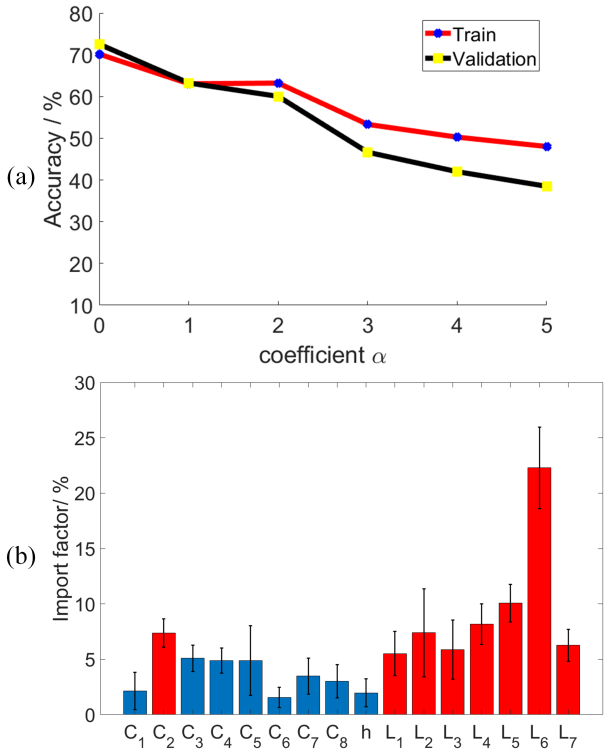

Fig. 12 The training result of the learning fuzzy anthropometric feature module. (a) the training result by inputting different fuzzy degree of samples; (b) the importance of coding elements

To verify the default interpolation estimation method proposed in Sec3.1, the coding elements with high importance are utilized to interpolating estimate the remaining default elements. The interpolation error rate under different sparsity of the fuzzy anthropometric feature matrix is analyzed in Table 4. It can be found that the error rate is about 30%, which fully meets the accuracy requirements of rough interpolation, but acquiring dense fuzzy anthropometric feature matrixes is more optimal.

The training result of this module with the input of different fuzzy degree samples are shown in Fig. 12(a). Taking accurate anthropometric features without Gaussian noise as inputs, the predicted accuracy of this module on the verification set reaches 72.54% after 160 epochs proving the possibility of reconstructing human body only with the anthropometric feature. When the sample is polluted with Gaussian noise, α in (10) is set to 0.1, the predicted accuracy of

Table 4 Interpolation error rate at various sparsity

Non-default elements number	Error rate(%)
8	28.1
9	27.0
10	28.6
11	29.4
12	30.1
13	28.5
14	34.1
15	26.6

this module on the validation set drops to 63.24%, and when the setting is 0.2, the predicted accuracy drops to 60.03%. Thus, the training result of the learning fuzzy anthropometric feature module is inversely related to the fuzzy degree of training sample. The learning ability of the learning fuzzy anthropometric feature module is still restricted by the variety of samples.

4.1.2 Training comparison of the learning multi-perspective silhouette feature module

To prove the superiority of this module in learning the human body shape feature of the silhouettes, all structures are trained only by multi-perspective silhouettes generated in Sec.3.3.3. As shown in Fig.13(a), Compared with only inputting a single silhouette, this module has higher predicted precision trained by multi-perspective silhouettes. And compared to merging RGB channels at the beginning of this module, learning the human shape features of each perspective, merging the learned features at the end of this module has a better and more stable training result.

The learning multi-perspective silhouette feature module is compared with Hs-Net and BfSNet as shown in Fig.13 (b)(c). The predicted accuracy of this module on validation set is close to BfSNet, while the training time is only 42.4% after 80 epochs. And compared with HS-Net, the predicted accuracy improves significantly by 16.4%. Thus, the learning multi-perspective silhouette feature module maintains a better balance between accuracy and training time compared to other networks.

4.1.3 Overall network comparison

To prove the superiority of input multi-dimensional fusion features for learning human body shape features, the fusion feature dataset generated in Sec.3.3 is taken as input for MDD-Net, the multi-perspective silhouettes are taken as input for other networks, and the prediction accuracy of networks are compared on four test sets generated in Sec 3.3.4.

As shown in Fig.14, in the unprocessed test set, the predicted accuracy of MDD-Net is about 89.5%, second only

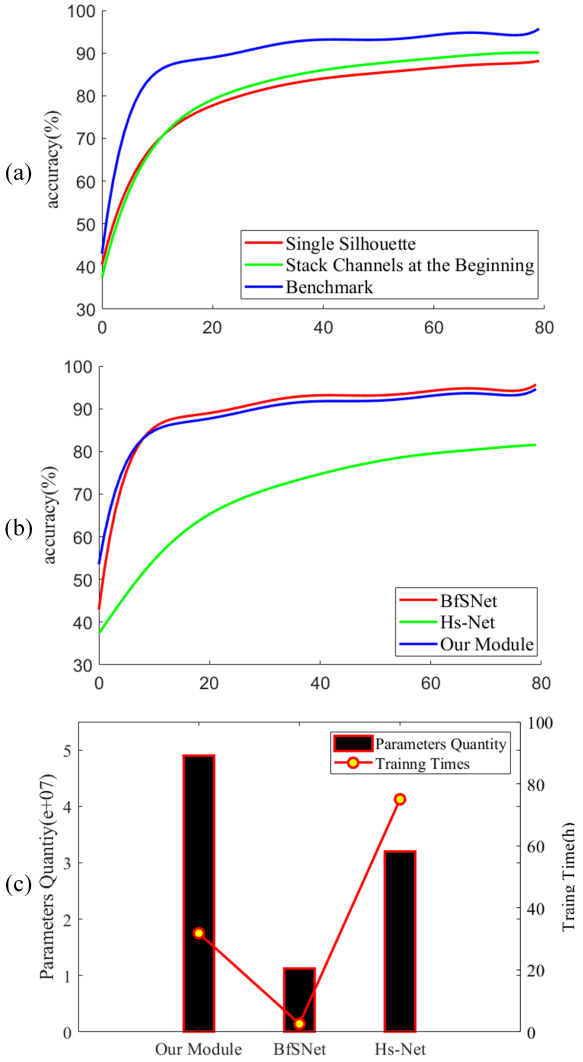

Fig. 13 Comparison between the learning multi-perspective silhouette feature module and other networks. (a)Training comparison. (b)Training parameters and time comparison.

to 92.52% of BfSNet, and far greater than 76.11% of Hs-Net. In the segmentation noise test set, MDD-Net has the best performance with a predicted accuracy closing to 50% and is 43.4% higher than BfSNet. MDD-Net and BfSNet are more sensitive to the noise in the foreground and background, resulting in a sharp decline in prediction ability. In the human body segmentation, the image noise usually appears on the shape boundary, MDD-Net could still deal with silhouettes with segmentation noise. In the nonstandard pose test set, MDD-Net improves the predicted accuracy by 37.2% compared to BfSNet. To eliminate the pose effect on the human body shape estimation, the consumers are demanded to keep T pose similar to the reference example, MDD-Net could relieve the pressure of consumer in acquiring multi-perspective silhouettes. In the perspectives error test set, MDD-Net still has 69.8% predicted accuracy and is

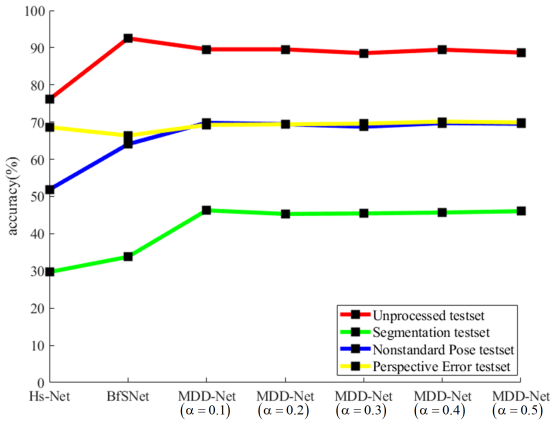

Fig. 14 The comparison between our method and other networks on four test sets.

improved by 9.0% compared to BfSNet, and the larger the perspective angle error range, the more robust of MDD-Net compared to other networks.

It founds that the performance of the MDD-Net is hardly affected by the fuzzy degree of fuzzy anthropometric feature , and the predicted accuracy is almost the same when is 0.1 or 0.5. In the ablation experiments of the learning fuzzy anthropometric feature module, the predicted accuracy decreases as increases. Thus, in the multidimensional feature fusion, MDD-Net tends to merge learned features rather than merge accumulated errors and does not guarantee global robustness of the network by limiting the sample generalization in a certain dimension. When the fuzzy anthropometric features are acquired on mobile terminal, the qualitatively features rather than quantitative features similar to guided-example are needed to prove the practicability of our method.

4.2 Anthropometry evaluation

To prove the application of our reconstructed model in the virtual try-on, the displacement error of the same instance predicted by different networks on four test sets are compared, and the features for virtual try-on are quantitatively evaluated further.

4.2.1 3D human body model displacement error

The point cloud displacement error of the same instance predicted by different networks on four test sets is shown in Fig.15. Hs-Net has the worst predicted result in the displacement error (e.g. the distorted shape in the belly and legs). On the unprocessed test set, the performance of MDD-Net is close to BfSNet and their predicted model could not be visually distinguished. However, on the segmentation noise, the nonstandard pose, and the perspective error test set, the displacement error predicted by MDD-Net is greatly smaller than BfSNet. MDD-Net could still predict a perfect human body model being close to the result of the unprocessed test set when there exists much segmentation noise, the arm is not raised, and the human body is oblique. The larger displacement error occurs on the feet and hands caused by its dense mesh, but it does not affect the expression of global human body shape. On the segmentation noise test set, segmentation noise has a great influence on the expression of local shape features, and there are large displacement errors in parts with complicated body curves (e.g. head and buttocks). On 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.

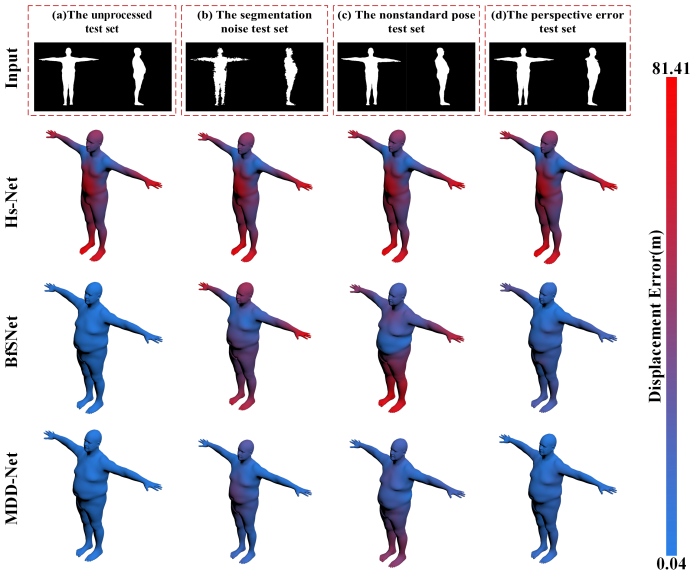

Fig. 15 The displacement error comparison.

4.2.2 Anthropometric error

The anthropomorphic estimation is to measure the human body parts values of the reconstructed 3D model as shown in Fig. 7. The purpose of it is to generate anthropomorphic features toward virtual try-on (e.g. BWH hand lengths leg lengths and shoulder breadth). In our 3D human body reconstruction, 3D human body models are utilized to generate 1d and 2d features, and then the fusion features are utilized to reconstruct 3D human body model. It is a process of nonlinear encoding and decoding of human body shape features. Anthropomorphic estimation could be taken as a standard to evaluate the encoding and decoding effect of our method.

The 3D human body models generated in Sec.3.3.1 are taken as the label and the average performance of MDD-Net under different is taken as reference. The average height

Table 5 The mean of anthropometric error.

Feature	Unprocessed			Segmentation Noise			Nonstandard Pose			Perspective Error		
	Hs	BfS	MDD	Hs	BfS	MDD	Hs	BfS	MDD	Hs	BfS	MDD
C_1	12.5	2.9	5.9	74.5	76.7	63.7	44.0	50.7	33.8	28.3	33.2	16.9
C_2	17.0	5.4	9.8	65.2	123.0	76.0	44.9	46.6	42.5	28.1	25.7	24.8
C_3	28.2	8.4	10.4	73.0	96.1	63.9	73.9	71.4	56.3	45.0	39.5	37.1
C_4	11.8	3.1	4.7	56.5	40.5	35.9	35.9	36.7	24.9	26.0	25.0	14.6
C_5	11.9	3.2	4.2	36.3	47.3	32.2	44.3	40.7	32.1	28.8	24.2	18.2
C_7	2.8	0.8	1.9	22.4	30.8	17.2	13.5	11.4	10.7	6.9	4.7	3.8
C_8	7.0	1.7	3.0	32.9	35.3	33.7	30.2	24.2	24.5	17.5	13.0	8.9
L_3+L_4	6.5	2.2	3.3	21.1	33.7	24.7	86.2	52.6	21.4	32.8	10.6	7.6
L_1+L_2	4.5	1.4	1.7	26.2	34.8	11.2	28.3	28.0	10.2	15.1	8.0	5.0
L_5	4.7	0.9	1.5	22.6	20.6	15.6	11.7	14.3	10.5	6.1	10.8	4.6

Table 6 The standard deviation of anthropometric error.

Feature	Unprocessed			Segmentation Noise			Nonstandard Pose			Perspective Error		
	Hs	BfS	MDD	Hs	BfS	MDD	Hs	BfS	MDD	Hs	BfS	MDD
C_1	0.25	0.11	0.15	0.78	0.76	0.68	0.72	0.78	0.54	0.46	0.57	0.32
C_2	0.39	0.21	0.28	0.72	1.27	1.26	0.77	0.79	0.66	0.61	0.49	0.55
C_3	0.71	0.39	0.44	1.36	1.38	1.29	1.30	1.26	1.04	0.92	0.85	0.84
C_4	0.28	0.18	0.21	0.56	0.59	0.51	0.60	0.55	0.45	0.43	0.42	0.35
C_5	0.41	0.21	0.27	0.51	0.71	0.54	0.83	0.82	0.71	0.66	0.60	0.54
C_7	0.08	0.04	0.06	0.18	0.26	0.22	0.23	0.20	0.18	0.15	0.13	0.10
C_8	0.19	0.09	0.13	0.43	0.49	0.44	0.54	0.44	0.44	0.44	0.38	0.24
L_3+L_4	0.11	0.04	0.06	0.35	0.52	0.42	1.33	0.92	0.35	0.67	0.22	0.13
L_1+L_2	0.08	0.02	0.03	0.32	0.45	0.20	0.44	0.46	0.20	0.27	0.15	0.08
L_5	0.09	0.02	0.04	0.26	0.29	0.25	0.19	0.23	0.18	0.10	0.19	0.08

Table 7 The truncation score of anthropometric error.

Feature	Unprocessed			Segmentation Noise			Nonstandard Pose			Perspective Error		
	Hs	BfS	MDD	Hs	BfS	MDD	Hs	BfS	MDD	Hs	BfS	MDD
C_1	0.49	0.98	0.85	0.02	0.02	0.07	0.14	0.11	0.17	0.22	0.21	0.38
C_2	0.43	0.91	0.77	0.12	0.01	0.06	0.13	0.13	0.14	0.28	0.28	0.31
C_3	0.37	0.86	0.66	0.07	0.06	0.09	0.10	0.10	0.14	0.20	0.25	0.29
C_4	0.33	0.93	0.79	0.01	0.06	0.07	0.09	0.08	0.15	0.12	0.15	0.29
C_5	0.40	0.91	0.75	0.04	0.02	0.07	0.07	0.07	0.13	0.15	0.19	0.30
C_7	0.58	0.97	0.74	0.01	0.01	0.07	0.12	0.13	0.13	0.24	0.39	0.43
C_8	0.32	0.93	0.75	0.03	0.05	0.05	0.07	0.08	0.09	0.17	0.25	0.30
L_3+L_4	0.65	0.98	0.93	0.21	0.13	0.22	0.06	0.10	0.23	0.20	0.51	0.59
L_1+L_2	0.63	0.99	0.97	0.07	0.06	0.29	0.11	0.11	0.34	0.23	0.43	0.58
L_5	0.48	0.99	0.86	0.05	0.09	0.15	0.20	0.15	0.23	0.37	0.27	0.51

h of 3D human body model on the unprocessed test set is 1.793, the average height H of Chinese adult males is 169.7 cm (Office, 2020), and the coefficient $c = H/h$ is utilized to map the point cloud error into real length. The anthropomorphic estimation is analyzed in terms of robustness from three metrics: mean value, standard deviation, and truncation score. The definition of truncation score ρ is shown in (11) where ξ is the cutoff coefficient, e , $\overline{f_0}$ are error and mean human body parts measured value, the $\varepsilon()$ is the activation function. Taking ξ as 0.01, the reconstructed 3D human body model is qualified on particular part and ρ equals 1 when the measured value error is less than 1% of the mean human body parts measured value, otherwise ρ equals 0.

$$\rho = \varepsilon\left(\xi \cdot \overline{f_0} - e\right) \tag{11}$$

As shown in Table 5,6,7, 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.

5 Discussion

This paper proposes a 3D human body reconstruction network: Multidimensional Dense Net (MDD-Net) that is toward virtual try-on on mobile terminal. It is a supervised deep learning network that inputs fuzzy anthropometric feature matrix and the multi-perspective silhouettes acquired on the mobile and estimates accurate human body shape. 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 data set based on SMPL model are proposed. The mobile terminal acquires fuzzy anthropometric features and multi-perspective silhouettes and displays the reconstructed 3D human body model. Based on SMPL, a fuzzy anthropometric feature data set with Gaussian noise, a silhouettes data set, and test sets affected by pose and position are generated. An interpolating estimation method for a fuzzy anthropometric feature matrix is proposed.

(2) The human body shape estimation network MDD-Net is proposed. The learning fuzzy anthropometric feature module and the learning multi-perspective silhouette feature module are proposed to learn the global and local human body features from multidimensional. And the feature fusion is utilized to merge the learned feature at the ratio of 4:1 to predict the shape coefficients of 3D human body.

(3) The MDD-Net is quantitatively analyzed and compared with the existing networks of Hs-Net and BfSNet. Experiments show that without increasing the fault-tolerant training samples, the accuracy of MDD-Net on the segmentation noise, nonstandard pose, and perspective error test sets are improved by 13.34%, 55.77%, 34.6%, and 43.4%, 37.2% 9.0% compared to Hs-Net and BfSNet. In the anthropometric evaluation, MDD-Net has a small error and standard deviation in critical anthropometric features, indicating the effectiveness of our method.

When acquiring the silhouette of the human body on mobile, the garments of consumers still affect the reconstruction accuracy of 3D human body model. In future work, we will generate dressed silhouettes using generation model (Lassner et al., 2017a), while collecting clothing prior features, further improve the robustness of 3D human body reconstruction network under the influence of pose and position.

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.

6 Compliance with ethical standards

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

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].

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

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

7 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.

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. Computer-Aided Design 44(1):56–67

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 LC, Papandreou G, Kokkinos I, Murphy K, Yuille AL (2017) Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE transactions on pattern analysis and machine intelligence 40(4):834–848

Chen Y, Cheng ZQ, Lai C, Martin RR, Dang G (2015) Realtime reconstruction of an animating human body from a single depth camera. IEEE transactions on visualization and computer 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, Oztireli 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

Hornung A, Dekkers E, Kobbelt L (2007) Character animation from 2d pictures and 3d motion data. ACM Transactions on Graphics (ToG) 26(1):1–es

Huang G, Liu Z, Van Der Maaten L, Weinberger KQ (2017) Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4700–4708

Jiang L, Ye J, Sun L, Li J (2019) Transferring and fitting fixed-sized garments onto bodies of various dimensions and postures. Computer-Aided Design 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

Leroy V, Franco JS, Boyer E (2017) Multi-view dynamic shape refinement using local temporal integration. In: Proceedings of the IEEE International Conference on Computer Vision, pp 3094–3103

Li J, Yu Q, Xu H, Lu G, Zhang D (2019) Measuring and modeling human bodies with a novel relocatable mechatronic sensor-net. Textile Research Journal 89(19-20):4131–4147

Lin G, Milan A, Shen C, Reid I (2017) Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1925–1934

Loper M, Mahmood N, Romero J, Pons-Moll G, Black MJ (2015) Smpl: A skinned multi-person linear model. ACM transactions on graphics (TOG) 34(6):1–16

Office TSCI (2020) The report on nutrition and chronic diseases in china (2020). URL `http://www.scio.gov.cn/xwfbh/xwbfbh/wqfbh/42311/44583/wz44585/Document/1695276/1695276.htm`

Omran M, Lassner C, Pons-Moll G, Gehler P, Schiele B (2018) Neural body fitting: Unifying deep learning and model based human pose and shape estimation. In: 2018 international conference on 3D vision (3DV), IEEE, pp 484–494

Popa AI, Zanfir M, Sminchisescu C (2017) Deep multitask architecture for integrated 2d and 3d human sensing. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 6289–6298

Rhodin H, Robertini N, Casas D, Richardt C, Seidel HP, Theobalt C (2016) General automatic human shape and motion capture using volumetric contour cues. In: European conference on computer vision, Springer, pp 509–526

Robinette KM, Blackwell S, Daanen H, Boehmer M, Fleming S (2002) Civilian american and european surface anthropometry resource (caesar), final report. volume 1. summary. Tech. rep., Sytronics Inc Dayton Oh

Smith BM, Chari V, Agrawal A, Rehg JM, Sever R (2019) Towards accurate 3d human body reconstruction from silhouettes. In: 2019 International Conference on 3D Vision (3DV), IEEE, pp 279–288

Tan J, Budvytis I, Cipolla R (2017) Indirect deep structured learning for 3d human body shape and pose prediction. In: British Machine Vision Conference 2017, BMVC 2017

Tung HYF, Tung HW, Yumer E, Fragkiadaki K (2017) Self-supervised learning of motion capture. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, pp 5242–5252

Uhm T, Park H, Park JI (2015) Fully vision-based automatic human body measurement system for apparel application. Measurement 61:169–179

Wang K, Xie J, Zhang G, Liu L, Yang J (2020) Sequential 3d human pose and shape estimation from point clouds. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 7275–7284

Weng CY, Curless B, Kemelmacher-Shlizerman I (2019) Photo wake-up: 3d character animation from a single photo. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 5908–5917

Wuhrer S, Shu C (2013) Estimating 3d human shapes from measurements. Machine vision and applications 24(6):1133–1147

Xiang D, Joo H, Sheikh Y (2019) Monocular total capture: Posing face, body, and hands in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 10965–10974

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

Xu L, Su Z, Han L, Yu T, Liu Y, Fang L (2019) Unstructuredfusion: realtime 4d geometry and texture reconstruction using commercial rgbd cameras. IEEE transactions on pattern analysis and machine intelligence 42(10):2508–2522

Yu T, Zheng Z, Guo K, Zhao J, Dai Q, Li H, Pons-Moll G, Liu Y (2018) Doublefusion: Real-time capture of human performances with inner body shapes from a single depth sensor. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7287–7296

Zanfir A, Marinoiu E, Sminchisescu C (2018) Monocular 3d pose and shape estimation of multiple people in natural scenes-the importance of multiple scene constraints. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 2148–2157

Zanfir A, Bazavan EG, Xu H, Freeman WT, Sukthankar R, Sminchisescu C (2020) Weakly supervised 3d human pose and shape reconstruction with normalizing flows. In: European Conference on Computer Vision, Springer, pp 465–481

Zhang Y, Zheng J, Magnenat-Thalmann N (2015) Example-guided anthropometric human body modeling. The Visual Computer 31(12):1615–1631

Zhao T, Li S, Ngan KN, Wu F (2018) 3-d reconstruction of human body shape from a single commodity depth camera. IEEE Transactions on Multimedia 21(1):114–123

Zollhöfer M, Nießner M, Izadi S, Rehmann C, Zach C, Fisher M, Wu C, Fitzgibbon A, Loop C, Theobalt C, et al. (2014) Real-time non-rigid reconstruction using an rgb-d camera. ACM Transactions on Graphics (ToG) 33(4):1–12