applied sciences

MDPI

Article

DeepProfile: Accurate Under-the-Clothes Body Profile Estimation

Shufang Lu [1,*], Funan Lu [1], Xufeng Shou [1] and Shuaiyin Zhu [2]

[1] College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou 310023, China; 2111912063@zjut.edu.cn (F.L.); 2294462474nona@gmail.com (X.S.)
[2] Shenzhen TOZI Technology Co., Ltd., Shenzhen 518052, China; syz@tozmart.com
* Correspondence: sflu@zjut.edu.cn

Abstract: Accurate human body profiles have many potential applications. Image-based human body profile estimation can be regarded as a fine-grained semantic segmentation problem, which is typically used to locate objects and boundaries in images. However, existing image segmentation methods, such as human parsing, require significant amounts of annotation and their datasets consider clothes as part of the human body profile. Therefore, the results they generate are not accurate when the human subject is dressed in loose-fitting clothing. In this paper, we created and labeled an under-the-clothes human body contour keypoint dataset; we utilized a convolutional neural network (CNN) to extract the contour keypoints, then combined them with a body profile database to generate under-the-clothes profiles. In order to improve the precision of keypoint detection, we propose a short-skip multi-scale dense (SMSD) block in the CNN to keep the details of the image and increase the information flow among different layers. Extensive experiments were conducted to show the effectiveness of our method. We demonstrate that our method achieved better results—especially when the person was dressed in loose-fitting clothes—than and competitive quantitative performance compared to state-of-the-art methods, while requiring less annotation effort. We also extended our method to the applications of 3D human model reconstruction and body size measurement.

Keywords: contour detection; convolutional neural network; image segmentation; 3D human model reconstruction

Citation: Lu, S.; Lu, F.; Shou, X.; Zhu, S. DeepProfile: Accurate Under-the-Clothes Body Profile Estimation. *Appl. Sci.* **2022**, *12*, 2220. https://doi.org/10.3390/app12042220

Academic Editor: Antonio Fernández-Caballero

Received: 28 December 2021
Accepted: 15 February 2022
Published: 21 February 2022

1. Introduction

Human body profiling can be widely used in many fields, such as ergonomics, clothing technology and computer graphics. The estimation of image-based human body profiles can be regarded as a fine-grained semantic segmentation problem. However, current image segmentation methods [1–3] have several drawbacks when applied to body profile estimation. On the one hand, the existing segmentation datasets usually label the contours of each person with clothes and the annotation result is shown in Figure 1b. Therefore, they cannot obtain an accurate result, as the precise body profile is invisible, being covered by clothes. On the other hand, although the human profile can be segmented by a closed boundary that is approximated by polygons, human labelers have to accurately click on numerous boundary points to obtain an accurate human profile, especially for invisible parts covered by clothes.

To overcome these drawbacks, we propose DeepProfile, a novel method to estimate accurate under-the-clothes body profiles. Our DeepProfile includes two stages, body contour keypoint extraction and profile generation. For the first stage, based on our observations in the real world, the tailor just measures several body parts to make well-fitting garments. We extract contour keypoints that are used for later body profile generation. Since there are no contour keypoint training and test datasets, we established a new dataset labeled with front-view and side-view contour keypoints. More details about this dataset are described in Section 3. Based on this dataset, we present an architecture named short-skip multi-scale

dense (SMSD) block to extract contour keypoints. Inspired by multi-scale dense networks (MSDNets) [4], in order to keep the details of the image and increase the information flow, we maintain high-resolution representations through the whole process and connect different layers with short skips in our SMSD block. Additionally, these connections strengthen the back propagation between the loss function and original input features, leading to an implicit deep supervision which makes our network easy to train. Due to the effectiveness of the multi-stage setting in Hourglass and MSPN, we also use the coarse-to-fine strategy by stacking two SMSD blocks with a different setting to obtain better performance. For the second stage, according to the contour keypoints detected in the first stage, we find several more similar profiles in an under-the-clothes body profile database [5] , which was extracted from a large database of 3D human scans. Finally, we generate the final under-the-clothes profile by interpolating these similar profiles.

(**a**) (**b**)

Figure 1. An example of profile in existing segmentation dataset. (**a**) Input image and (**b**) corresponding profile.

In summary, our contributions are three-fold, as follows:

- We constructed an under-the-clothes contour keypoint dataset including a total of 9016 different persons, each person having one front-view image and one side-view image. There were, in total, 45 contour keypoints, 28 for the front view and 17 for the side view. We trained a separate model for each view.
- We put forward DeepProfile, an effective method to estimate accurate under-the-clothes human body profiles, which is generated by contour keypoints extracted from images. Compared with image segmentation methods, it reduced data labeling time and cost by requiring only several contour keypoints to be labeled.
- We applied our body profiles in two scenarios, including accurate 3D human model construction and body size measurement. The 3D human model was generated from front-view and side-view under-the-clothes profiles and the body size measurement results satisfy the criteria of the clothing industry.

2. Related Works

In this section, we focus on techniques targeting image-based human segmentation, 3D body shape estimation and keypoint detection, the topics that are most relevant to our work.

2.1. Image-Based Human Segmentation

General image segmentation is typically used to locate objects and boundaries in images. Long et al. [6] introduced one of the first deep learning works for semantic image

segmentation using a fully convolutional network. Later, deep-learning based semantic segmentation models, such as encoder–decoder architectures [7], regional convolutional networks (R-CNN) [8], Dilated Convolutional Models (ASPP) [9], Gated shape CNNs [10], Attention-Based Models [11], few-shot models [12] and Pyramid Networks [13] have been proposed. General image segmentation methods can be easily applied for extracting human body contours with the corresponding datasets.

One of the popular topics in image segmentation is human segmentation, which can be classified into two categories, human parsing and pose-based human segmentation. On the one hand, human parsing is to segment a human image into different fine grained semantic parts, such as head, torso, arms and legs. Yamaguchi et al. [14] first proposed solving the problem of human parsing as a labeling problem, where images are segmented into superpixels. Dong et al. [15] introduced a method to seamlessly integrate human parsing and pose estimation within a unified framework by utilizing Parselets and Mixture of joint-group templates. Liu et al. [16] proposed a quasi-parametric human parsing model with a specially designed matching convolutional neural network (M-CNN). Liang et al. [2] labeled the human parsing dataset LIP and proposed the JPPNet, which uses multi-scale features and iterative refinement for human parsing. To increase the accuracy of boundary area segmentation results, Ruan et al. [1] added an edge-perceiving module to integrate the characteristic of object contour to refine the boundaries of parsing. On the other hand, pose-based human segmentation approaches generate human instance segmentation with the help of pose estimation. Pose2Instance [17] is a human pose-conditioned segmentation model that adopts a cascade network to improve instance-level person segmentation. However, it relies on human detection and the performance drops when bounding boxes have large overlaps. PersonLab [18] is an approach for the tasks of pose estimation and human instance segmentation using an efficient single-shot model. It treats human body segmentation as a pixel-wise clustering and employs human pose to refine the clustering results. Pose2Seg [19] concatenates the human pose skeleton feature to the image feature in the network to improve human instance segmentation, especially in the case of occlusion. It detects human body key parts and then builds up a segmentation mask on top of the key parts. However, the human body segmentation results generated by the aforementioned methods include information of the clothes the person wears, which does not make for accurate human profiles.

Our method is to estimate the under-the-clothes body profile even when people are dressed in loose-fitting clothing.

2.2. Three-Dimensional Body Shape Estimation

Recently, parametric body models have been commonly used to reconstruct body shape and pose. Kanazawa et al. [20] proposed the HMR method, which uses a 3D regression module to generate parameters of an SMPL [21] body model from a single image. DensePose [22] also uses the SMPL model for dense pose estimation in the wild. It recovers highly accurate dense poses, but it is not suitable for predicting the body shape. NBF [23] takes advantage of both deep learning-based and traditional model-based methods; additionally, it directly predicts the parameters of the model from an RGB image or a semantic segmentation of the image. Smplify-x [24] improves SMPLify [25] by detecting 2D features corresponding to the face, hands and feet to fit the full SMPL-X model. However, these methods address the human pose and shape jointly.

To estimate an accurate 3D body shape, some under-clothing shape estimation methods have been proposed. "The naked truth" [26] is one of the earliest works on recovering the underlying shape; it fits the SCAPE model to a set of calibrated multi-view images or video sequences. The drawback of this method is that it requires the subjects to be captured in several different poses or in a long dynamic sequence. Wuhrer et al. [27] presented a representation that models human body shape and posture independently. et al. [28] proposed a learning-based framework, Body PointNet, to estimate body shape and pose under clothing from a 3D scan. The estimated body shape is output in a point cloud by

operating on a single scan of a dressed person. However, both of these two methods estimate the shape under clothing directly from 3D scans instead of images, making them not as convenient as image-based methods. Streuber et al. [29] introduced Body Talk, which creates a plausible 3D body from standard linguistic descriptions of a 3D shape. This approach mostly uses semantic values that are difficult to quantify; therefore, it may not achieve metric precision. Shigeki et al. [30] proposed an approach to estimate 3D under-clothing human body shapes from a single RGB image. Their approach optimizes an SMPL [21] model using a cloth–skin displacement model, silhouette shape and joint locations. Zhu et al. [5] proposed a method to predict realistic and precise under-the-clothes human body models based on two orthogonal-view photos. However, it requires tedious manual feature extraction operation to be performed on the images.

2.3. Keypoint Detection Techniques

Human keypoint detection usually refers to human pose estimation and it aims to obtain the spatial coordinates of human body joints within a person's image. Similarly, these human pose keypoint detection techniques can be used in our human contour keypoint estimation.

Mainstream keypoint detection methods calculate the position of each keypoint by estimating its heatmap, then choose the coordinate with the highest heat value as its position. "SimpleBaseline" [31] is a simple and effective baseline method to evaluate new methods for keypoint detection. Tompson et al. [32] used CNNs and graphical models to estimate the keypoint offset location. Newell et al. [33] proposed a Stacked Hourglass Network for human pose estimation which fuses low-level and high-level features and improves accuracy by increasing the number of Hourglass units. SPM [34] is a fast and efficient single-stage method implemented with CNNs. Multi-stage methods can decompose complicated problems and improve the accuracy and robustness. G-RMI [35] is a top-down pose estimation approach; it predicts the location and scale of boxes in the first stage and estimates the keypoints in the second stage. Chen et al. [36] presented a two-stage algorithm CPN to relieve the occlusion and complex background problems. Li et al. [37] improved the multi-stage keypoint detection by single-stage module design, cross-stage feature aggregation and coarse-to-fine supervision. Sun et al. [38] proposed a network to maintain high-resolution representations through the whole process. Recently, Zhang et al. [39] paid attention to the investigation of the representation of the human pose using heatmaps. They proposed a more principled distribution-aware decoding method and improved the standard coordinate encoding process. To regress the keypoint positions accurately, Geng et al. [40] presented disentangled keypoint regression (DEKR) to learn disentangled representations through two simple schemes, adaptive convolutions and a multi-branch structure. The first stage of our profile estimation is contour keypoint estimation and its accuracy directly affects the accuracy of profile estimation. In our method, we create and label an under-the-clothes human contour keypoint dataset and propose a CNN network with short-skip multi-scale dense (SMSD) blocks to predict contour keypoints.

3. Dataset

According to anthropometry, ergonomics and product design, the human body shape can be represented by several body parts (neck, shoulder, bust, waist, crotch, knee, calf and ankle, etc.). To better capture the regional shape of the human body, we expanded these features to 45 contour keypoints, including 28 keypoints in the front-view image and 17 keypoints in the side-view image. Figure 2 shows under-the-clothes human body contour keypoint annotation in front-view images and side-view images.

The under-the-clothes contour keypoint dataset covers 9016 persons and each person has one front-view image and one side-view image. There were no restrictions to clothing types, so the human subjects of the dataset were dressed in arbitrary clothing, including tight-fitting, normal-fitting or even loose-fitting clothes. There were 4615 male subjects

and 4401 female subjects and the gender distribution was even. The height distribution and weight distribution of the dataset are shown in Figure 3. We randomly split the 9016 subjects into 8016 for training, 500 for validation and 500 for testing. The front-view and side-view images of the subjects were taken with the subject assuming a standard standing pose, as shown in Figure 1a . There was a small number of front-view images in which the subject was partly obscured by something. In the side-view images, the faces of all the subjects were facing right.

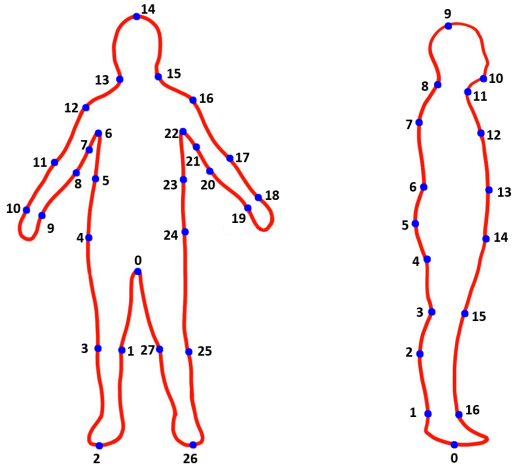


Figure 2. Human body contour keypoints for front-view images (**a**) and side-view images (**b**).

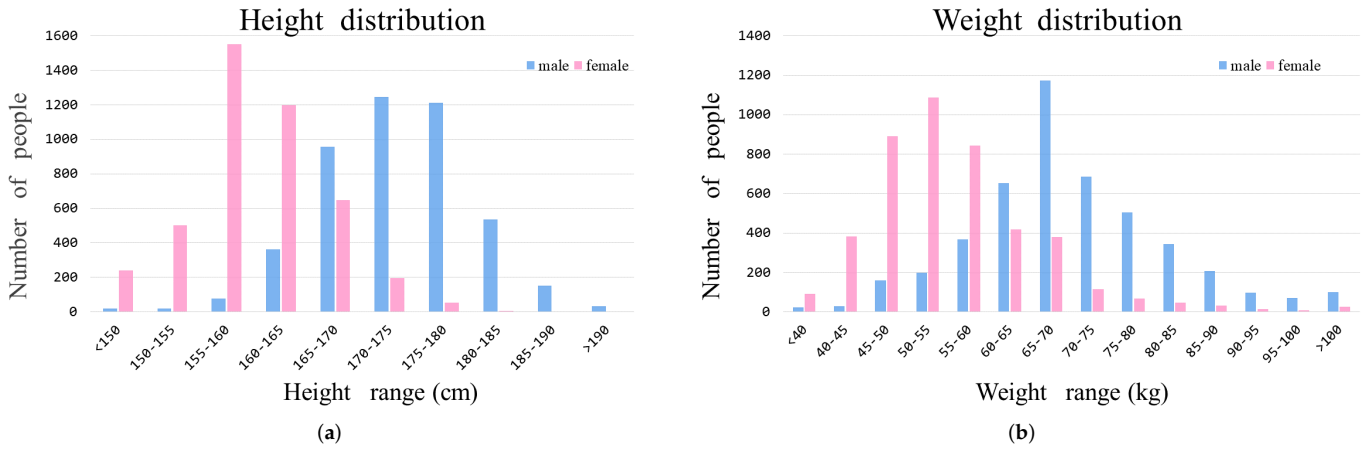


Figure 3. Analysis of the contour keypoint dataset: (**a**) is height distribution and (**b**) is weight distribution.

4. Our Approach

Our DeepProfile method involves two stages. First, we utilize a CNN with SMSD blocks to extract high-precision human body contour keypoints. Then, we combine both contour keypoints and the under-the-clothes profile database [5] to generate accurate under-the-clothes body profiles. The overview of our method is shown in Figure 4.

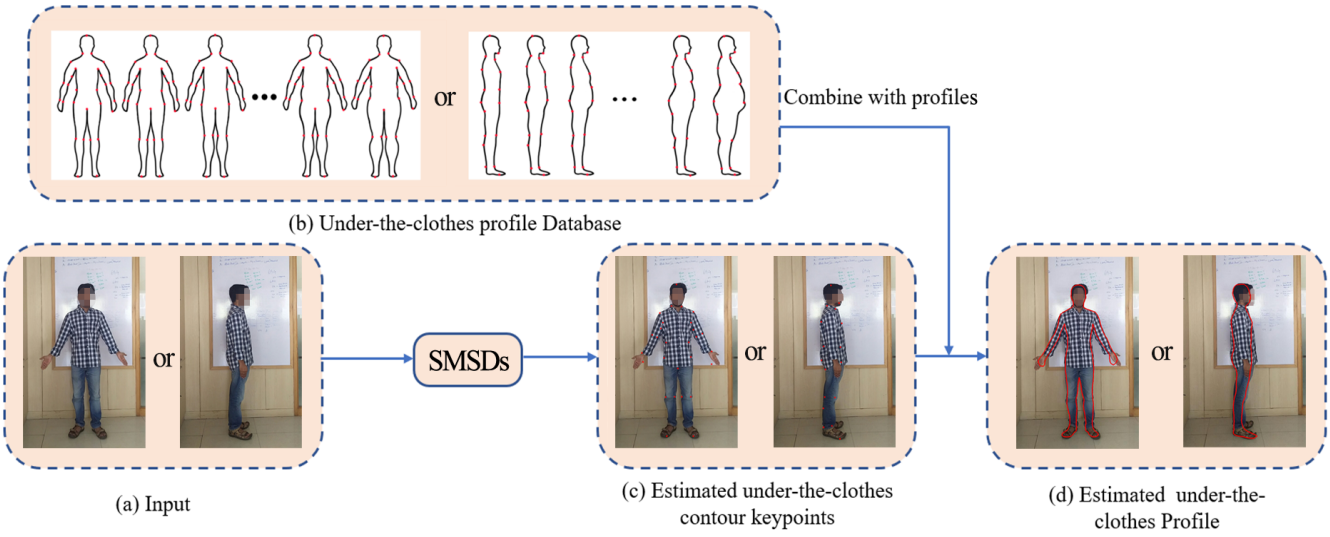


Figure 4. Overview of our DeepProfile. Given the input front- and side-view images (**a**) and a human profile database (**b**), the body contour keypoints are estimated (**c**) to generate the human body profile (**d**). Here, we employ a CNN with SMSD blocks to estimate contour keypoints and compare the difference among them and relevant feature points in profile database, then we interpolate the most similar profiles to synthesize the under-the-clothes profile. Images of different views are trained separately.

4.1. Contour Keypoint Extraction

The contour keypoint extraction problem is defined as giving an RGB image $\boldsymbol{I} \in R^{(W \times H \times 3)}$; we need to estimate the contour keypoints $\boldsymbol{P} \in R^{(N_P \times D)}$ with N_P points, D dimensions and a regression function f_r, represented as follows:

$$\boldsymbol{P} = f_r(\boldsymbol{I}, \boldsymbol{\theta}), \tag{1}$$

where $\boldsymbol{\theta}$ is a set of trainable parameters of the function f_r. Our goal is to optimize the parameters $\boldsymbol{\theta}$ so that we can obtain high-precision contour keypoints. To reliably detect contour keypoints, we use a heatmap $\boldsymbol{H}$ to encode the probabilities of each keypoint. The heatmap $\boldsymbol{H}$ is constructed by modeling the contour keypoints' position as Gaussian peaks. Specifically, for a position (x, y) in the given image $\boldsymbol{I}$, $\boldsymbol{H}(x, y)$ is calculated by

$$\boldsymbol{H}(x, y) = \sum_{i=1}^{N_P} e^{\frac{(x-x_i)^2+(y-y_i)^2}{2\sigma^2}}, \tag{2}$$

where (x_i, y_i) represents the ith contour keypoint and σ is an empirically chosen constant to control the variance of Gaussian distribution.

Network architecture. As shown in Figure 5, we mainly use two short-skip multi-scale dense (SMSD) blocks to predict the contour keypoints from coarse to fine. For the first SMSD block, we utilize ResNet directly to extract features at four different scales from the input image. Then, for features at each scale, we use a 1×1 convolution to decrease the number of channels to 256. Here, we define features from the largest resolution to the smallest resolution as $m1$, $m2$, $m3$ and $m4$. The remaining operation contains the following three steps:

1. In general, a smaller resolution feature has a larger receptive field and a larger resolution feature has a smaller receptive field but keeps more details. In order to take advantage of different resolution features, we upsample $m4$ with nearest neighbor sampling and add it to $m3$ to generate new features, $n3$. We then repeat the process with ($m3$, $m2$) and ($m2$, $m1$) to generate features $n2$ and $n1$, respectively.

2. With the input of $n3$, $n2$ and $n1$, two features of $p2$ and $p1$ are produced by using the same operation in step 1. Unlike FPN [41], which directly upsamples the high-level features and combines them with the low-level features, we add dense connections from $m1$, $m2$, $m3$ and $m4$ to $p1$ and $p2$. These short-skip connections not only strengthen the information flow in our network, but also strengthen the back propagation between the loss function and current features, which makes our network easy to train.
3. Finally, we fuse two features, $p2$ and $p1$, to obtain the heatmaps of the first SMSD block.

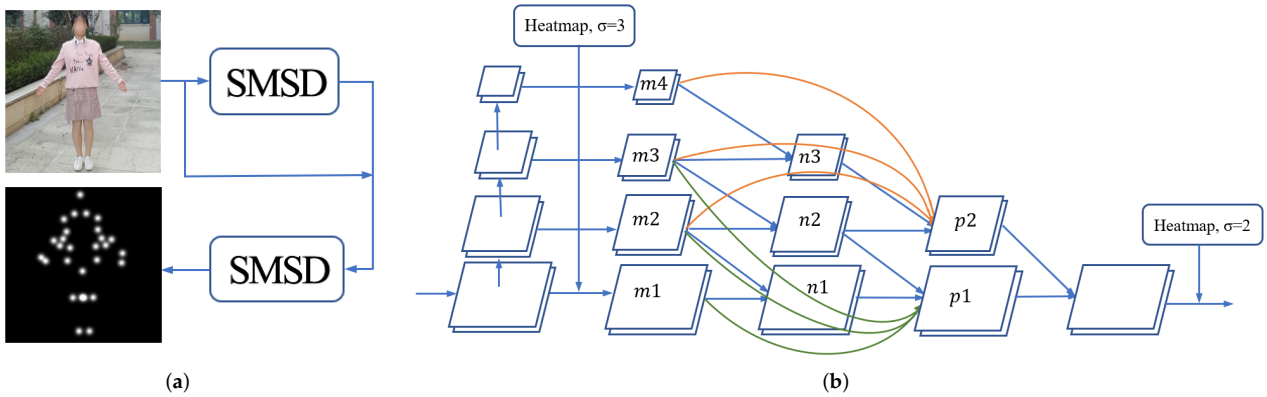


Figure 5. Our human body contour keypoint estimation network (**a**) and structure of SMSD block (**b**).

In order to extract higher-accuracy contour keypoints, we add coarse-to-fine supervision into the SMSD block. In step 1, we connect the output of ResNet to one feature; then, a heatmap ($\sigma = 3$) is employed to compute loss, while, in step 3, σ is set as 2 to obtain more accurate keypoint positions. Figure 6 shows two different heatmaps. Besides, unlike MSDNet [4], we use multi-stage modules to obtain better performance. After the first SMSD block, we fuse its output with $m1$ as the input of the second SMSD block. Especially, for the second SMSD block, we remove the first layer of ResNet.

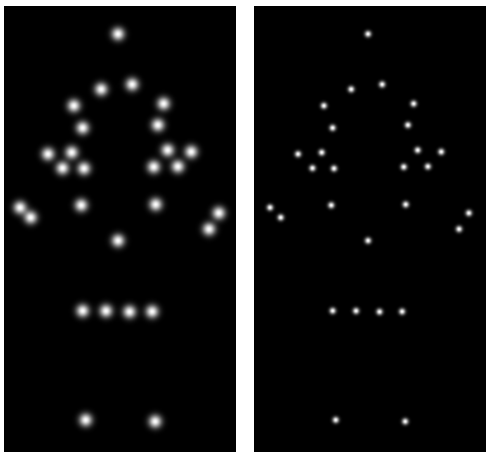


Figure 6. Heatmaps with different values of σ. The value of σ in (**a**) equals 3 and, in (**b**), σ equals 2. These two heatmaps allow our network to estimate the contour keypoints from coarse to fine.

Loss function. There are four losses in our contour keypoint extraction network; all of them come from the first and the third steps of two SMSD blocks. They are computed by the loss function L2, written as follows:

$$L2 = \frac{1}{N_P}\sum_{i=1}^{N_P}\omega|\hat{h}_i - h_i|^2, \tag{3}$$

where $\hat{h}_i$ indicates the ground truth heatmap, h_i is the predicted heatmap and i is for ith keypoint. Since some keypoints may be invisible due to occlusion, we add a weight ω to the loss. We set $\omega = 0$ for invisible keypoints and $\omega = 1$ for those visible ones.

4.2. Profile Estimation from Contour Keypoints

Based on the contour keypoints extracted from the images in Figure 4c, we then combine them with an under-the-clothes body profile database [5] (see Figure 4b) to generate the final human body profile shown in Figure 4d. The profiles in this database are constructed from a large database of human scans and cover a wide range of body shapes. Each profile is composed of about 1000 points with even vertical spacing. As there are only 7–9 boundary keypoints in [5], in order to match our 45 contour keypoints with it, we manually selected the corresponding 28 front-view keypoints and 17 side-view keypoints from the database of [5]. As all the profiles had the same number of points and topology, manual selection was required only once.

The under-the-clothes profile is synthesized by interpolating the most similar profiles m. Keypoint and image features are both available to search similar profiles. In our experiments, we searched with all keypoints and set m as 15.

5. Experimental Results and Discussions

5.1. Data Preprocessing and Evaluation of Contour Keypoints

In the contour keypoint extraction pre-process, data augmentations include random rotation $[-45°, 45°]$ and random scale $[0.7, 1]$, which are applied to ensure the whole body is in the image. The resolution of the input image for the detector was (256×256). For test and evaluation, the scale factor was set as 0.85 and no rotation was used. For a comprehensive evaluation, we used a simplified object keypoint similarity (OKS) that is represented as

$$OKS = \frac{\sum_i exp(-d_i^2/2s^2k_i^2)\delta(v_i > 0)}{\sum_i \delta(v_i > 0)}. \quad (4)$$

where d_i is the Euclidean distance between the detected ith contour keypoint and the corresponding ground truth, s is the object scale and v_i is the visibility of the ith keypoint in the ground truth. k_i is per-keypoint constant that controls falloff and it was set as 0.025 in our method. For the predicted heatmap, we selected the highest heat value location with a quarter offset in the direction of the second highest value. To obtain a detailed result, such as COCO [42], we compared the average precision (AP) at OKS = 0.5, 0.55, . . . , 0.90, 0.95 and mAP (the mean of 10 AP scores).

5.2. Training and Comparison for Contour Keypoints

To compare with state-of-the-art methods, we chose Hourglass [33], CPN [36], HR-Net [38], SimpleBaseline [31], MSPN [37], DARK [39] and DEKR [40], seven existing mainstream methods, to train on front-view and side-view images separately. We used their publicly available codes and trained them on our proposed training set for a fair comparison. For front-view images, we generated a bounding box based on the contour keypoints to remove some background. Especially, for CPN, σ was set as 2,3,4 and 5 to generate the heatmaps, because these values obtained higher accuracy results in our dataset. For our method, we used the Adam optimizer. The learning rate started with 2.5×10^{-4} and decreased to 2.5×10^{-5} at the 180th epoch, then decreased to 2.5×10^{-6} at the 220th epoch and, finally, stopped at the 240th epoch. The batch size for training was 24. For side-view images, we trained these methods the same way as we trained them on front-view images. The result comparison of contour keypoint estimation between our method and other methods is shown in Table 1; our method outperformed the second one by 0.53 in respect to front-view images and 0.94 in respect to side-view images.

Table 1. Quantitative comparisons with other state-of-the-art methods based on our keypoint dataset.

Method	Input Size	Orient	Params	*mAP*	AP^{65}	AP^{75}	AP^{85}
8-stage Hourglass [33]	256 × 256	Front	98.7*M*	78.56	99.0	94.4	67.6
CPN [36]	256 × 256	Front	178.1*M*	77.14	99.4	94.2	63.8
SimpleBaseline [31]	256 × 256	Front	262.6*M*	77.48	99.4	94.8	64.0
HRNet [38]	256 × 256	Front	243.2*M*	77.71	99.0	94.2	62.8
MSPN [37]	256 × 256	Front	462.1*M*	78.64	99.6	94.8	68.2
DARK(HRNet-W32) [39]	256 × 256	Front	108.9*M*	78.89	99.4	93.8	68.8
DEKR(HRNet-W32) [40]	256 × 256	Front	113.3*M*	78.85	99.2	94.8	68.2
Ours	256 × 256	Front	248.6*M*	79.42	99.8	95.4	70.2
8-stage Hourglass [33]	256 × 256	Side	98.6*M*	69.02	95.0	81.0	43.6
CPN [36]	256 × 256	Side	177.6*M*	66.62	93.6	78.2	35.4
SimpleBaseline [31]	256 × 256	Side	262.6*M*	67.76	94.6	78.4	41.6
HRNet [38]	256 × 256	Side	243.2*M*	68.9	95.2	81.4	44.2
MSPN [37]	256 × 256	Side	460.5*M*	66.72	94.0	77.4	36.8
DARK(HRNet-W32) [39]	256 × 256	Side	108.9*M*	69.16	94.6	79.8	42.4
DEKR(HRNet-W32) [40]	256 × 256	Side	112.8*M*	68.22	95.6	79.2	41.2
Ours	256 × 256	Side	248.5*M*	70.1	95.4	84.0	46.0

To further prove the effectiveness of our SMSD block, we designed an ablation experiment to show the results with different numbers of SMSD blocks in the network. As shown in Table 2, the single-SMSD-block model performance was 79.24 with front-view images and 69.16 with side-view images. By contrast, the model with two SMSD blocks led to a 0.18 improvement with front-view images and a 0.94 improvement with side-view images. However, when we increased the number of SMSD blocks to three, it only led to a 0.02 improvement with front-view images and a 0.1 improvement with side-view images, while more parameters were required.

Moreover, we also tried to remove the dense connections from $m1$, $m2$, $m3$ and $m4$ to $p1$ and $p2$ in the SMSD block to test the effectiveness of the dense connections. From Table 2, we can see thaet a single SMSD block with dense connections led to a 0.46 improvement with front-view images and a 0.6 improvement with side-view images. In addition, we compared our SMSD block with Res2Net-50 [43]. Without these dense connections, Res2Net-50 outperformed the one-SMSD-block model by 0.44 with front-view images and 0.46 with side-view images. However, with these dense connections, the one-SMSD-block model outperformed Res2Net-50 by 0.04 with front-view images and 0.12 with side-view images.

Table 2. Component analysis. "Front" and "Side" denote the front- and side-view images. The accuracy was improved with the increase in the number of SMSD blocks. "1 SMSD w/o dense" means that the dense connections from $m1$, $m2$, $m3$ and $m4$ to $p1$ and $p2$ in the SMSD block were removed.

Method	mAP/Front	mAP/Side
Res2Net-50 [43]	79.20	69.02
1 SMSD w/o dense	78.78	68.56
1 SMSD w/ dense	79.24	69.16
2 SMSDs w/ dense	79.42	70.1
3 SMSDs w/ dense	79.44	70.2

5.3. Body Profile Results

For the body profile, we compared the contour line generated by our method with several semantic segmentation methods, including CE2P [1], DeepLabv3+ [7], Gated-

SCNN [10], DenseASPP [9], SPNet [13] and RePRI [12]. In order to obtain a fair result, we labeled our dataset with region segmentation annotation and all these segmentation methods were retrained on this newly created dataset. The comparison results are shown in Figure 7. Most of the methods obtained good visual results if the human subject was dressed in tight-fitting clothing and the background image was simple, such as the results shown in the first and second rows. However, if the human subject was dressed in loose-fitting clothing, the accuracy of the semantic segmentation methods decreased. As shown in the third row, because of the long skirt, the results of CE2P, Deeplabv3+, DenseASPP, Gated-SCNN and SPNet were not satisfactory around the crotch area compared with the ground truth. Furthermore, as shown in the fourth row, due to the loose coat, most of the semantic segmentation methods could not precisely locate some parts (such as the armpit), which affected the precision of the body profile. By contrast, our method was able to obtain more accurate results because the body profile was generated through extracted contour keypoints. Moreover, our method required less data annotation than these semantic segmentation approaches.

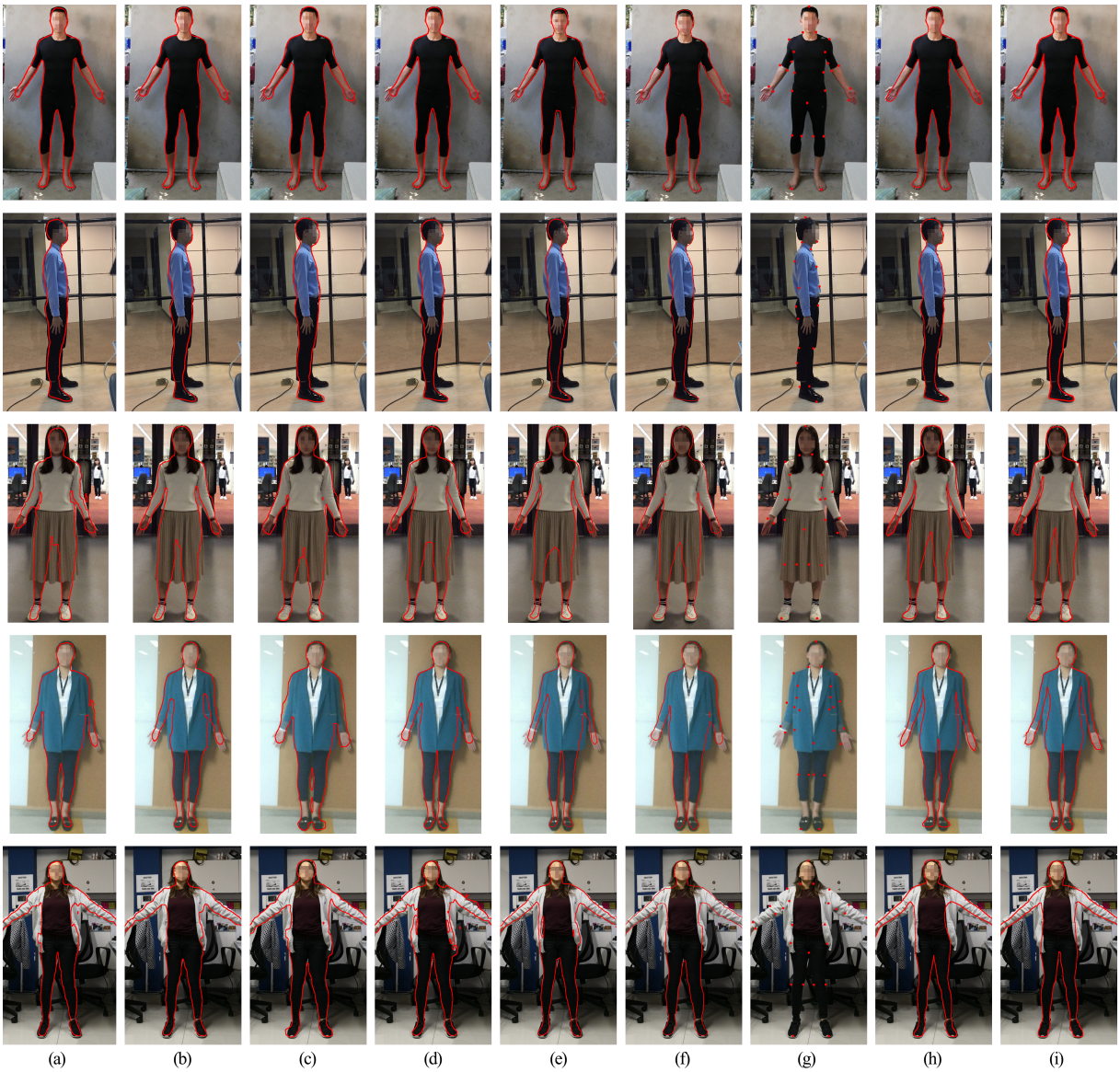


Figure 7. The contour prediction result comparison between other methods and our method, from left to right: (**a**) CE2P, (**b**) Deeplabv3+, (**c**) DenseASPP, (**d**) Gated-SCNN, (**e**) SPNet, (**f**) RePRI, (**g**) our contour keypoint, (**h**) our profile and (**i**) ground truth.

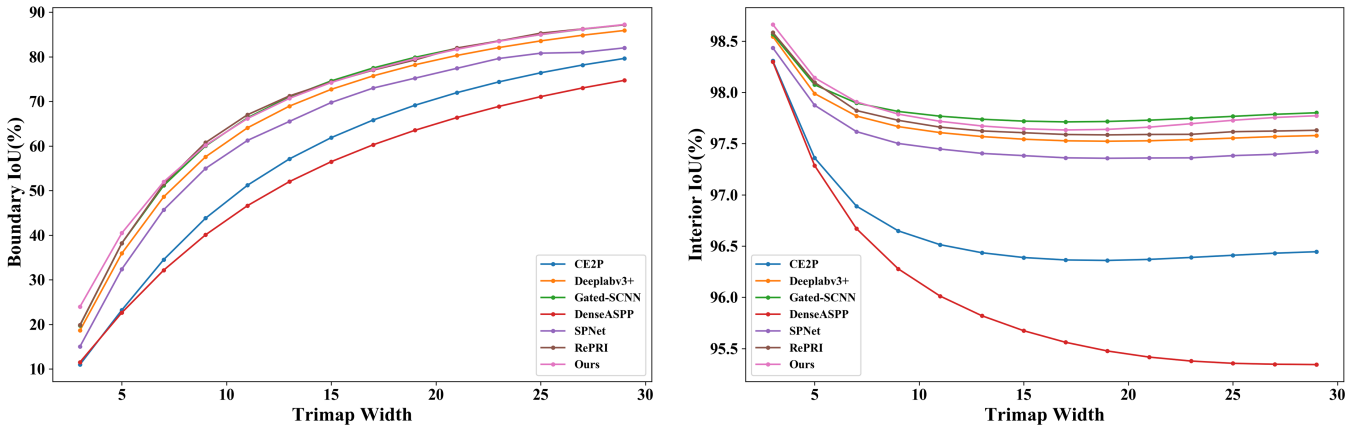
In semantic segmentation, intersection over union (IoU) and pixel accuracy are two frequent metrics for evaluating the effectiveness of a method. However, these two metrics consider all the pixels in both background area and surrounding area. For the contours or profiles, we propose a new metric which focuses on the pixels located on the contours. For each pixel of the predicted profile, we calculate the shortest distance from its coordinates to the coordinates of the pixels on the labeled profile. Then, a mean value is calculated based on the shortest distance among all pixels, which indicates the difference between ground truth and predicted profile. The mean value of distances is represented as

$$dis_m = \frac{1}{n}\sum_{i=1}^{n} \arg\min(||p_i - \hat{p}_j||) \tag{5}$$

where n is the number of pixels of the predicted under-the-clothes profile. There were about 2000 pixels and 1200 pixels in front- and side-view images, respectively. p_i is the ith pixel of the predicted profile and $\hat{p}_j$ is the jth pixel of the ground truth profile. The comparison between our method and other six segmentation methods is shown in Table 3. Our method achieved similar accuracy in terms of mIoU and pixel accuracy to those of state-of-the-art segmentation methods. The mean distance of our method was shorter than the second one by 0.057 pixels. Besides, as shown in Figure 8, we also evaluated segmentation performance in terms of the boundary IoU and interior IoU [44] using trimap widths from 3 pixels to 29 pixels. Our method achieved competitive results compared with Gated-SCNN in boundary IoU and interior IoU, but required quite less data annotation.

Table 3. Result comparison between state-of-the-art segmentation methods and our method. "Pixel acc" is pixel accuracy, "mIoU" represents mean intersection over union, "dis_m" denotes the mean shortest distance between the locations of pixels of the predicted under-the-clothes profile and the locations of pixels of the labeled profile.

Method	Backbone	Pixel Acc	mIoU	dis_m
DenseASPP [9]	DenseNet121	97.72	88.36	6.193
Gated-SCNN [10]	ResNet101	98.87	93.80	2.676
SPNet [13]	ResNet101	98.83	93.65	2.951
CE2P [1]	ResNet50	98.81	93.54	3.02
Deeplabv3+ [7]	ResNet50	98.88	93.92	2.656
RePRI [12]	ResNet50	98.61	93.92	2.419
Our method	ResNet50	98.86	94.01	2.362



(**a**) (**b**)

Figure 8. Error analysis on test set. Boundary IoU (**a**) and Interior IoU (**b**) have different trimap widths.

6. Body Profile Applications

Three-dimensional human model reconstruction. We adopted the 3D human model reconstruction method [5] to create an accurate customized human model from the front-view and side-view profiles extracted in Section 4. We briefly review the process as follows (for more details, refer to [5]): We introduce a 3D shape representation by a 30-layer mesh structure representing shape characteristics from neck to ankle. Each layer represents a cross-sectional shape of the subject's body corresponding the girth of a feature, which includes 2D size features and 3D shape features. Then, the 3D model is reconstructed in the following three steps:

(1) Extract the cross-sectional 2D size features from the subject's profiles estimated in Section 4;
(2) Predict 3D shape features from 2D size features for each layer, which is based on relationship models pre-learned between 2D size features and cross-sectional 3D shape features from a large scale of real human scanned models;
(3) A template model is then deformed with the predicted cross-sectional 3D shapes.

We further fit the SMPL model to our 3D human model to compare with other reconstruction methods based on SMPL [21]. Figure 9 shows two examples generated by body profiles detected by our method and the reconstruction result comparison among different approaches, including HMR [20], NBF [23] and Smplify-x [24]. Compared with other methods, the pose of the 3D model generated by our method was closer to the human pose in the input images, especially around the joints. Besides, the sizes of the limbs generated by other methods were larger than the ground truth, while the corresponding results were improved in our method.

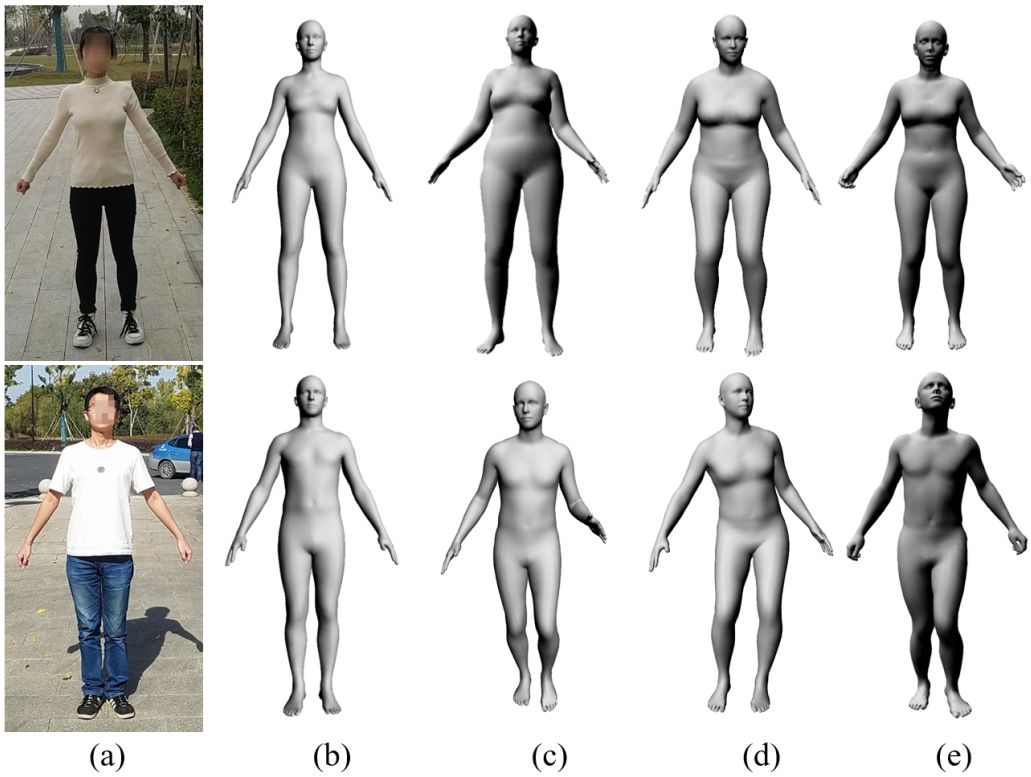


Figure 9. Three-dimensional human model generation and comparison with other methods. From left to right: (**a**) input front-view image, (**b**) 3D model generated by our method, (**c**) HMR [20], (**d**) NBF [23] and (**e**) Smplify-x [24].

Non-contact body size measurement. Human size measurement and fit recommendations play very important roles in the clothing industry and online shopping. We also

applied our estimated under-the-clothes profile method to obtain body size via the generated 3D human model. To test the accuracy of our method, we selected three body parts, including chest, abdomen and calf, which are from upper, middle and lower sections of the body, respectively, for size measurements. The girth measurements of these parts could be easily obtained from the parallel layered 3D shape representation. Table 4 shows the results and the error analysis of size measurements between the estimated size and the ground truth, where the ground truth of each body part was obtained by averaging three manual measurements. The measurement deviation between our estimated size and the manual measured ground truth was about 2 cm, which satisfies the size tolerance of the clothing industry.

Table 4. Results and error analysis of size measurement: chest girth, abdominal girth and calf girth. (Gt) means ground truth; (Err) means the deviation between the estimated size and the ground truth.

ID	Sex	Chest (Gt)/cm	Chest (Err)/cm	Abdominal (Gt)/cm	Abdominal (Err)/cm	Calf (Gt)/cm	Calf (Err)/cm
1	M	83	1.85	75	1.00	38	−1.13
2	M	89	1.52	83	1.27	37	0.16
3	M	87	−1.87	80	1.76	34	−0.22
4	M	101	−0.55	104	−1.71	42	−2.03
5	M	93	1.45	83	1.25	37	−2.03
6	F	82	−2.26	67	0.56	32	1.91
7	F	83	1.03	62	1.82	29	1.14
8	F	80	1.28	66	−0.72	32	−0.38
9	F	80	−0.61	72	0.3	33	−0.13
10	F	86	1.67	82	−1.34	35	1.97
11	M	89	−1.65	84	0.91	38	−1.56

7. Conclusions and Future Work

In this work, we propose a human body profile estimation method which generates accurate under-the-clothes human body profiles via deep learning. We established and labeled under-the-clothes body contour keypoints, based on which the precise human body profile was generated. To improve the precision of keypoint detection, we propose an SMSD block, which keeps the details of the image and increases the information flow among different layers. Extensive experimental results clearly demonstrate significant performance gain from the proposed method over state-of-the-art methods. Moreover, the body profiles obtained with our method were extended to the applications of 3D human model reconstruction and non-contact body size measurement.

Since the images of human subjects in the dataset were captured in a standard standing pose, the body profile extraction results were not satisfactory if the tested pose deviated greatly from the standard pose. In the future, we aim to improve the robustness of the proposed method by expanding our dataset with more, different human poses. Moreover, more potential applications based on our profiling, such as personalized outfit recommendation, could be developed [45,46].

Author Contributions: Conceptualization, S.L., X.S. and S.Z.; methodology, S.L., F.L. and X.S.; data curation, S.Z.; writing—original draft preparation, F.L. and X.S.; writing—review and editing, S.L. and F.L.; supervision, S.L. and S.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by the Natural Science Foundation of Zhejiang Province(LY19F020027) and Zhejiang Provincial Science and Technology Planning Key Project of China under Grant No. 2022C01120.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: To be supplied upon request.

Conflicts of Interest: The authors declare no conflict of interest.

Reference

1. Ruan, T.; Liu, T.; Huang, Z.; Wei, Y.; Wei, S.; Zhao, Y. Devil in the details: Towards accurate single and multiple human parsing. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual, 17 July 2019; Volume 33, pp. 4814–4821.
2. Liang, X.; Gong, K.; Shen, X.; Lin, L. Look into person: Joint body parsing & pose estimation network and a new benchmark. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *4*, 871–885.
3. Sun, K.; Zhao, Y.; Jiang, B.; Cheng, T.; Xiao, B.; Liu, D.; Mu, Y.; Wang, X.; Liu, W.; Wang, J. High-resolution representations for labeling pixels and regions. *arXiv* **2019**, arXiv:1904.04514.
4. Huang, G.; Chen, D.; Li, T.; Wu, F.; van der Maaten, L.; Weinberger, K.Q. Multi-scale dense networks for resource efficient image classification. *arXiv* **2017**, arXiv:1703.09844.
5. Zhu, S.; Mok, P.Y. Predicting realistic and precise human body models under clothing based on orthogonal-view photos. *Procedia Manuf.* **2015**, *3*, 3812–3819. [CrossRef]
6. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
7. Chen, L.-C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 833–851.
8. He, K.; Gkioxari, G.; Dollár , P.; Girshick, R. Mask r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2980–2988.
9. Yang, M.; Yu, K.; Zhang, C.; Li, Z.; Yang, K. Denseaspp for semantic segmentation in street scenes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 3684–3692.
10. Takikawa, T.; Acuna, D.; Jampani, V.; Fidler, S. Gated-scnn: Gated shape cnns for semantic segmentation. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 5228–5237.
11. Huang, Z.; Wang, X.; Wei, Y.; Huang, L.; Shi, H.; Liu, W.; Huang, T.S. CCNet: Criss-cross attention for semantic segmentation. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 603–612.
12. Boudiaf, M.; Kervadec, H.; Masud, Z.I.; Piantanida, P.; Ayed, I.B.; Dolz, J. Few-Shot Segmentation Without Meta-Learning: A Good Transductive Inference Is All You Need? In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 13974–13983.
13. Hou, Q.; Zhang, L.; Cheng, M.-M.; Feng, J. Strip Pooling: Rethinking Spatial Pooling for Scene Parsing. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 4002–4011.
14. Yamaguchi, K.; Kiapour, M.H.; Ortiz, L.E.; Berg, T.L. Parsing clothing in fashion photographs. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 3570–3577.
15. Dong, J.; Chen, Q.; Shen, X.; Yang, J.; Yan, S. Towards Unified Human Parsing and Pose Estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 843–850.
16. Liu, S.; Liang, X.; Liu, L.; Shen, X.; Yang, J.; Xu, C.; Lin, L.; Cao, X.; Yan, S. Matching-CNN meets KNN: Quasi-parametric human parsing. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1419–1427.
17. Tripathi, S.; Collins, M.; Brown, M.; Belongie, S. Pose2instance: Harnessing keypoints for person instance segmentation. *arXiv* **2017**, arXiv:1704.01152.
18. Papandreou, G.; Zhu, T.; Chen, L.C.; Gidaris, S.; Tompson, J.; Murphy, K. PersonLab: Person Pose Estimation and Instance Segmentation with a Bottom-Up, Part-Based, Geometric Embedding Model. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 282–299.
19. Zhang, S.H.; Li, R.; Dong, X.; Rosin, P.; Cai, Z.; Han, X.; Yang, D.; Huang, H.; Hu, S.M. Pose2Seg: Detection Free Human Instance Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 889–898.
20. Kanazawa, A.; Black, M.J.; Jacobs, D.W.; Malik, J. End-to-end recovery of human shape and pose. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7122–7131.
21. Loper, M.; Mahmood, N.; Romero, J.; Pons-Moll, G.; Black, M.J. SMPL: A skinned multi-person linear model. *ACM Trans. Graph.* **2015**, *34*, 1–16. [CrossRef]
22. Güler, R.A.; Neverova, N.; Kokkinos, I. DensePose: Dense Human Pose Estimation in the Wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7297–7306.
23. Omran, M.; Lassner, C.; Pons-Moll, G.; Gehler, P.; Schiele, B. Neural body fitting: Unifying deep learning and model based human pose and shape estimation. In Proceedings of the 2018 International Conference on 3D Vision, Verona, Italy, 5–8 September 2018; pp. 484–494.

24. Pavlakos, G.; Choutas, V.; Ghorbani, N.; Bolkart, T.; Osman, A.A.; Tzionas, D.; Black, M.J. Expressive body capture: 3d hands, face, and body from a single image. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 10967–10977.
25. Bogo, F.; Kanazawa, A.; Lassner, C.; Gehler, P.; Romero, J.; Black, M.J. Keep It SMPL: Automatic Estimation of 3D Human Pose and Shape from a Single Image. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 561–578.
26. Bălan, A.O.; Black, M.J. The Naked Truth: Estimating Body Shape Under Clothing. In Proceedings of the European Conference on Computer Vision, Marseille, France, 12–18 October 2016; pp. 15–29.
27. Wuhrer, S.; Pishchulin, L.; Brunton, A.; Shu, C.; Lang, J. Estimation of human body shape and posture under clothing. *Comput. Vis. Image Underst.* **2021**, *17*, 3793–3802. [CrossRef]
28. Hu, P.; Kaashki, N.N.; Dadarlat, V.; Munteanu, A. Learning to Estimate the Body Shape Under Clothing From a Single 3-D Scan. *IEEE Trans. Ind. Informatics* **2021**, *17*, 3793–3802. [CrossRef]
29. Streuber, S.; Quiros-Ramirez, M.A.; Hill, M.Q.; Hahn, C.A.; Zuffi, S.; O'Toole, A.; Black, M.J. Body talk: Crowdshaping realistic 3D avatars with words. *Acm Trans. Graph.* **2016**, *35*, 1–14. [CrossRef]
30. Shigeki, Y.; Okura, F.; Mitsugami, I.; Yagi, Y. Estimating 3D human shape under clothing from a single RGB image. *Ipsj Trans. Comput. Vis. Appl.* **2018**, *10*, 1–6. [CrossRef]
31. Xiao, B.; Wu, H.; Wei, Y. Simple baselines for human pose estimation and tracking. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 472–487.
32. Tompson, J.; Goroshin, R.; Jain, A.; LeCun, Y.; Bregler, C. Efficient object localization using Convolutional Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 648–656.
33. Newell, A.; Yang, K.; Deng, J. Stacked hourglass networks for human pose estimation. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 483–499.
34. Nie, X.; Feng, J.; Zhang, J.; Yan, S. Single-stage multi-person pose machines. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Korea , 27 October–2 November 2019; pp. 6950–6959.
35. Papandreou, G.; Zhu, T.; Kanazawa, N.; Toshev, A.; Tompson, J.; Bregler, C.; Murphy, K. Towards accurate multi-person pose estimation in the wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 3711–3719.
36. Chen, Y.; Wang, Z.; Peng, Y.; Zhang, Z.; Yu, G.; Sun, J. Cascaded pyramid network for multi-person pose estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7103–7112.
37. Li, W.; Wang, Z.; Yin, B.; Peng, Q.; Du, Y.; Xiao, T.; Yu, G.; Lu, H.; Wei, Y.; Sun, J. Rethinking on multi-stage networks for human pose estimation. *arXiv* **2019**, arXiv:1901.00148.
38. Sun, K.; Xiao, B.; Liu, D.; Wang, J. Deep high-resolution representation learning for human pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 5686–5696.
39. Zhang, F.; Zhu, X.; Dai, H.; Ye, M.; Zhu, C. Distribution-Aware Coordinate Representation for Human Pose Estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 7091–7100.
40. Geng, Z.; Sun, K.; Xiao, B.; Zhang, Z.; Wang, J. Bottom-Up Human Pose Estimation via Disentangled Keypoint Regression. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 14671–14681.
41. Lin, T.-Y.; Dollár , P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 936–944.
42. Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; pp. 740–755.
43. Gao, S.-H.; Cheng, M.-M.; Zhao, K.; Zhang, X.-Y.; Yang, M.-H.; Torr, P. Res2net: A new multi-scale backbone architecture. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *43*, 652–662. [CrossRef] [PubMed]
44. Arnab, A.; Zheng, S.; Jayasumana, S.; Romera-Paredes, B.; Larsson, M.; Kirillov, A.; Savchynskyy, B.; Rother, C.; Kahl, F.; Torr, P.H.S. Conditional random fields meet deep neural networks for semantic segmentation: Combining probabilistic graphical models with deep learning forstructured prediction. *IEEE Signal Process. Mag.* **2018**, *35*, 37–52. [CrossRef]
45. Verma, D.; Gulati, K.; Goel, V.; Shah, R.R. Fashionist: Personalising outfit recommendation for cold-start scenarios. In Proceedings of the 28th ACM International Conference on Multimedia, Seattle, WA, USA, 12–16 October 2020; pp. 4527–4529.
46. Sagar, D.; Garg, J.; Kansal, P.; Bhalla, S.; Shah, R.R.; Yu, Y. Pai-bpr: Personalized outfit recommendation scheme with attribute-wise interpretability. In Proceedings of the 2020 IEEE Sixth International Conference on Multimedia Big Data, New Delhi, India, 24–26 September 2020; pp. 221–230.