Parametric Shape Estimation of Human Body Under Wide Clothing

Yucheng Lu, Jin-Hyuck Cha, Se-Kyoung Youm, and Seung-Won Jung, *Senior Member, IEEE*

***Abstract*—The shape of the human body plays an important role in many applications, such as those involving personal healthcare and virtual clothing try-ons. However, accurate body shape measurements typically require the user to be wearing a minimal amount of clothing, which is not practical in many situations. To resolve this issue using deep learning techniques, we need a paired dataset of ground-truth naked human body shapes and their corresponding color images with clothes. As it is practically impossible to collect enough of this kind of data from real-world environments to train a deep neural network, in this paper, we present the Synthetic dataset of Human Avatars under wiDE gaRment (SHADER). The SHADER dataset consists of 300,000 paired ground-truth naked and dressed images of 1,500 synthetic humans with different body shapes, poses, garments, skin tones, and backgrounds. To take full advantage of SHADER, we propose a novel silhouette confidence measure and show that our silhouette confidence prediction network can help improve the performance of state-of-the-art shape estimation networks for human bodies under clothing. The experimental results demonstrate the effectiveness of the proposed approach. The code and dataset are available at https://github.com/YCL92/SHADER.**

***Index Terms*—Silhouette confidence, convolutional neural network, human shape estimation, synthetic dataset.**

I. Introduction

In THE field of healthcare research, body shape is often used to evaluate overall health. For example, waist size can be used to estimate the amount of belly fat and the waist-to-hip ratio, which has been introduced as risk indicator of certain health problems [1], [2]. Aside from healthcare monitoring, body shape also plays an important role in interactive applications such as virtual try-on and avatar synthesis [3], [4]. In such applications, customers can virtually try on clothes and immediately see their appearance on a screen without having to put on the chosen clothes. These applications can also allow clothing stores to provide customers with better recommendations for garments based on their body shapes. Virtual avatars that have body shapes similar to the user can be useful in immersive games and virtual reality applications.

Although body shape plays an important role in these applications, direct measurements can be time-consuming and often involve privacy issues. Since the overall body shape of a person can be roughly estimated at a glance, many algorithms have been developed to estimate body shapes directly from images. Early attempts in this field used an optimization framework which iteratively minimizes the error between the projection of the estimated 3D human body models and 2D annotations in the form of silhouettes and joint locations [5], [6]. The human body models reconstructed using these methods are mostly accurate, but at the expense of huge computational complexity. In addition, precise initialization of 3D models is essential to avoid getting stuck in local minima.

With the success of convolutional neural networks (CNNs) in many computer vision tasks, they have recently come to be used for human body shape estimation. A straightforward approach is to reconstruct a 3D mesh directly from an input color image by a single forward inference pass [7], [8]. However, the resulting resolution in terms of the number of point clouds or voxels is low, making the result less practical for further use. Since the human body shape is symmetric and structural, it can be represented by parametric models such as shape completion and animation of people (SCAPE) [9] as well as skinned multi-person linear (SMPL) models [10]. The deep neural networks that have recently been proposed [11]–[16] are trained to predict the parameters of these models, and they have demonstrated superiority over direct 3D mesh estimation methods.

Although significant progress has been made in human body shape estimation, the estimation of body shapes under wide clothing remains a challenge. This is particularly true because most of the existing datasets [17]–[19] only provide either naked human body scans or bodies with minimal clothing, and some of them even use images that were captured in controlled environments. If shape estimation networks are trained using these datasets to minimize the error between the projection of 3D body models and 2D annotations such as silhouettes and joint locations, they cannot be well generalized to unseen human bodies wearing wide and loose garments.

In this paper, in an attempt to make estimations of body shape under wide clothing feasible, we present the Synthetic dataset of Human Avatars under wiDE gaRment (SHADER). The

Manuscript received June 15, 2020; revised September 8, 2020; accepted October 5, 2020. Date of publication October 9, 2020; date of current version October 19, 2021. This work was supported by the Samsung Research Funding & Incubation Center of Samsung Electronics under Project No. SRFC-IT1801-11. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Wen-Huang Cheng. *(Corresponding author: Seung-Won Jung).*

Yucheng Lu and Jin-Hyuck Cha are with the Department of Multimedia Engineering, Dongguk University, Seoul 04620, South Korea (e-mail: yucheng.l@outlook.com; ckwlsgur20@gmail.com).

Se-Kyoung Youm is with the Department of Industrial and Systems Engineering, Dongguk University, Seoul 04620, South Korea (e-mail: skyoum@dgu.edu).

Seung-Won Jung is with the Department of Electrical Engineering, Korea University, Seoul 02841, Korea (e-mail: swjung83@korea.ac.kr).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TMM.2020.3029941.

Digital Object Identifier 10.1109/TMM.2020.3029941

SHADER dataset contains 300,000 paired naked and dressed color images of 1,500 synthetic human avatars with different shapes, poses, skin tones, background environments, and most importantly, wide clothing with realistic deformations. It also includes 2D silhouettes for both dressed and naked humans, 2D and 3D joint locations, and SMPL parameters. To build SHADER, we first design a semi-automatic pipeline that applies garments to 3D synthetic human avatars according to their sizes and simulates realistic cloth deformations in response to different poses. These 3D garments and avatar meshes are then used to construct the dataset. Moreover, to take full advantage of SHADER, we propose a novel silhouette confidence and train a network to predict it given an input color image. The results of extensive experiments show that the proposed silhouette confidence can improve the performance of state-of-the-art methods in estimating body shapes under wide clothing.

The rest of this paper is organized as follows: Section II reviews the existing research related to our work. Section III introduces the proposed SHADER dataset. Section IV gives more details about the generation of SHADER. Section V presents our proposed neural network for estimating human body shape. Section VI offers the implementation details and experimental results. Section VII discusses the limitation and future work. Finally, Section VIII summarizes our work.

II. Related Works

A. Human Pose Estimation

Human pose estimation has been extensively studied due to the promise it shows for applications such as gesture recognition and human behavior analysis [20], [21]. While we focus on human shape estimation, many human shape estimation techniques are motivated by or extended from human pose estimation techniques.

Compared to classical pose estimation methods [22]–[26], recent studies tend to predict human poses in an end-to-end manner using CNNs. The pioneering work of DeepPose, as proposed by Toshev *et al.* [27], formulated human pose estimation as a joint regression problem that used a deep CNN to directly predict 2D joint coordinates. In another study, Tompson *et al.* [28] applied their network multiple times using images with different spatial resolutions to obtain joint heat maps. Similarly, Fan *et al.* [29] used not only the local part appearance but also a holistic view of each local part as the inputs of their dual-source CNN.

The stacked hourglass architecture proposed by Newell *et al.* [30] has received particularly extensive attention. This architecture has shown promising results in predicting structural information such as human pose. Each hourglass module in the network predicts joint heat maps with intermediate supervision, while the stacked structure allows high level features to go back and forth across different scales. Dabral *et al.* [31] further extended the stacked hourglass network for 3D pose estimation by first training a 2D pose estimation sub-network and then training the whole network to predict both 2D joint locations and depth maps. In addition, Zhou *et al.* [32] trained their network in a weakly supervised manner by adding a constraint on bone length ratio to the training loss. Luvizon *et al.* [33] re-used the visual features and joint probability maps during pose estimation and proposed a multitask network for joint pose estimation and action recognition. Yang *et al.* [34] replaced commonly used handcrafted constraints with a multi-input discriminator to enforce the generator to predict anthropometrically valid human pose.

B. Human Shape Estimation

Human shape estimation techniques can be classified by the types of inputs they take: 3D data, such as volumetric scans or depth maps, or 2D images. Human shape estimation can be more easily performed using 3D data, since they inherently provide geometrical information about the human body. Yu *et al.* [35] made estimations from a sequence of depth maps and updated the SMPL model at every time step using a non-linear regressor. Similarly, Mishra *et al.* [36] first fitted the SMPL model using depth maps and then generated a consensus mesh with clothing details. In another study, to better estimate body shape under loose clothing, Yang *et al.* [37] proposed a clothing energy term to constrain the predicted shape to be within the volumetric observation.

Predicting human shape solely from a 2D image is obviously more challenging than doing so from 3D data, but it is more desirable in many practical applications. One intuitive approach is to use the human body silhouette, as it heavily reflects a persons body shape. Song *et al.* [38] collected pairs of dressed silhouette and naked body landmarks to train a 3D body landmark regressor, which was later used to fit the SMPL model. Instead of a single-view silhouette, Ji *et al.* [39] presented two parallel networks that could predict body shapes using a pair of silhouettes from two orthogonal viewpoints. Tan *et al.* [40] trained a network to directly predict the SMPL model and pose vectors from a color image. They also designed a decoder network to generate human silhouettes using the predicted SMPL model, which were then used to measure a silhouette projection loss such that the network could be trained in an end-to-end manner. Pavlakos *et al.* [11] proposed a feature extractor based on stacked hourglass modules to simultaneously generate human silhouettes and 2D joint maps, which were then processed by two sub-networks to obtain the final SMPL model. In contrast to Tan et al. they used a partially differentiable renderer to compute the projection error and used joint projection error to increase the accuracy.

To further improve performance, Omran *et al.* [12] replaced human silhouettes with body-part segmentation. Varol *et al.* [13] used the combination of a color image, body-part segmentation, 2D pose, and 3D pose to achieve better performance. Xu *et al.* [15] estimated a body pixel-to-surface correspondence map called IUV map using the network architecture from DensePose [41]. The IUV map was treated as an intermediate representation of 3D shape and used to train the SMPL regression network. Rong *et al.* [16] investigated the effectiveness of different types of annotations including the IUV map and body-part segmentation. Taking a different approach, Kolotouros *et al.* [14], [42] focused on the unification of optimization-based and regression-based approaches. In [42], 3D locations of the body mesh vertices were directly regressed

through a Graph-CNN, which were then fitted to SMPL parameters to obtain a smoother mesh. In [14], they used an iterative optimization-based method from SMPLify [43] to supervise the SMPL regression network, which yielded state-of-the-art performance. Pavlakos *et al.* [44] introduced cloth texture consistency across multi-view color images as a natural form of supervision signal and used it as the loss term to train the SMPL regression network, which also led to state-of-the-art performance.

C. Publicly Available Datasets

Although recent CNN-based human shape estimation techniques have shown great robustness, accuracy, and efficiency, these CNNs must be trained using a massive amount of training data with the annotated ground-truth. Among the publicly available human body-related datasets, only a few provide both pose and shape annotations. Human3.6M [17] is one of the most widely used large-scale real-world datasets that provides pose annotations, silhouettes, depth maps, and body scans. Even though Human3.6 M contains sequential multi-view observations of each subject, which are particularly suitable for human dynamics prediction [45]–[48] and human pose estimation [31]–[34], [49], the small number of subjects (six males and five females) included makes it less useful in training human shape estimation networks. Lassner *et al.* [18] presented a dataset called Unite the People that has richer annotations from a large variety of body shapes found in real photos. It has been used in training body shape prediction networks [11], [12], [16], [40] and has proven to be practical. However, the ground-truth shapes in this dataset were obtained using an extended version of the SMPLify method [43], which does not specifically address the effect of clothes, and thus the reliability of the ground-truth for wide clothing is not guaranteed.

In general, it is nearly impossible to collect real human photos with accurate ground-truth body shapes under wide clothing. It is thus more convenient to generate synthetic data using parametric human models. Pishchulin *et al.* [49] used morphable model deformation to obtain a large variety of shapes and applied random real-world backgrounds to make the synthesized images appear more realistic. In a different approach, Chen *et al.* [50] focused on enriching the texture diversity of clothing by applying combinations of garments to the computer-generated SCAPE models according to contour matching. Varol *et al.* [19] proposed a dataset called SURREAL that uses 3D motion capture (MoCap) data, randomly chosen SMPL shape vectors, skin patterns, clothing textures, and background images to generate realistic 3D models and their corresponding 2D rendered images. Owing to its synthetic nature, the SURREAL dataset provides more accurate ground-truth shapes with a larger variety of annotations in terms of 2D/3D joint locations, depth maps, and body partitions, leading to improved performance in both pose and shape estimation, as reported in [13].

Although these synthesized datasets provide more accurate ground-truth shapes, they cannot be used for the precise modeling of human bodies under wide clothing because of the lack of natural deformation in their simulated garments. People in daily life often wear wide and loose garments for which the distance to skin is not negligible. To improve estimation performance for such situations, a dressed human dataset is necessary. Several recent studies applied clothes to naked body meshes. For example, Gundogdu *et al.* [51] fed a naked body mesh with a roughly aligned garment mesh to a network to produce a deformed garment mesh and applied it to obtain a dressed body mesh. Bhatnagar *et al.* [52] proposed a method that estimates a garment mesh from multi-view color images of a person in the standard T-pose. Although these automatic garment mesh generation methods [51], [52] may seem promising, estimated garment meshes can have less physically correct garment deformations. Keeping this in mind, we used physics-based simulation software [53] with human interactions to obtain garment meshes that are as accurate as possible for our dataset. To our knowledge, 3DPeople [54] is the work most closely related to ours. However, although the 3DPeople dataset is rich in terms of having a variety of garments that even includes sunglasses, hats, and caps, it was mainly developed to model the geometry of dressed humans, and lacks diversity in its body shapes (40 subjects for each gender). By contrast, our dataset includes diverse body shapes (750 subjects for each gender), and thus the naked body shape can be estimated from a single color image without the need for subsidiary shape representations.

(a) (b) (c)

Fig. 1. Image examples from three different datasets: (a) Human 3.6M [17], (b) SURREAL [19], and (c) Unite the People [18] (left: original color image, right: corresponding ground-truth).

III. The SHADER Dataset

We first describe our motivation for the estimation of body shape under wide clothing, then explain the silhouette confidence.

A. Motivation

When a person is wearing minimal and tight clothing, the gap between his or her clothes and skin is marginal, and the body and clothes experience similar transformations when the person is moving. For human shape estimation in such a case, special care may not need to be taken for clothing. Fig. 1 shows some examples of several publicly available datasets [17]–[19]. It appears that two of the datasets, namely Human 3.6M [17] and SURREAL [19], focus more on the diversity of body poses and background scenes, as their clothing settings are rather simple. In particular, SURREAL has no real clothing structures on bodies. Another dataset called Unite the People [18] provides a rich variety of garments, but the ground-truth body shapes are estimated by an extended version of the SMPLify method [43] so that the accuracy is often limited for subjects under wide clothing, as shown in Fig. 1(c). Consequently, most of the previous

TABLE I
COMPARISON OF DATASETS

Dataset	SMPL parameters	Large shape space	MoCap data	Dressed images	Accurate shapes	Accurate 2D joints	Accurate 3D joints
Human3.6M [17]	✓	✗	✓	✗	✓	✓	✓
Unite the People [18]	✓	✓	✗	✓	✗	✗	✗
SURREAL [19]	✓	✓	✓	✗	✓	✓	✓
3DPeople [55]	✗	✗	✓	✓	✓	✓	✓
SHADER	✓	✓	✓	✓	✓	✓	✓

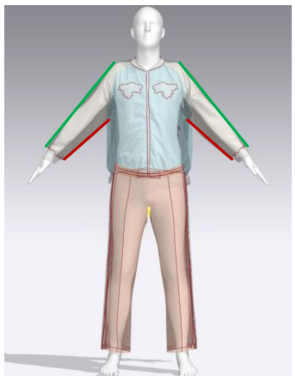


Fig. 2. Illustration of body shape under wide clothing.

deep learning-based methods [11]–[16] were trained and tested using these limited datasets, and are thus not optimized for more general situations. A more detailed comparison of the datasets is provided in Table I.

When we observe a person's body shape from the outfit, our intuition is to focus more on the tightness of clothes to skin. For example, one can easily tell the shape difference between an obese person with tight clothes and a slim person with loose clothes even if their dressed body silhouettes are similar. In other words, the inner-body boundary of the former case is tightly aligned with the silhouette, whereas there are sufficient gaps between the inner-body boundary and the silhouette in the latter case. We therefore attempt to exploit the capability of the neural network to learn this difference. Fig. 2 shows an example of a body under wide clothing. As shown in the figure, the gaps between clothes and skin are not marginal and may not even remain constant in all cases. The upper boundaries of the silhouette, marked in green, are closer to the skin, whereas the lower boundaries, marked in red, are distant from the skin due to gravity and the pose taken. For this example, it is clear that the upper boundaries should be treated as more reliable than the lower boundaries when estimating the body shape; otherwise, the body shape can be over-estimated. Since SHADER includes silhouettes for both dressed and naked humans as well as realistic garment deformation, we can transform our intuition to quantitative measurements such that a neural network can precisely estimate a body shape under wide clothing.

B. Silhouette Confidence

Based on the discussion in Section III-A, we define silhouette confidence, which is dependent on the distance from the silhouette of the naked body to that of the dressed body. Specifically, we extend the commonly used binary silhouette, whose pixel values are ones and zeros for pixels inside and outside the body region, respectively. Considering the availability of both dressed and naked silhouettes in SHADER, we can assign zero to the pixels outside the dressed silhouette and one to the pixels inside the naked silhouette. However, the pixels that are outside the naked silhouette but inside the dressed silhouette require a special designation, because their reliability cannot be the same due to the ambiguity of cloth deformation in certain areas, e.g., the edges of the sleeves marked in red in Fig. 1. Therefore, a more flexible method would involve weighting pixels according to their probability of being the body shape.

Let P_{body}, P_{cloth}, and P_{bg} denote a set of pixels inside the binary silhouette of the naked human body, a set of pixels inside the binary silhouette of the dressed human body but outside the silhouette of the naked human body, and a set of background pixels, respectively. The confidence values for the pixels in P_{body} and P_{bg} are first set as 1 and 0, respectively. We then apply distance transform [55] to all pixels in P_{cloth} to obtain their distances to the closest neighbors in P_{body}. For each pixel $p_{i,j} \in P_{cloth}$ at location (i, j), its distance $d_{i,j}$ is obtained as follows:

$$d_{i,j} = \min_{p_k \in P_{body}} \|p_{i,j} - p_k\|_1 , \tag{1}$$

where $\|p_{i,j} - p_k\|_1$ measures the L_1 distance between two pixel coordinates. The distance $d_{i,j}$ is ultimately mapped to a confidence value $c_{i,j}$ using a Gaussian kernel as follows:

$$c_{i,j} = \frac{1}{\sqrt{2\pi\sigma_c^2}} \exp\left(-\frac{d_{i,j}^2}{2\sigma_c^2}\right), \tag{2}$$

where the standard deviation σ_c controls the decay rate of the confidence values.

It is worth mentioning that the motivation of the silhouette confidence is to extract reliable parts of the human body under wide clothing rather than to estimate the full body shape, whereas the binary silhouettes were frequently used in representing full shapes in the previous methods. In the regions where the margins between cloth and skin are not negligible, simply setting the pixels in P_{cloth} to zero can make the confidence estimation difficult. Due to the symmetric structure of human bodies, the shape can still be recovered even when only one side of the body part has high reliability. We will demonstrate the effectiveness of the proposed silhouette confidence in Section VI.

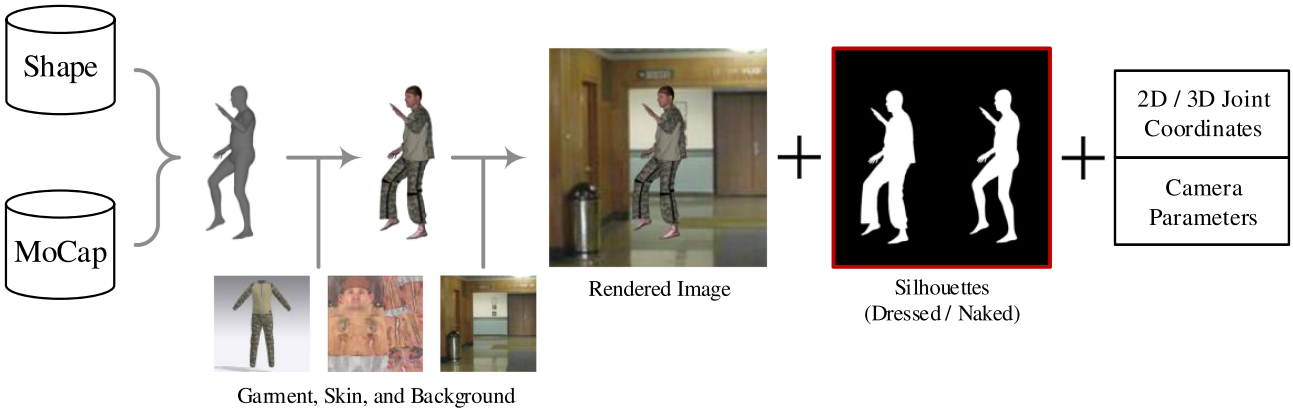


Fig. 3. Overview of the data generation pipeline of the SHADER dataset.

IV. Dataset Generation

Fig. 3 shows the dataset generation pipeline of SHADER. First, 750 naked human body shapes are generated for each gender using SMPL shape parameters converted from the CAESAR dataset [56], [57]. To facilitate cloth simulation, the body models are placed in the standard T-pose to generate the initial naked body mesh. We then apply wide clothing to the naked body mesh using physics-based cloth simulation software. Several types of commercial software have been evaluated for physics-based 3D simulation, including Maya [58] and 3ds Max [59]. Unfortunately, they focus more on the visual effects of 3D surfaces rather than the complicated movements of clothing; in addition, the 3D modeling of clothes is technically a complex work that requires cross-field backgrounds, thus making it infeasible for our task. We instead chose CLO3D [53], a virtual cloth design and simulation software, to generate 3D garment deformations. CLO3D can simulate realistic garment deformations on virtual avatars with a rich number of adjustable parameters based on real-world physics such as gravity, wind strength, hardness, and coefficient of friction. Once a suit is dressed on an avatar, the garments' movement can be further simulated given a per-frame animation file of the avatar mesh without any human interaction. Moreover, CLO3D provides an online store [60] of commercially available garment sets made by professional designers, which obviates the need for garment making by non-professionals.

Even though CLO3D can simulate sequential movements of garments, the initial fitting still requires manual adjustment. Theoretically, we would need to apply clothing to every different body shape, which could be laborious and time consuming. Therefore, to minimize the amount of human labor necessary, based on the fact that in real life there are only a few sizes available in stores, e.g. S, M, L, etc., the whole-body shape space of each gender is first divided according to height. Each sub-space is then further divided according to shape. Specifically, we first obtain the maximum and minimum z-coordinate values from all reconstructed body meshes, then uniformly divide the whole range into five intervals, resulting in five sub-spaces. In a similar manner, for all body meshes in each sub-space, we compute the maximum and minimum volumes and uniformly divide the whole volume range into five intervals. We thus obtain 25 sub-spaces, and consequently 25 centroid body meshes, which define our templates. In other words, for each garment, we only apply garment fitting to 25 templates. For other non-template human body models, the garment mesh from the corresponding template is directly applied.

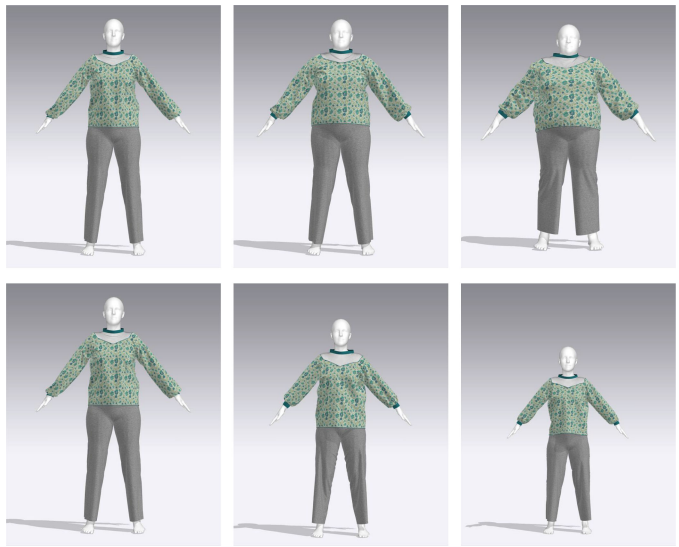


Fig. 4. Template examples. Top: three templates with various body types but similar height; Bottom: three templates with various height but similar body type.

Fig. 4 shows some templates used in this semi-automatic fitting procedure, which cover a wide range of body types from slim to over-weight, as well as various heights from tall to short. In addition, since CLO3D [53] supports physics-based simulation, the simulated garments have varying interactions with different body shapes and cloth materials. To make our dataset cover a large variety of dress codes, we use 30 commercial 3D garment models (15 upper garments and 15 lower garments) from [60] and apply different garment combinations to the naked body meshes. Fig. 5 shows several garment fitting results for the same body mesh. The garment fitting process took two paid trainees about one month to complete.

While the semi-automatic garment fitting is able to successfully apply garment templates to most human avatars in the

Fig. 5. Cloth fitting examples of the same body mesh from the male (top) and the female (bottom) part of the SHADER dataset.

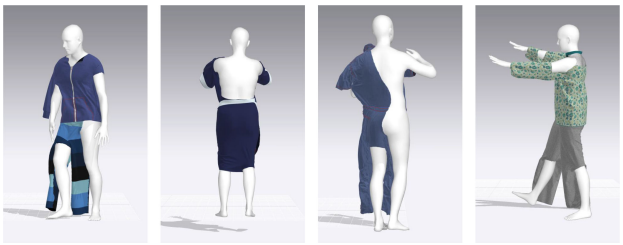


Fig. 6. Some failure cases of the semi-automatic fitting.

dataset, there are still a small number of failure cases where the garments are not properly fitted to the bodies (some examples are shown in Fig. 6). These samples are manually corrected to guarantee their quality. Upon completion of the garment fitting for the standard T-posed body mesh, the garment mesh is transferred to other naked body meshes with arbitrary body poses using MoCap sequences in SURREAL [19]. As CLO3D supports automatic garment simulation given per-frame animation of body poses, we generate an intermediate transition between the standard T-pose and the body pose of a randomly chosen starting frame of a MoCap sequence using linear interpolation. Because SMPL pose parameters have axis-angle representation, pose parameter interpolation is performed after converting axis-angle representation to Euler angle representation [61]. In this process, 200 consecutive frames from the initial frame of MoCap data are used, resulting in 200 differently deformed garment meshes. Finally, Blender [62] is used to project 3D models into 2D images.

In summary, SHADER can provide SMPL shape and pose parameters, synthetic but realistic 2D image renderings, 2D and 3D joint locations, naked and dressed 2D silhouettes, and their SMPL shape parameters. Our released dataset contains 750 human body shapes for each gender, and 200 different body poses for each human shape. Each mesh is colored by a randomly chosen skin texture from SURREAL [19]. When projecting 3D meshes to 2D images, the camera parameters are also randomly applied. Note that the number of samples in the dataset can be further increased by changing the background images, which is performed on-the-fly during the training stage as data augmentation.

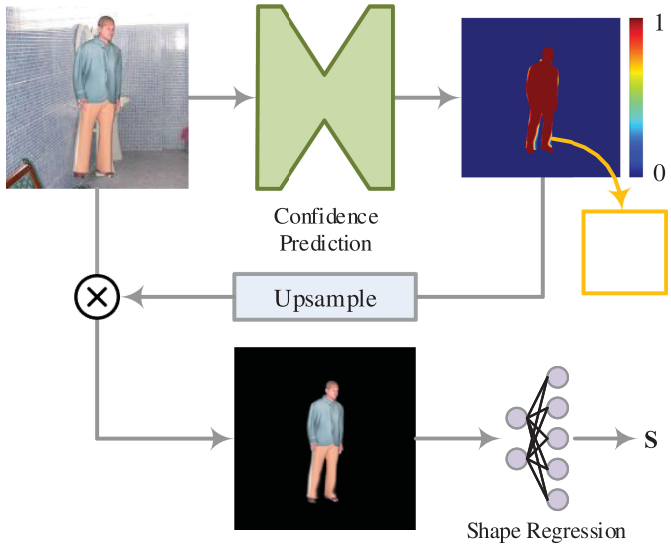


Fig. 7. Network structure overview. An input RGB image is first fed to the confidence prediction network, then the estimated confidence channel is element-wise multiplied with the input image. The weighted input image is used to estimate the body shape parameters, denoted as **S**.

V. Human Shape Estimation

This section presents the network structure used to evaluate the effectiveness of the proposed SHADER dataset. Fig. 7 shows an overview of the network, which mainly consists of two subnets: a confidence prediction network and a shape estimation network.

A. *Confidence Prediction Network*

The pixel values in the silhouette confidence map indicate their reliability to the human shape representation. To make the best use of this silhouette confidence measure, we train a network that can predict the silhouette confidence given a color image. As shown in Fig. 8, the network structure is embodied with the widely used Hourglass module [30]. The input image is first examined by several convolutional and residual layers to extract features with 256 channels. These feature maps are then used to generate silhouette confidence maps by six repeated hourglass modules. Each module is followed by a convolution layer to produce the intermediate silhouette confidence, which is then re-mapped to 256-channel features and fused with the original feature maps before proceeding to the next module. This design encourages each hourglass module to learn only a small number of features and progressively refine the prediction of the previous module. The final estimation is obtained using the last Hourglass module and its output convolutional layers. Batch normalization and ReLU activation are applied to all convolutional layers, and each Hourglass module is trained under the same supervision to retain the gradients during back-propagation.

We found that although the expected output of the network is a single-channel confidence map, including the inverse version as an additional output channel can make it easier and faster for the network to converge. Thus, we define the loss function as the sum of the mean squared error of two silhouette confidence

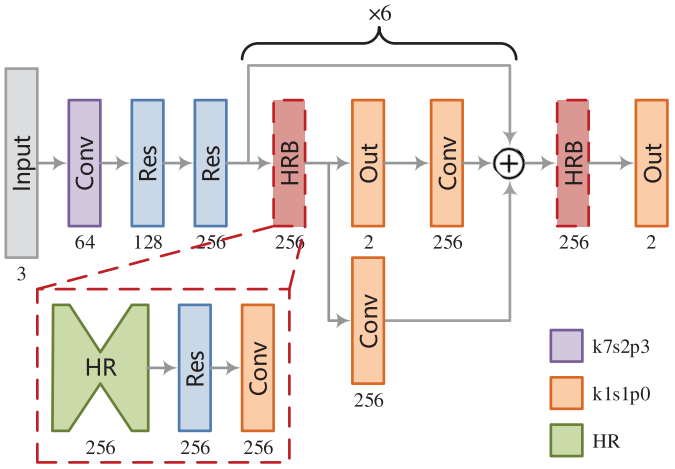


Fig. 8. Structure of the confidence prediction network. Each rectangle represents a processing block, and the number below each block is the size of output channels. The color of each block indicates its convolution settings, i.e., k·s·p·, where "k" represents kernel size, "s" represents kernel stride, and "p" represents padding. Res and HR represent the residual module [63] and the hourglass module [30], respectively.

maps, and only use the first channel in the inference stage:

$$l\left(p,\hat{p}\right)=\sum_{i=1}^{N}\sum_{j=1}^{M}\left\|c_{i,j}p_{i,j}-\hat{p}_{i,j}^{1}\right\|_{2}^{2}+\sum_{i=1}^{N}\sum_{j=1}^{M}\left\|c_{i,j}p_{i,j}-\left(1-\hat{p}_{i,j}^{2}\right)\right\|_{2}^{2},\qquad(3)$$

where $c_{i,j}$ is the silhouette confidence defined in (2), $p_{i,j}$ is the naked silhouette at location (i,j), and $\hat{p}_{i,j}^{1}$ and $\hat{p}_{i,j}^{2}$ are the predicted confidence values from the first and second output channels with a size of $N\times M$, respectively.

B. Shape Estimation Network

Our purpose is to demonstrate how the silhouette confidence helps improve the performance of predicting body shapes under wide clothing. We thus assume that any existing human shape estimation network can be used with the confidence prediction network. Consequently, the proposed method is not dependent on specific human shape estimation network architectures, and we will present the experimental results using three state-of-the-art human shape estimation networks [14], [42], [44] in Section V.

The subsequent human shape estimation network can take advantage of the estimated confidence map in estimating the SMPL parameters. Specifically, to show the effectiveness and generalizability of the proposed method, we use the human shape estimation networks with pre-trained weights. To apply these networks to images with wide clothing, each input channel of the color image is element-wise multiplied with the predicted confidence map; this shrinks the region between the inner body and the clothes. We use this pre-processed color image as the new input of the human shape estimation networks and obtain the SMPL shape and pose parameters.

Fig. 9. Examples of the synthetic test set (top: male, bottom: female). The backgrounds are removed for improved clarity.

Fig. 10. Examples of the real human test set (top: male, bottom: female). Face regions are masked out for privacy. The backgrounds are removed for improved clarity.

VI. Experiments and Analysis

A. Test Sets and Implementation Details

The confidence prediction network was implemented using PyTorch. Among the 200 consecutive body poses for each body shape in SHADER, we collected the last frame to construct the validation set. To evaluate the performance, we generated a synthetic test set containing 500 body models (5 subjects × 100 poses) for each gender. Each subject was dressed with garments that were not used in the training dataset. We also collected several video clips from YouTube that show real subjects with underwear as well as with different clothes of varying fashions. From these videos, we captured screenshots for each subject and used them as the real test set. This real test set includes four persons for each gender, and for each person, 20-60 screenshots with different garments and body poses were obtained. Figs. 9 and 10 show several examples from the two test sets.

Data augmentation was applied during the training of the confidence prediction network to prevent overfitting. We used background images from [64] while excluding images containing humans. When generating training samples, we randomly selected

and cropped a background image and blended it with the foreground human body. To train the confidence prediction network, all input images were resized to 224×224. We cropped the human body according to its bounding box with 10% margins. The network was trained for 5 epochs using RMSprop optimizer [65] with an initial learning rate of 3e-4 and a learning rate decay of 0.9 at every epoch.

B. Parameter Analysis

The silhouette confidence was defined using the Gaussian kernel, where the standard deviation in (2) determines the decay rate of the confidence value from skin to garment. In theory, a smaller value of standard deviation results in sharper transition between skin and garment and thus makes the generated silhouette confidence closer to the naked body shape. It is however more challenging for the network to predict such silhouette confidence. In comparison, a silhouette confidence map generated using a larger standard deviation is easier to predict but less accurate since it assigns higher weights to garment pixels nearby. To leverage this issue and find a trade-off between them, we performed an experiment to evaluate the effect of this Gaussian kernel by training multiple confidence prediction networks with standard deviations ranging from 0.59.5 and with a step size of 0.5. We used human mesh recovery [66] as the reference shape estimation network for this task since it estimates body shape in an iterative regression manner and is thus more robust to misalignment between models of different genders (i.e., male, female, and neutral). The dressed bodies of the synthetic test set without a background were used as the inputs of the confidence prediction network, and the body shapes obtained using different confidence prediction networks were compared with the ground-truth 3D meshes. Specifically, we extracted all 6,890 3D vertices from the body mesh generated by the estimated SMPL parameters, then measured the Euclidean distances of the vertices between the predicted and ground-truth shapes, and took the average value as the measurement result of this sample. We also evaluated the performance on both the naked and dressed test sets without using the confidence prediction network. Because the SMPL body shape space is defined in a normalized cube, we directly used their original output rather than the values re-scaled to the actual sizes.

Table II shows the comparison results of the average vertex distance to the ground-truth for all samples in the synthetic test dataset along with the margin to the best result obtained from the naked images. Compared to the shape prediction results obtained using the dressed set directly, i.e., $\sigma = \infty$, the proposed method significantly reduced the shape prediction error. Based on the results, we chose $\sigma = 1.0$ as the default setting for the rest of the experiments.

TABLE II
COMPARISON OF THE VERTEX DISTANCE OBTAINED FROM NETWORKS TRAINED USING DIFFERENT STANDARD DEVIATION VALUES

σ	Mean Distance	Margin to the best	σ	Mean Distance	Margin to the best
0.5	471.2776	34.8827	5.5	474.8851	38.4902
1.0	459.6334	23.2386	6.0	474.9834	38.5885
1.5	469.2827	32.8878	6.5	478.6912	42.2963
2.0	467.5021	31.1072	7.0	475.2919	38.8970
2.5	471.0144	34.6195	7.5	474.7984	38.4035
3.0	473.1663	36.7714	8.0	474.9197	38.5248
3.5	474.5434	38.1485	8.5	476.4842	40.0893
4.0	473.9833	37.5884	9.0	473.3571	36.9622
4.5	473.2936	36.8987	9.5	474.9585	38.5636
5.0	476.2043	39.8094	∞	478.7165	42.3216

C. Performance Comparisons

To evaluate the effectiveness of SHADER as well as the proposed silhouette confidence, we used three state-of-the-art shape estimation networks [14], [42], [44]. Note that only the confidence prediction network was trained using the SHADER dataset, and the pre-trained shape estimation networks were directly used. We evaluated the performance on both the synthetic test set and the real test set. To begin, we applied only the shape estimation network to naked images (for the synthetic set) and underwear worn images (for the real set). The estimated body shapes using these images reflect the best possible performance of the state-of-the-art networks [14], [42], [44]. Note that the SHADER dataset is gender specific, as male and female datasets have individual SMPL models, whereas the three state-of-the-art methods are neutral to gender. Therefore, rather than using gender-specific 3D meshes in the SHADER dataset, we regard these estimated shapes as the target shapes. Next, we tested the whole pipeline shown in Fig. 7 on the dressed images and obtained the body shapes.

Figs. 11 and 12 depict several examples of color images that were pre-processed using the predicted silhouette confidence on the male and female real test set, respectively. Fig. 13 shows some shape estimation results obtained with and without using the predicted silhouette confidence. For improved clarity, the constant gray-colored images were alpha-blended with the test images, where the confidence values were used as alpha values. It can be seen that the pixels inside the silhouette of the dressed human body but outside the silhouette of the naked human body were multiplied with low confidence values, resulting in the boundaries of the human bodies in the pre-processed color images being more closely fitted to the body parts with higher reliability. This demonstrates that the confidence prediction network effectively predicted the importance of the boundary pixels and was thus able to control the contribution of the boundary pixels in body shape estimation.

We also compared the shape estimated using the dressed images with the target shape. The shape estimation performance was measured as the Euclidean distances of the 3D vertices between the estimated shapes and the target shape. We present the results of this measurement in the form of average values and their standard deviations for all of the image samples in each subject. Table III shows the results obtained using the state-of-the-art shape estimation networks [14], [42], [44] both with and without the use of the proposed confidence prediction network for the synthetic test set. The results show that the

TABLE III
PERFORMANCE COMPARISON OF THE VERTEX DISTANCE FOR THE SYNTHETIC TEST SET

Subject	Vanilla [42]		Weighted [42]*		Vanilla [14]		Weighted [14]*		Vanilla [44]		Weighted [44]*	
	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std
Male 1	7.1025	3.8458	3.2358	2.0883	2.3217	1.1003	1.2276	0.6559	0.1870	0.0289	0.1648	0.0198
Male 2	7.5571	3.8769	4.6218	3.0450	3.1535	1.6635	2.1167	1.4307	2.8777	4.0194	2.7295	4.2583
Male 3	10.8591	3.8074	8.3206	2.9802	3.4463	1.9037	2.3366	1.0961	0.7709	0.3995	0.7055	0.4485
Male 4	7.9193	3.1672	5.0712	3.2655	3.5441	2.3278	2.9451	2.2581	3.1051	3.6983	2.1127	2.5901
Male 5	6.5856	3.8859	4.1444	2.9436	1.6762	1.0058	1.5811	1.0607	1.0953	1.3400	0.8354	1.1297
Average	**8.0047**	**3.7166**	**5.0788**	**2.8645**	**2.8284**	**1.6002**	**2.0414**	**1.3003**	**1.6072**	**1.8972**	**1.3096**	**1.6893**
Female 1	8.9366	3.7161	6.2989	3.6680	4.8258	1.9672	3.3440	1.7733	0.6697	0.4255	0.6322	0.3592
Female 2	20.2045	18.0453	15.9210	12.1655	4.1225	2.3994	4.3461	2.7117	3.1038	1.4740	2.5964	1.1212
Female 3	8.5496	3.4797	7.1536	3.2002	3.4182	2.1873	2.9828	1.9264	2.3399	3.2056	2.2044	3.0512
Female 4	13.5703	7.8873	10.9811	7.5982	6.5546	3.2576	6.4508	3.8410	1.6692	2.2858	1.5553	2.4077
Female 5	11.8438	4.9013	8.7023	3.7865	7.5816	5.1901	7.1582	4.7151	1.1792	0.6309	1.0966	0.6808
Average	**12.6210**	**7.6059**	**9.8114**	**6.0837**	**5.3005**	**3.0003**	**4.8564**	**2.9935**	**1.7924**	**1.6044**	**1.6170**	**1.5240**

*With input images weighted by silhouette confidence from the confidence prediction network.

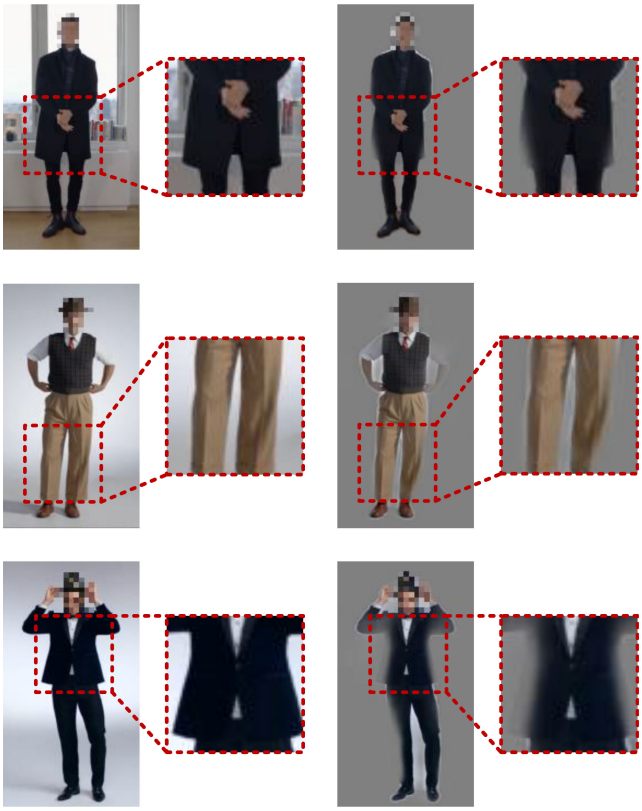


Fig. 11. Examples of the male real test images (left) and the corresponding confidence-weighted images (right).

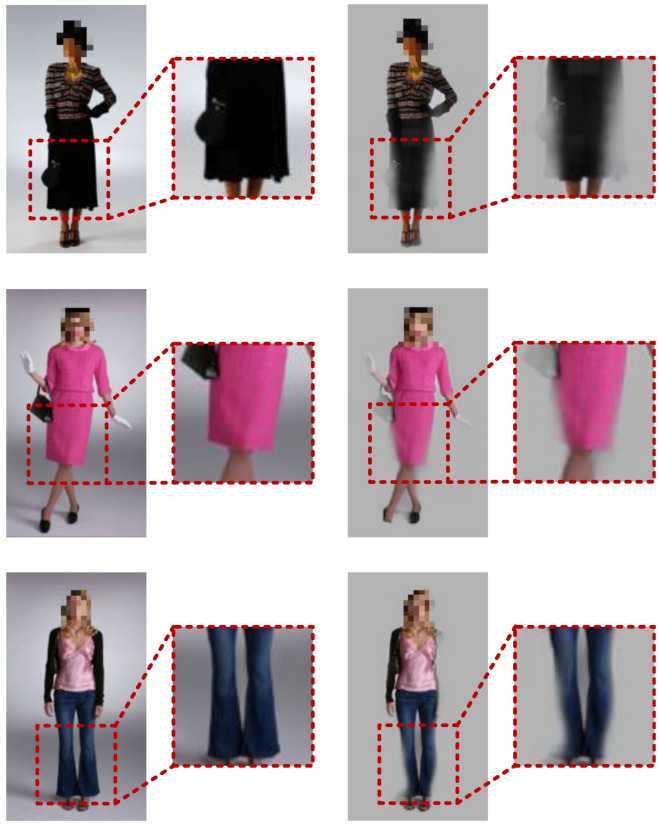


Fig. 12. Examples of the female real test images (left) and the corresponding confidence-weighted images (right).

proposed method decreased the average vertex distance by 36.5% and 22.3% of [42], by 27.8% and 8.4% of [14], and by 18.5% and 9.8% of [44] for the synthetic male and female sets, respectively. Table IV shows the results for the real test set, which indicate the proposed method decreased the average vertex distance by 15.6% and 12.6% of [42], by 7.8% and 12.4% of [14], and by 10.4% and 1.0% of [44] for the real male and female sets, respectively. Furthermore, the proposed method also decreased the average standard deviation of the vertex distances, indicating the consistency and robustness of the proposed method against different poses and types of clothes.

VII. FUTURE WORK

Due to the fact that the garment deformations in the SHADER dataset are simulated using physics-based software, the rendered images are competitive with those taken in the real-world with respect to visual fidelity. Through experiments we showed that

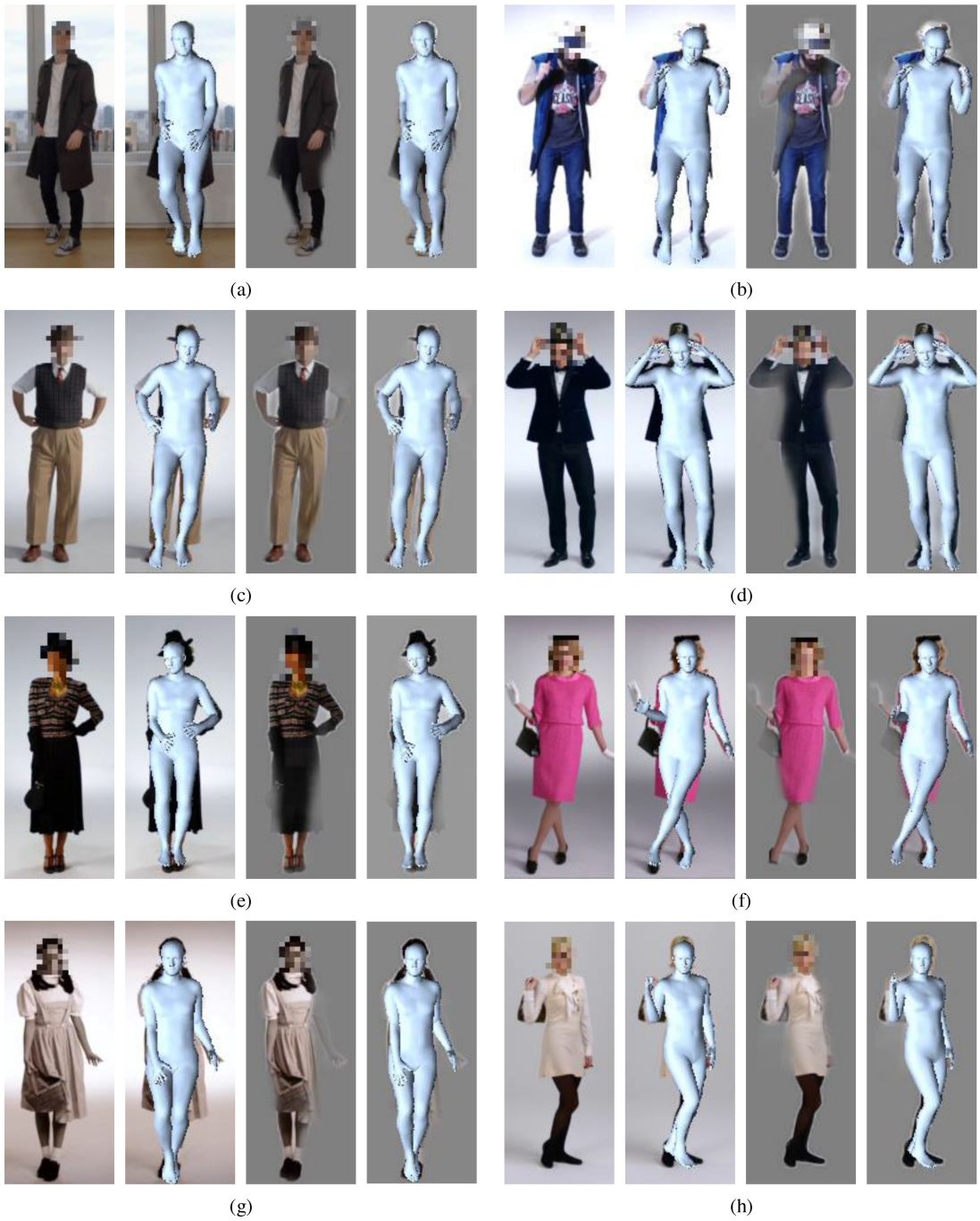


Fig. 13. Experiment results of the proposed method on the real test set. (a)-(d) and (e)-(h) correspond to male 1-4 and female 1-4 in Table IV, respectively. One sample is shown for each subject. From the left to the right, the images in each sub-figure correspond to input image, estimated body mesh overlaid on the input image, confidence-weighted image, estimated body mesh overlaid on the confidence-weighted image, respectively.

although the confidence prediction network is trained with synthetic data only, it remains effective when being adopted to the real-world images. Still, there is a certain gap between two domains. We observed that the generated garment meshes are locally smooth in most surfaces, whereas creases and folds are widely noticed in the real-world fashion style. This misalignment might lead to performance drop in predicting silhouette confidence. A possible solution can be data augmentation via applying real-world garment patterns with pose-invariant creases and folds to the synthetic garment meshes, which is left for future work.

In addition to the aforementioned limitation, the separated network structure used in our experiments could witness a decline in performance since the shape prediction models are pre-trained on other datasets. A dedicated network is preferred in our future work to estimate body shapes directly from the input images, where the silhouette confidence could be learned implicitly with intermediate supervision.

TABLE IV
THE PERFORMANCE COMPARISON OF THE VERTEX DISTANCE FOR THE REAL TEST SET

Subject	Vanilla [42]		Weighted [42]*		Vanilla [14]		Weighted [14]*		Vanilla [44]		Weighted [44]*	
	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std
Male 1	17.5551	7.7140	15.3342	8.2327	5.0352	2.7294	4.6876	2.6689	3.1216	3.6486	2.6271	2.9206
Male 2	15.4212	9.1781	11.9815	8.4181	9.0399	5.4580	8.4787	5.3260	4.3858	5.4222	4.2014	4.9472
Male 3	17.5433	10.3456	16.6805	10.6579	6.5587	2.9693	5.7584	3.2164	1.3032	1.9712	1.0767	0.7118
Male 4	20.9976	10.4047	16.3904	9.0083	8.2003	3.8411	7.6510	3.7941	2.0352	2.3156	1.8145	2.1798
Average	**17.8793**	**9.4106**	**15.0967**	**9.0793**	**7.2085**	**3.7495**	**6.6439**	**3.7514**	**2.7115**	**3.3394**	**2.4299**	**2.6899**
Female 1	16.7365	12.3477	14.4285	11.2156	4.8909	2.2897	4.3702	1.9086	3.3985	4.0964	3.2911	3.9558
Female 2	16.2413	6.4727	15.0179	5.5773	7.3385	4.1480	6.2004	3.2210	1.3099	1.8969	1.5368	2.0418
Female 3	25.3741	9.6916	21.6950	9.4968	6.9594	4.9840	6.5297	4.7396	2.4440	2.7125	2.2692	2.3479
Female 4	16.1930	8.8197	14.0396	9.2567	4.4241	2.1006	3.5833	1.9668	3.3462	3.8219	3.2936	4.0950
Average	**18.6362**	**9.3329**	**16.2953**	**8.8866**	**5.9032**	**3.3806**	**5.1709**	**2.9590**	**2.6247**	**3.1319**	**2.5977**	**3.1101**

*With input images weighted by silhouette confidence from the confidence prediction network.

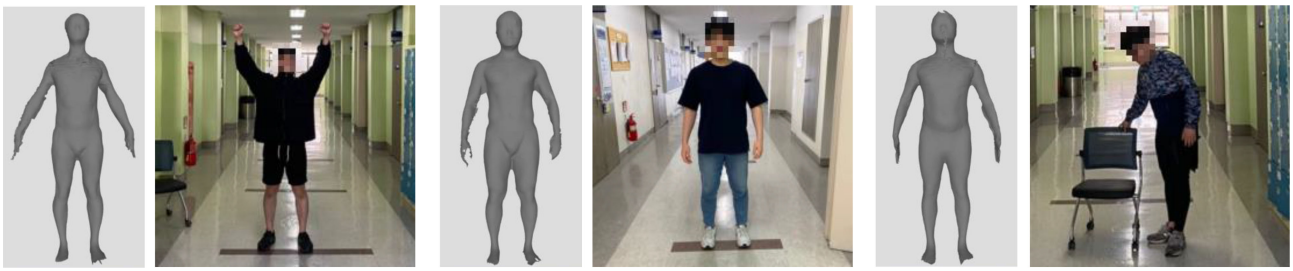


Fig. 14. Some examples of the 3D scan test set. For each subject, left: the ground-truth 3D scan; right: a color image example.

TABLE V
THE PERFORMANCE COMPARISON OF THE HAUSDORFF DISTANCE FOR THE REAL 3D SCAN TEST SET

Subject	Vanilla [42]		Weighted [42]*		Vanilla [14]		Weighted [14]*		Vanilla [44]		Weighted [44]*	
	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std
♯ 1	7.0765	0.0096	7.0717	0.0095	7.1024	0.0078	7.0986	0.0073	7.1037	0.0075	7.0820	0.0064
♯ 2	9.7007	0.0113	9.6951	0.0097	9.7299	0.0138	9.7269	0.0129	9.7269	0.0131	9.6995	0.0043
♯ 3	8.7582	0.0108	8.7524	0.0125	8.7827	0.0092	8.7802	0.0093	8.7870	0.0058	8.7624	0.0031
Average	**8.5118**	**0.0106**	**8.5064**	**0.0106**	**8.5383**	**0.0103**	**8.5352**	**0.0098**	**8.5392**	**0.0088**	**8.5146**	**0.0046**

*With input images weighted by silhouette confidence from the confidence prediction network.

VIII. CONCLUSION

In this paper, we presented SHADER, a newly constructed synthetic dataset of realistic human bodies under wide clothing. We designed a full pipeline around this dataset generation that includes semi-automatic garment fitting for a large shape space from the CAESAR dataset [56], [57], pose variation using the MoCap dataset [19], and the extraction of dressed and naked body silhouettes. The rich diversity of shapes, poses, garments, and background scenes in the SHADER dataset makes it possible to estimate shape parameters from a single color image of a subject under wide clothing. To reveal the body shape underlying wide clothing, we defined the silhouette confidence representation and demonstrated its effectiveness through a confidence prediction network. Experiment results showed that using this network in combination with state-of-the-art shape estimation networks can lead to improved shape estimation performance for bodies wearing wide clothing.

APPENDIX

We also performed an extended experiment to further evaluate the effectiveness of the proposed SHADER dataset as well as the confidence prediction network using real 3D scan data. Three male subjects participated in this experiment. For each subject, we obtained his normalized 3D mesh using a commercially available 3D scanner, and took 10-15 color images with different clothes and poses. Fig. 14 shows the 3D scans as well as some image examples of the subjects. Similar to the experiments in Section VI, we tested the shape estimation network with and without the help of the confidence prediction network on all dressed images, then computed the vertex distances between the estimated body mesh and the ground-truth 3D scan. Note that because the body mesh generated by SMPL parameters and the ground-truth 3D body scan have different numbers of vertices, we used the Hausdorff distance [67] as the metric instead of the vertex-wise Euclidean distance. As shown in

Table V, the proposed confidence network trained on the SHADER dataset helped reduce the body reconstruction error, which further strengthens the aforementioned conclusions.

REFERENCES

[1] E. F. Elsayed *et al.*, "Waist-to-hip ratio and body mass index as risk factors for cardiovascular events in CKD," *Am. J. Kidney Dis.*, vol. 52, no. 1, pp. 49–57, Jul. 2008.

[2] T. A. Welborn, S. S. Dhaliwal, and S. A. Bennett, "Waist–hip ratio is the dominant risk factor predicting cardiovascular death in australia," *Med. J. Aust.*, vol. 179, no. 11, pp. 580–585, Dec. 2003.

[3] M. Yuan, I. R. Khan, F. Farbiz, S. Yao, A. Niswar, and M.-H. Foo, "A mixed reality virtual clothes try-on system," *IEEE Trans. Multimedia*, vol. 15, no. 8, pp. 1958–1968, Dec. 2013.

[4] Y. A. Sekhavat, "Privacy preserving cloth try-on using mobile augmented reality," *IEEE Trans. Multimedia*, vol. 19, no. 5, pp. 1041–1049, May 2017.

[5] L. Sigal, A. Balan, and M. J. Black, "Combined discriminative and generative articulated pose and non-rigid shape estimation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2008, pp. 1337–1344.

[6] P. Guan, A. Weiss, A. O. Balan, and M. J. Black, "Estimating human shape and pose from a single image," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2009, pp. 1381–1388.

[7] H. Fan, H. Su, and L. J. Guibas, "A point set generation network for 3D object reconstruction from a single image," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 605–613.

[8] M. Tatarchenko, A. Dosovitskiy, and T. Brox, "Octree generating networks: Efficient convolutional architectures for high-resolution 3D outputs," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2088–2096.

[9] D. Anguelov, P. Srinivasan, D. Koller, S. Thrun, J. Rodgers, and J. Davis, "SCAPE: Shape completion and animation of people," *ACM Trans. Graph.*, vol. 24, no. 3, pp. 408–416, Jul. 2005.

[10] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "SMPL: A skinned multi-person linear model," *ACM Trans. Graph.*, vol. 34, no. 6, pp. 1–16, Oct. 2015.

[11] G. Pavlakos, L. Zhu, X. Zhou, and K. Daniilidis, "Learning to estimate 3D human pose and shape from a single color image," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 459–468.

[12] M. Omran, C. Lassner, G. Pons-Moll, P. Gehler, and B. Schiele, "Neural body fitting: Unifying deep learning and model based human pose and shape estimation," in *Proc. Int. Conf. 3D Vis.*, 2018, pp. 484–494.

[13] G. Varol *et al.*, "BodyNet: Volumetric inference of 3D human body shapes," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 20–36.

[14] N. Kolotouros, G. Pavlakos, M. J. Black, and K. Daniilidis, "Learning to reconstruct 3D human pose and shape via model-fitting in the loop," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 2252–2261.

[15] Y. Xu, S.-C. Zhu, and T. Tung, "DenseRaC: Joint 3D pose and shape estimation by dense render-and-compare," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 7760–7770.

[16] Y. Rong, Z. Liu, C. Li, K. Cao, and C. C. Loy, "Delving deep into hybrid annotations for 3D human recovery in the wild," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 5340–5348.

[17] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, "Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 7, pp. 1325–1339, Jul. 2014.

[18] C. Lassner, J. Romero, M. Kiefel, F. Bogo, M. J. Black, and P. V. Gehler, "Unite the people: Closing the loop between 3D and 2D human representations," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6050–6059.

[19] G. Varol *et al.*, "Learning from synthetic humans," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 109–117.

[20] P. Wei, H. Sun, and N. Zheng, "Learning composite latent structures for 3D human action representation and recognition," *IEEE Trans. Multimedia*, vol. 21, no. 9, pp. 2195–2208, Sep. 2019.

[21] F. Angelini, Z. Fu, Y. Long, L. Shao, and S. M. Naqvi, "2D pose-based real-time human action recognition with occlusion-handling," *IEEE Trans. Multimedia*, vol. 22, no. 6, pp. 1433–1446, Jun. 2020.

[22] V. Ferrari, M. Marin-Jimenez, and A. Zisserman, "Progressive search space reduction for human pose estimation," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2008, pp. 1–8.

[23] L. Bourdev and J. Malik, "Poselets: Body part detectors trained using 3D human pose annotations," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2009, pp. 1365–1372.

[24] S. Johnson and M. Everingham, "Learning effective human pose estimation from inaccurate annotation," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2011, pp. 1465–1472.

[25] L. Ladicky, P. H. Torr, and A. Zisserman, "Human pose estimation using a joint pixel-wise and part-wise formulation," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 3578–3585.

[26] B. Sapp and B. Taskar, "MODEC: Multimodal decomposable models for human pose estimation," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 3674–3681.

[27] A. Toshev and C. Szegedy, "DeepPose: Human pose estimation via deep neural networks," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 1653–1660.

[28] J. J. Tompson, A. Jain, Y. LeCun, and C. Bregler, "Joint training of a convolutional network and a graphical model for human pose estimation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 1799–1807.

[29] X. Fan, K. Zheng, Y. Lin, and S. Wang, "Combining local appearance and holistic view: Dual-source deep neural networks for human pose estimation," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1347–1355.

[30] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 483–499.

[31] R. Dabral, A. Mundhada, U. Kusupati, S. Afaque, A. Sharma, and A. Jain, "Learning 3D human pose from structure and motion," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 668–683.

[32] X. Zhou, Q. Huang, X. Sun, X. Xue, and Y. Wei, "Towards 3D human pose estimation in the wild: A weakly-supervised approach," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 398–407.

[33] D. C. Luvizon, D. Picard, and H. Tabia, "2D/3D pose estimation and action recognition using multitask deep learning," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 5137–5146.

[34] W. Yang, W. Ouyang, X. Wang, J. Ren, H. Li, and X. Wang, "3d human pose estimation in the wild by adversarial learning," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 5255–5264.

[35] T. Yu *et al.*, "DoubleFusion: Real-time capture of human performances with inner body shapes from a single depth sensor," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7287–7296.

[36] G. Mishra, S. Saini, K. Varanasi, and P. Narayanan, "Human shape capture and tracking at home," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2018, pp. 390–399.

[37] J. Yang, J.-S. Franco, F. Hétroy-Wheeler, and S. Wuhrer, "Estimation of human body shape in motion with wide clothing," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 439–454.

[38] D. Song *et al.*, "3D body shapes estimation from dressed-human silhouettes," *Comput. Graph. Forum*, vol. 35, no. 7, pp. 147–156, Oct. 2016.

[39] Z. Ji, X. Qi, Y. Wang, G. Xu, P. Du, and Q. Wu, "Shape-from-mask: A deep learning based human body shape reconstruction from binary mask images," 2018, *arXiv:1806.08485*.

[40] I. B. Vince Tan and R. Cipolla, "Indirect deep structured learning for 3D human body shape and pose prediction," in *Proc. Brit. Mach. Vis. Conf.*, 2017, pp. 15.1–15.11.

[41] R. Alp Güler, N. Neverova, and I. Kokkinos, "DensePose: Dense human pose estimation in the wild," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7297–7306.

[42] N. Kolotouros, G. Pavlakos, and K. Daniilidis, "Convolutional mesh regression for single-image human shape reconstruction," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 4501–4510.

[43] F. Bogo*et al.*, "Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 561–578.

[44] G. Pavlakos, N. Kolotouros, and K. Daniilidis, "TexturePose: Supervising human mesh estimation with texture consistency," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 803–812.

[45] Y.-W. Chao, J. Yang, B. Price, S. Cohen, and J. Deng, "Forecasting human dynamics from static images," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 548–556.

[46] C. Li, Z. Zhang, W. Sun Lee, and G. Hee Lee, "Convolutional sequence to sequence model for human dynamics," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 5226–5234.

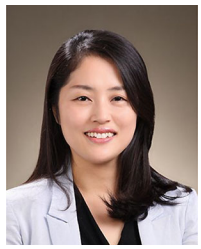
[47] A. Kanazawa, J. Y. Zhang, P. Felsen, and J. Malik, "Learning 3D human dynamics from video," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 5614–5623.

[48] Y. Wang, J. Zhang, H. Zhu, M. Long, J. Wang, and P. S. Yu, "Memory in memory: A predictive neural network for learning higher-order non-stationarity from spatiotemporal dynamics," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 9154–9162.

[49] G. Rogez, P. Weinzaepfel, and C. Schmid, "LCR-Net++: Multi-person 2D and 3D pose detection in natural images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 5, pp. 1146.1161, May 2019.
[50] L. Pishchulin, A. Jain, C. Wojek, M. Andriluka, T. Thormählen, and B. Schiele, "Learning people detection models from few training samples," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2011, pp. 1473–1480.
[51] W. Chen *et al.*, "Synthesizing training images for boosting human 3D pose estimation," in *Proc. Int. Conf. 3D Vis.*, 2016, pp. 479–488.
[52] E. Gundogdu *et al.*, "GarNet: A two-stream network for fast and accurate 3D cloth draping," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 8739–8748.
[53] B. L. Bhatnagar, G. Tiwari, C. Theobalt, and G. Pons-Moll, "Multi-garment net: Learning to dress 3D people from images," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 5420–5430.
[54] CLO Virtual Fashion LLC. (2019) CLO3D software. [Online]. Available: https://www.clo3d.com/, Accessed on: Oct. 14, 2020.
[55] A. Pumarola, J. Sanchez-Riera, G. Choi, A. Sanfeliu, and F. Moreno-Noguer, "3DPeople: Modeling the geometry of dressed humans," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 2242–2251.
[56] R. Kimmel and A. M. Bruckstein, "Subpixel distance maps and weighted distance transforms," in *Proc. Geometric Methods Comput. Vis. II*, vol. 2031, 1993, pp. 259–268.
[57] S. Blackwell *et al.*, "Civilian American and European surface anthropometry resource (CAESAR), Volume 2: Descriptions," AFRL-HE-WP-TR-2002-0169, AirForce Res. Lab., Wright-Patterson Air Force Base, Dayton, OH 45432, Tech. Rep., 2002.
[58] K. M. Robinette, S. Blackwell, H. Daanen, M. Boehmer, and S. Fleming, "Civilian American and European surface anthropometry resource (CAESAR), Volume 1: Summary," AFRL-HE-WP-TR-2002-0169, AirForce Res. Lab., Wright-Patterson Air Force Base, Dayton, OH 45432, Tech. Rep., 2002.
[59] Autodesk. (2019) Maya software. [Online]. Available: https://www.autodesk.com/products/maya, Accessed on: Oct. 14, 2020.
[60] Autodesk. (2019) 3ds max software. [Online]. Available: https://www.autodesk.com/products/3ds-max, Accessed on: Oct. 14, 2020.
[61] CLO Virtual Fashion LLC. (2019) MD store (online 3D garment mesh store). [Online]. Available: https://www.marvelousdesigner.com/store/, Accessed on: Oct. 14, 2020.
[62] G. G. Slabaugh, "Computing Euler angles from a rotation matrix," *Retrieved August*, vol. 6, no. 2000, pp. 39–63, 1999, Accessed on: Oct. 14, 2020.
[63] Blender Foundation. (2019) Blender software. [Online]. Available: https://www.blender.org/, Accessed on: Oct. 14, 2020.
[64] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
[65] A. Quattoni and A. Torralba, "Recognizing indoor scenes," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 413–420.
[66] G. Hinton, N. Srivastava, and K. Swersky, "Lecture 6a overview of mini-batch gradient descent," *Coursera Lecture Slides*, 2012. [Online]. Available: https://class. coursera. org/neuralnets-2012-001/lecture
[67] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik, "End-to-end recovery of human shape and pose," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7122–7131.
[68] W. Rucklidge, *The Hausdorff Distance*. Springer, Berlin, Heidelberg, 1996.

Yucheng Lu received the B.S. degree in optical information science and technology from Hangzhou Dianzi University, Hangzhou, China, in 2016. He is currently working toward the joint M.S. and Ph.D. degree in multimedia engineering with the Department of Multimedia Engineering, Dongguk University, Seoul, South Korea. His main research interests include image segmentation, 3D model reconstruction, and machine learning based computer vision applications.

Jin-Hyuck Cha received the B.S. and M.S. degrees in multimedia engineering from Dongguk University, Seoul, South Korea, in 2017 and 2019, respectively. He is currently working as a Research Engineer with Beyless, Seongnam-si, South Korea. His current research interests include deep learning and computer vision applications.



Se-Kyoung Youm received the Ph.D. degree from Dongguk University in 2007. From 2008 to 2010, she was a Postdoctoral Fellow with the University of Illinois, Urbana-Champaign, USA. Since 2010, she has been with Dongguk University. Her current research interests include healthcare engineering, gerontechnology, u-health service, and interdisciplinary research.

Seung-Won Jung (Senior Member, IEEE) received the B.S. and Ph.D. degrees in electrical engineering from Korea University, Seoul, South Korea, in 2005 and 2011, respectively. From 2011 to 2012, he was a Research Professor with the Research Institute of Information and Communication Technology, Korea University. He was a Research Scientist with the Samsung Advanced Institute of Technology, Yongin-si, South Korea, from 2012 to 2014. From 2014 to 2020, he was an Assistant Professor with the Department of Multimedia Engineering, Dongguk University, Seoul, South Korea. In 2020, he joined the Department of Electrical Engineering, Korea University, where he is currently an Associate Professor. He has authored or coauthored more than 60 peer-reviewed articles in international journals. His current research interests include image processing and computer vision. He received the Hae-Dong Young Scholar Award from the Institute of Electronics and Information Engineers in 2019.