Learning to Infer Inner-Body under Clothing from Monocular Video

Xiongzheng Li[†], Jing Huang[†], Jinsong Zhang, Xiaokun Sun, Haibiao Xuan, Yu-Kun Lai, *Member, IEEE*, Yingdi Xie, Jingyu Yang, *Senior Member, IEEE*, and Kun Li[*], *Member, IEEE*

Abstract—Accurately estimating the human inner-body under clothing is very important for body measurement, virtual try-on and VR/AR applications. In this paper, we propose the first method to allow everyone to easily reconstruct their own 3D inner-body under daily clothing from a self-captured video with the mean reconstruction error of 0.73cm within 15s. This avoids privacy concerns arising from nudity or minimal clothing. Specifically, we propose a novel two-stage framework with a Semantic-guided Undressing Network (SUNet) and an Intra-Inter Transformer Network (IITNet). SUNet learns semantically related body features to alleviate the complexity and uncertainty of directly estimating 3D inner-bodies under clothing. IITNet reconstructs the 3D inner-body model by making full use of intra-frame and inter-frame information, which addresses the misalignment of inconsistent poses in different frames. Experimental results on both public datasets and our collected dataset demonstrate the effectiveness of the proposed method. The code and dataset is available for research purposes at http://cic.tju.edu.cn/faculty/likun/projects/Inner-Body.

Index Terms—Inner-body, under clothing, reconstruction, single RGB camera, transformer.

1 Introduction

Estimating a personalized body shape is very important for virtual try-on and body measurement, which enables personalized avatar generation in VR/AR [1], [2], [3], [4]. Although it is straightforward to capture the personalized body shape when the person is naked or nearly naked, most people cannot accept this form of collection. Therefore, our work proposes a new problem: estimating an accurate inner-body under clothing from a self-captured video, to avoid privacy concerns arising from nudity or minimal clothing.

Although one can estimate the inner-body from a clothed 3D model reconstructed by a scanner or a multi-view studio [5], [6], [7] high cost and large set up size prevent the wide-spread applications of such systems. For users, it is more convenient and cheaper to adopt a widely-used RGB camera. Some methods [8], [9], [10], [11], [12], [13], [14], [15], [16], [17] estimated a parametric human model [18], [19] as the inner-body from an RGB image. However, the reconstruction is limited to the parametric space and cannot represent personalized shapes of different people. Therefore, the parametric model is not sufficient for accurate inner-bodies. Some other methods [20], [21], [22] directly estimate 3D vertex positions of human bodies from images. These non-parametric methods can recover more complex shapes, but large distortion or unstable results may occur due to the uncertainty in high dimensions. Moreover, all the methods above directly extract global features or local features from the image of a dressed person to predict the body shape. As a result, the estimated inner-body can be significantly influenced by the clothing. This problem is amplified when the person wears very loose clothes, leading to a fatter result than the actual inner-body.

Therefore, there is an urgent need for estimating an accurate inner-body from a single RGB camera. We propose to estimate the inner-body from a few frames (2-8) of a monocular video captured by the dressed users themselves. There are two main challenges to achieve this: First, it is a highly uncertain and complex problem due to the influence of the clothes. Second, in the self-captured video, the person is moving and hence different frames have different human poses (subtle differences exist even if users try to keep the same pose). How to make better use of pose-inconsistent multi-view information is a key but difficult problem.

In this paper, we propose a two-stage framework to infer an accurate inner-body model of a clothed person from a few frames (2-8) of a monocular video in which the person is moving. To reduce the ambiguities brought in by directly estimating the 3D inner-body from the video, we propose a Semantic-guided Undressing Network (SUNet) to first estimate the 2D inner-body mask, which can learn semantically related body features. In order to deal with changing poses in different frames, instead of averaging the features of multi-frames, an Intra-Inter Transformer Network (IITNet) is proposed to make full use of intra-frame and inter-frame information. The intra-transformer is proposed to pay attention to the effective features within the frame by calculating the correlation in the feature map. The inter-transformer can better integrate the features of different frames and obtain the personalized body shape. To the best of our knowledge, our method is the first work to estimate an accurate inner-body model of a clothed person from a single RGB cam-

- *[†] Equal contribution.*
- *[*] Corresponding author: Kun Li (Email: lik@tju.edu.cn)*
- *Xiongzheng Li, Jing Huang, Jinsong Zhang, Xiaokun Sun, Haibiao Xuan, and Kun Li are with the College of Intelligence and Computing, Tianjin University, Tianjin 300350, China.*
- *Yu-Kun Lai is with the School of Computer Science and Informatics, Cardiff University, Cardiff CF24 4AG, United Kingdom.*
- *Yingdi Xie is with VRC Inc., Tokyo 1920046, Japan.*
- *Jingyu Yang is with the School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China.*

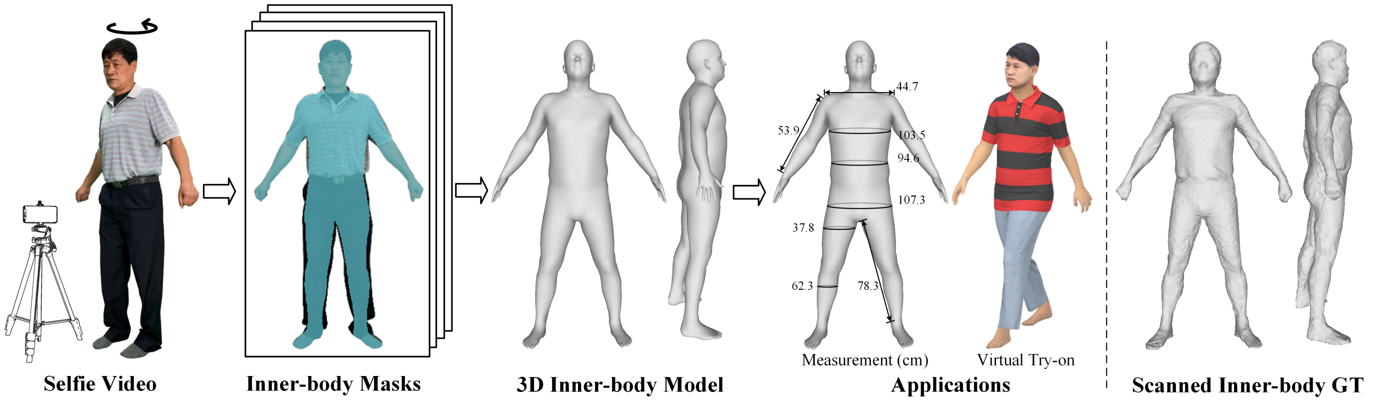


Fig. 1. Given a self-captured video of a clothed person, our method can infer the inner-body masks under clothing and further reconstruct the 3D inner-body model with high accuracy, which enables convenient body measurement and virtual try-on applications.

era. Experimental results on both public dataset and our collected dataset demonstrate the accuracy of the proposed method. Our method makes it easy for everyone to monitor their own body measurement indicators every day, and also to experience virtual try-on easily. An example is given in Fig. 1. *The code and dataset is available for research purposes at http://cic.tju.edu.cn/faculty/likun/projects/Inner-Body*.

Our main contributions can be summarized as:

- We propose a novel two-stage framework to reconstruct the 3D inner-body model of a clothed person from a few frames (2-8) of a monocular video captured by users themselves. This allows for the first time extracting accurate inner-body models of clothed people from a single RGB camera.
- We propose a Semantic-guided Undressing Network (SUNet) to estimate the 2D inner-body mask from the clothed human image, which alleviates the complexity and uncertainty of directly estimating 3D inner-bodies.
- We propose an Intra-Inter Transformer Network (IITNet) to learn the importance of intra-frame and inter-frame features, which addresses the challenge of the changed poses in different frames.
- Experimental results demonstrate that our method can infer the 3D inner-body model to a mean reconstruction error of 0.73cm within 15 seconds. Our method also enables convenient body measurement and virtual try-on applications.

2 Related Work

Although some work uses expensive scanners or multi-camera systems [5], [6], [7] for inner-body reconstruction, such techniques are not widely accessible. In this section, we focus on work most related to ours, where a single RGB camera is used.

Parametric Methods: Parametric methods estimate parametric human models such as SCAPE [19], SMPL [18], and SMPL-X [23] as the human inner-body from a single RGB input, which are efficient and flexible. Given the manually annotated 2D landmarks and smooth shading, Guan *et al.* [24] recovered the human shape and pose by optimizing the parameters of the SCAPE model. Bogo *et al.* [9] proposed an optimization-based method called SMPLify to estimate a SMPL model from a single RGB image. Subsequently, Pavlakos *et al.* [23] proposed SMPLify-X to estimate a SMPL-X model that extended SMPL with fully articulated hands and an expressive face. However, these traditional optimization-based methods are slow and sensitive to the initialization. Kanazawa *et al.* [11] proposed an end-to-end human mesh recovery (HMR) framework to infer the parameters of SMPL directly from image features. Based on HMR, Luan *et al.* [25] suggested to use 3D pose to calibrate the human mesh and developed two new pose calibration frameworks, namely serial PC-HMR and parallel PC-HMR. Kolotouros *et al.* [10] proposed SPIN, which trained a neural network through a tight collaboration of a regression method and an iterative optimization method. Instead of inference from a single RGB image, Kocabas *et al.* [26] proposed a video-based method called VIBE for human pose and shape estimation, which extended SPIN over time by extending SMPLify to video and leveraged the motions of AMASS dataset for adversarial training. Ferrari *et al.* [16] propose a strategy that first trains the network on synthetic datasets and then fine-tunes the network on the real datasets by using the trained network as supervision to estimate the minimal parameterized human body. Yang *et al.* [17] recover the human body shape by relying on the joints and human body silhouette input from the user to obtain a parameterized human body.

Parametric models are not accurate inner-bodies because the pose and shape parameters have limited capacity to represent the body of an arbitrary person. We instead estimate the offsets over the parametric model to better represent personalized body shapes, given the estimated inner-body masks.

Non-Parametric Methods: Non-parametric approaches directly predict the 3D representation from an RGB image or video. Some methods [27], [28], [29] directly regressed voxels to represent human bodies by convolutional neural networks. However, this representation requires intensive memory and has low resolution. To avoid high memory requirements, implicit function representations are proposed for human reconstruction. Saito *et al.* [30] proposed a pixel-aligned implicit function representation called PIFU for high-quality mesh reconstructions with fine geometry details (*e.g.*, clothing wrinkles) from images. However, PIFU [30] and its variants [31], [32], [33] mainly focused on recovering the details of clothes rather than the shape of inner-body. Kolotouros *et al.* [20] considered retaining the

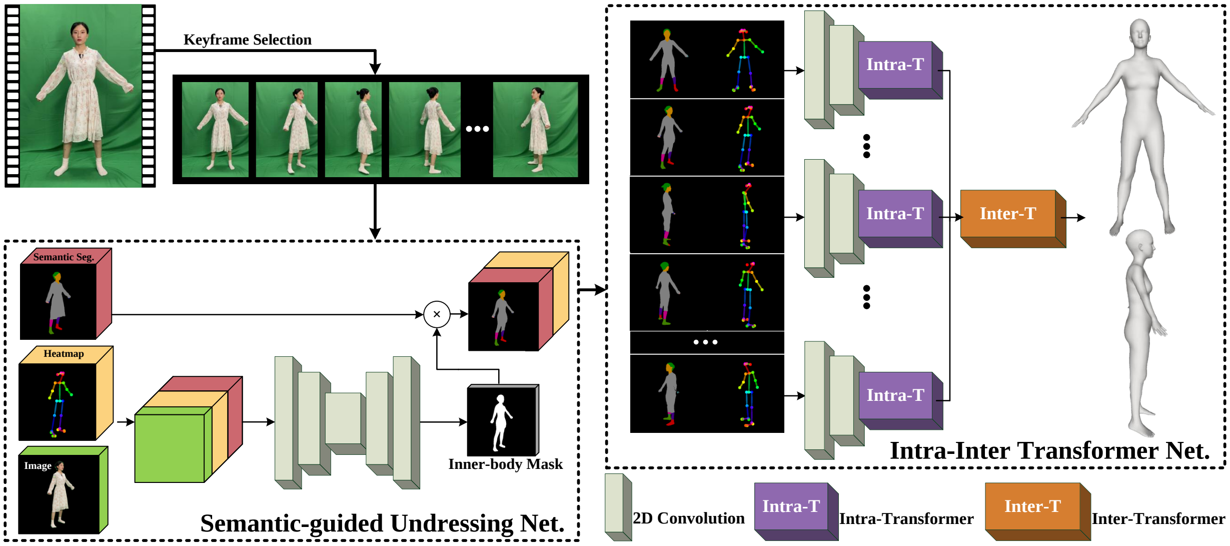


Fig. 2. Overview of our method.

topology of SMPL and directly estimated the positions of the 3D vertices by a graph convolutional network (GCN). Choi *et al.* [34] proposed Pose2Mesh based on graph convolution, which can recover the 3D human mesh from 2D human pose. Similarly, Moon *et al.* [21] proposed an image-to-lixel (line+pixel) network called I2LMeshNet, to estimate the per-lixel likelihood on 1D heatmaps for each vertex, which preserves the spatial relationship in the input image and models the uncertainty of prediction. However, these methods tend to estimate a body shape that is consistent with the clothing, and hence are difficult to obtain the accurate inner-body when wearing loose clothes. Given a monocular video in which a person is rotating in front of a single RGB camera, Thiemo *et al.* [35] proposed a learning-based method to reconstruct a 3D shape including SMPL and clothing. The SMPL model cannot be regarded as the accurate inner-body, and the final reconstructed 3D shape contains the geometry of clothes. Moreover, their method does not deal with the misalignment problem caused by the changed poses and shapes in different frames. All the above methods directly extract global features or local features from the images of dressed people to predict the human shapes, which leads to the influence of clothing on the inner-body estimation.

In this paper, we propose the first method (to our best knowledge) to estimate an accurate personalized inner-body under clothing from a few frames (2-8) of a monocular video captured by the users themselves. To reduce the ambiguities of directly estimating the 3D inner-body from the images of dressed people, we propose a two-stage framework by estimating the inner-body masks in 2D and reconstructing the final inner-body model in 3D. To better deal with the inconsistent poses in different frames, we propose intra-inter transformers to pay more attention to the effective features within the frame and integrate the features between different frames.

3 Method

We aim to solve a new but valuable problem: reconstructing a personalized inner-body under clothing from a self-captured monocular video. The user only needs to turn around with a rough A-pose in front of an RGB camera to take a short video. To alleviate the complexity and uncertainty, we propose a two-stage framework with a Semantic-guided Undressing Network (SUNet) and an Intra-Inter Transformer Network (IITNet). To deal with the pose-inconsistency in different frames, an intuitive idea is to employ SMPL fitting algorithm that estimates a single body shape parameter and different pose parameters for different frames. However, this requires to optimize multi-frame poses at once and store multi-frame models in memory during optimization, which makes it computationally expensive. More importantly, this non-linear optimization is time-consuming which limits its practical applications. Finally, the SMPL fitting algorithm is susceptible to local minima when not initialized properly. Therefore, we propose *intra-transformer* and *inter-transformer* which are low computational cost, fast and stable. As shown in Fig. 2, we first automatically extract 2-8 keyframes according to the poses and remove the ordinary background (Sec. 3.1), and then estimate the inner-body masks using the proposed SUNet (Sec. 3.2). Finally, we obtain the 3D inner-body model by the proposed IITNet (Sec. 3.3).

3.1 Keyframe Selection

When the subject is rotating in front of a single camera, there is no need to use all the frames in the video. Nevertheless, manually selecting the keyframes of the video is not feasible. Therefore, we adopt a simple but effective strategy to select important perspectives as the keyframes according to the keypoints (especially the shoulder keypoints) in the video. The rotation starts with the person facing the camera, where the distance between two shoulder keypoints is the largest D_{max}. We consider it to be the front view and use it as the first keyframe. Taking 8 keyframes for example, we select the keyframes when the distance Dis satisfies the following equation:

$$Dis = D_{max} \cdot |\cos(\frac{\pi}{4}n)|, \quad n \in \{0, 1, ..., 7\}\,, \tag{1}$$

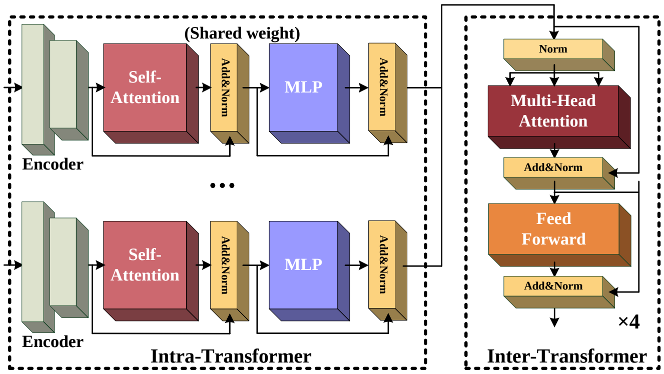


Fig. 3. Details of our Intra-Inter Transformer Network.

where n is the index of keyframes. After the keyframe selection, the ordinary background is removed by MODNet [36].

3.2 Semantic-guided Undressing Network

To alleviate the ambiguities brought in by directly estimating the 3D inner-body from the video, we propose to first estimate 2D inner-body masks and then reconstruct the 3D inner-body model. The first stage can learn semantically related inner-body features.

Given the keyframes $F = \{\mathbf{F}_0, ..., \mathbf{F}_{N-1}\}$ (N is the number of keyframes), SUNet adopts an encoder-decoder structure based on U-Net [37] to estimate the 2D inner-body masks $M = \{\mathbf{M}_0, ..., \mathbf{M}_{N-1}\}$. Instead of taking the source images as the input, we concatenate the source images, the pose heatmaps $H = \{\mathbf{H}_0, ..., \mathbf{H}_{N-1}\}$, and the semantic segmentation maps $S = \{\mathbf{S}_0, ..., \mathbf{S}_{N-1}\}$ as the inputs to predict more accurate inner-body masks. The semantic segmentation map $\mathbf{S}_n$ of each keyframe n is calculated by SCHP [38], and the heatmap $\mathbf{H}_n$ of each keyframe n is obtained from the joint coordinates detected by OpenPose [39]:

$$\mathbf{H}_n = \exp\left(-\frac{(x - \mathbf{J}_n^x)^2 + (y - \mathbf{J}_n^y)^2}{2\sigma^2}\right), \tag{2}$$

where $\mathbf{J}_n^x$ and $\mathbf{J}_n^y$ are the x-axis and y-axis coordinates of joints at keyframe n, respectively. σ is a hyper-parameter, and is set to 10 in our experiments. The pose information helps to maintain the rationality of the output mask, and the semantic information is beneficial to distinguish the clothes and the body, which allows the network to estimate human inner-body masks with higher accuracy.

The above process can be expressed by a function G as follows:

$$\mathbf{M}_n = \mathrm{G}(\mathbf{F}_n, \mathbf{S}_n, \mathbf{H}_n), \quad n \in \{0, 1, ..., N-1\}. \tag{3}$$

We multiply the inner-body masks M by the image semantic segmentation maps S to get the semantic inner-body masks $I = \{\mathbf{I}_0, ..., \mathbf{I}_{N-1}\}$.

3.3 Intra-Inter Transformer Network

Given the estimated semantic inner-body masks $I = \{\mathbf{I}_0, ..., \mathbf{I}_{N-1}\}$ and the corresponding 2D joints, we learn to infer the SMPL+O parameters including the body shape $\boldsymbol{\beta}$, personalized body offset $\boldsymbol{O}$, and 3D poses $P = \{\boldsymbol{\theta}_0, ..., \boldsymbol{\theta}_{N-1}\}$. The pose $\boldsymbol{\theta}$ of each frame is dynamic and changing during rotation, but the shape parameters $\boldsymbol{\beta}$ and personalized body offset $\boldsymbol{O}$ should be the same. Therefore, we estimate consistent shape parameters and dynamic pose parameters for all the frames. In order to deal with changing poses in different frames, different from [35] that averages the features of multi-frames, we take the input images as temporal signals to select more useful information. A natural idea is to use Recurrent Neural Networks (RNN) for image sequence fusion. However, RNN cannot be calculated in parallel, *i.e.*, the current moment depends on the output of the previous moment and cannot well fuse the features from multiple perspectives at the same time. Therefore, we design an *intra-inter transformer* network to make better use of pose-inconsistent multi-view information. As shown in Fig. 3, we first extract the feature C of the inner-body semantic image and pose features by a 2D convolutional encoder. After that, an *intra-transformer* is proposed to pay attention to the effective features within the frame by calculating the correlation between the features of each position i and any position j in the feature map. The intra-transformer consists of a single self-attention and a multi-layer perceptron (MLP), and computes the output feature map in a non-local manner:

$$\begin{aligned} U_{ij} &= \psi(\frac{1}{\sqrt{d_0}} q(C_i) k(C_j)^T), \\ C^{'} &= U \cdot v(C) + C, \\ U^{'} &= \mathcal{F}_{\text{BN}}(\mathcal{F}_{\text{MLP}}(\mathcal{F}_{\text{BN}}(C^{'}))), \end{aligned} \tag{4}$$

where ψ represents the softmax operator, and d_0 is the scaling factor. $q(\cdot)$, $k(\cdot)$ and $v(\cdot)$ are learnable query, key and value embedding functions of the intra-transformer, respectively. U is the attention map, and U_{ij} presents the value at position (i, j) on the attention map. $\mathcal{F}_{\text{BN}}$ and $\mathcal{F}_{\text{MLP}}$ represent batch normalization and multi-layer perception, respectively. Note that the intra-frame transformer is applied to individual frames so the frame index is omitted for simplicity.

To better integrate the features of different frames and obtain the consistent body shape parameters and personalized body offset, we design an *inter-transformer*. Similar to ViT [40], we flatten 2D features U_n of each keyframe and concatenate all the features in depth axis as $[Z_0, ..., Z_{N-1}]$. Then, we add a learnable embedding Z_{head} to $[Z_0, ..., Z_{N-1}]$. Finally, we input $Z_{\text{total}} = [Z_{\text{head}}; Z_0, ... Z_{N-1}]$ into the inter-transformer to extract the most important image features from different frames to reconstruct more accurate results. The values of fusion features Z_{fuse} are weighted based on their respective contributions. The inter-transformer can be formulated as:

$$\begin{aligned} Z_{\text{total}}^{'} &= \mathcal{F}_{\text{MSA}}(\mathcal{F}_{\text{LN}}(Z_{\text{total}})) + Z_{\text{total}}, \\ Z_{\text{fuse}} &= \mathcal{F}_{\text{LN}}(\mathcal{F}_{\text{FFN}}(\mathcal{F}_{\text{LN}}(Z_{\text{total}}^{'}))), \end{aligned} \tag{5}$$

where $\mathcal{F}_{\text{MSA}}$ and $\mathcal{F}_{\text{LN}}$ represent multi-head self-attention and layer normalization, respectively. $\mathcal{F}_{\text{FFN}}$ is a feed-forward network including two fully connected layers. Finally, based on Z_{fuse}, we calculate $\boldsymbol{\beta}$ and $\boldsymbol{O}$ through a fully connected layer (FC) and a four-step Graph Convolution

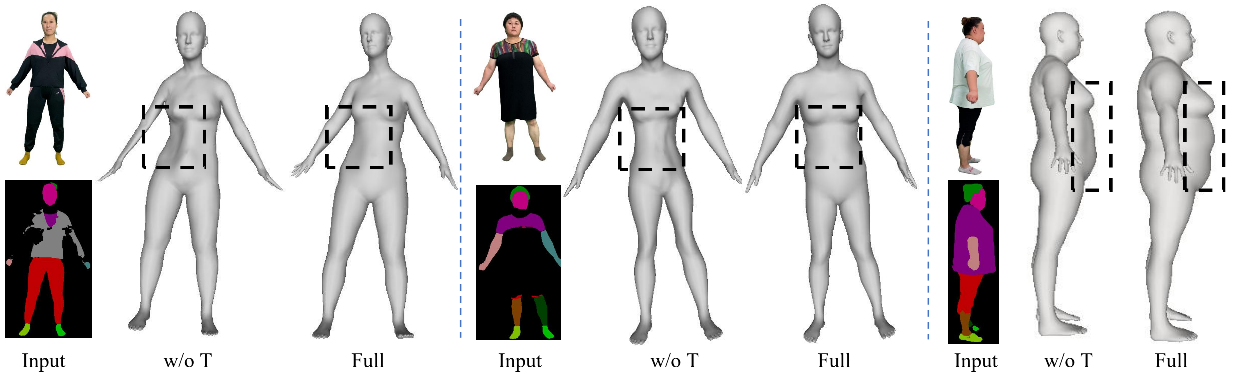


Fig. 4. The results with intra-inter transformers (Full) are better than those without intra-inter transformers (w/o T), regardless of inaccurate segmentation (first two cases) or overweight people (last case).

Network (GCN) with Chebyshev filters [41], respectively. Our intra-inter transformers can well deal with challenging cases with inaccurate segmentation and overweight people as shown in Fig. 4.

3.4 Loss Function

We introduce the losses of SUNet and IITNet, respectively. Our SUNet only uses 2D supervision, and IITNet uses both 3D and 2D supervision. We define $\tilde{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{O}}$ to be the ground-truth of the body shape and personalized body offset, respectively, and $\tilde{P} = \left\{\tilde{\boldsymbol{\theta}}_0, ..., \tilde{\boldsymbol{\theta}}_{N-1}\right\}$ is the ground-truth of 3D poses.

- **SUNet Loss.** To train the SUNet, we use Binary Cross Entry (BCE) loss function defined as:

$$L_{\mathrm{SU}} = -\tilde{M} \cdot \log M + (1 - \tilde{M}) \cdot \log(1 - M), \tag{6}$$

where M is the predicted inner-body mask and $\tilde{M}$ is the ground-truth.

- **IITNet Loss.** We use the following loss to train the IITNet:

$$\begin{aligned} L_{\mathrm{IIT}} = \lambda_{\mathrm{T}} L_{\mathrm{T}} + \lambda_{\mathrm{P}} L_{\mathrm{P}} + \lambda_{\mathrm{3D}} L_{\mathrm{3D}} + \\ \lambda_{\mathrm{2D}} L_{\mathrm{2D}} + \lambda_{\mathrm{seg}} L_{\mathrm{seg}} + \lambda_{\mathrm{smpl}} L_{\mathrm{smpl}}, \end{aligned} \tag{7}$$

where λ_{T}, λ_{P}, λ_{3D}, λ_{2D}, λ_{seg} and λ_{smpl} are the weights that balance the contributions of individual loss terms. The losses are defined in detail as follows.

3D Vertex Loss in the Canonical T-pose ($\boldsymbol{\theta} = \boldsymbol{\theta}_T$). This loss supervises human body shape independently of pose:

$$L_{\mathrm{T}} = ||\mathcal{F}_{\boldsymbol{O}}(\boldsymbol{\beta}, \boldsymbol{\theta}_T, \boldsymbol{O}) - \mathcal{F}_{\boldsymbol{O}}(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\theta}}_T, \tilde{\boldsymbol{O}})||_2, \tag{8}$$

where $\boldsymbol{\theta}_T$ is the canonical T-pose, and $\mathcal{F}_{\boldsymbol{O}}(\cdot)$ represents SMPL+O model similar to [42].

3D Vertex Loss in Posed Space. This loss forces the posed mesh to be as close as possible to the ground-truth:

$$L_{\mathrm{P}} = \sum_{i=1}^{N} ||\mathcal{F}_{\boldsymbol{O}}(\boldsymbol{\beta}, \boldsymbol{\theta}_i, \boldsymbol{O}) - \mathcal{F}_{\boldsymbol{O}}(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\theta}}_i, \tilde{\boldsymbol{O}})||_2. \tag{9}$$

2D Segmentation Loss. We also regularize the body shape training by projecting the human body onto the images:

$$L_{\mathrm{seg}} = \sum_{i=1}^{N} ||\mathcal{F}_{\mathrm{render}}(\mathcal{F}_{\boldsymbol{O}}(\boldsymbol{\beta}, \boldsymbol{\theta}_i, \boldsymbol{O}), c) - b(\tilde{\mathrm{I}}_i)||_2, \tag{10}$$

where $\mathcal{F}_{\mathrm{render}}(\cdot)$ is a differentiable renderer which renders the human mesh onto the frame i by camera c, and $b(\tilde{\mathrm{I}}_i)$ is the binary segmentation inner-body mask.

3D Joint Loss. We regularize the pose training by imposing a loss on the joints in Euclidean space:

$$\begin{aligned} \boldsymbol{X}_i &= \boldsymbol{J}_{\mathrm{regressor}}(\mathcal{F}_{\boldsymbol{O}}(\boldsymbol{\beta}, \boldsymbol{\theta}_i, \boldsymbol{O})), \\ L_{\mathrm{3D}} &= \sum_{i=1}^{N} ||\boldsymbol{X}_i - \tilde{\boldsymbol{X}}_i||_2, \end{aligned} \tag{11}$$

where $\boldsymbol{X}$ represents 3D joints which are computed from the body vertices with a pretrained linear regressor $\boldsymbol{J}_{\mathrm{regressor}}$.

2D Joint Loss. Similar to the 2D segmentation projection loss L_{seg}, we further project 3D joints to the images to supervise 2D joins:

$$L_{\mathrm{2D}} = \sum_{i=1}^{N} ||\boldsymbol{x}_i - \tilde{\boldsymbol{x}}_i||_2, \tag{12}$$

where $\boldsymbol{x}$ represents 2D joints which are the 2D projections of 3D joints.

Parameter Loss. In addition, we train the model to optimize the pose and shape using a direct loss on the predicted parameters $\boldsymbol{\theta}, \boldsymbol{\beta}$ by:

$$L_{\mathrm{smpl}} = \lambda_\theta \sum_{i=1}^{N} ||\boldsymbol{\theta}_i - \tilde{\boldsymbol{\theta}}_i||_2 + \lambda_\beta ||\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}||_2, \tag{13}$$

where λ_θ and λ_β are the balance weights.

3.5 Neural Optimization

Although we can train our model on a large dataset with diverse people and clothes, due to the domain gap, such a model still lacks the ability of generalization to the inputs out of the domain of training dataset. Therefore, we apply the one/few-shot learning, similar to [35], [43], to push the network to focus on each individual by several steps of fast personal adaptation using the input frames. Note that this optimization requires no 3D annotation as the network is fine-tuned using only 2D input. Specifically, given the estimated 2D joints and inner-body masks at the test time, by freezing most parameters of the IITNet, we use L_{seg} as well as L_{2D} described in Sec. 3.4 to optimize FC and GCN described in Sec. 3.3, which improves the generalization of the model.

3.6 Implementation Details

SUNet: We train SUNet using the Adam optimizer [44] with a learning rate of 1×10^{-4}, the batch size of 4, and 25 epochs. The learning rate is decayed by a factor of 0.5 every 10 epochs. Besides, we rotate, scale and crop the input images to enhance the robustness of the model. The training of SUNet can be achieved within 12 hours with an RTX 3090 GPU.

IITNet: The intra-transformer consists of a single self-attention module and a multi-layer perceptron (MLP), and the inter-transformer consists of four modules with a feed-forward network followed by an attention layer. The multi-head self-attention has four attention heads and the embedding dimension D is set to 256. We train IITNet using the Adam optimizer [44] with a learning rate of 1×10^{-4}, batch size of 8, decay of 0.01, and 60 epochs. For better performance, we train and test the male model and the female model separately. We train the male model and the female model with an RTX 3090 GPU, which takes 36 hours totally.

4 Experiments

4.1 Dataset

Because we aim to solve a new problem, there is no available datasets with self-captured videos of clothed persons and the corresponding 3D inner-bodies. Therefore, we collect a dataset called *Inner-Body Under Clothing (IBUC)*, which contains 1203 pairwise 3D inner-bodies (scans with tight clothes) and 3D clothed-bodies (scans with daily clothes) of 344 persons (159 males and 185 females) together with their self-captured A-pose videos. The subjects are collected from the talent market and each subject signs a license agreement. Our dataset contains over 500 different garments, the styles of which include T/long shirt, short/long/down coat, hooded jacket, pants, and skirt/dress, ranging from tight to loose. We randomly select 80 models for test, 40 for validation and the remaining for training.

In order to obtain topology-consistent inner-bodies for supervision, we non-rigidly register SMPL+O model to the inner-body scans. Specifically, we first render each inner-body scan from 32 viewpoints and find 2D keypoints of body, face and hands using OpenPose [39]. Then, we obtain 3D landmarks by minimizing the 2D re-projection error to the detected 2D keypoints, and optimize the pose and shape parameters of the SMPL model. Finally, we adopt ED graph-based non-rigid deformation and per-vertex refinement [5] to obtain a fitted mesh. To deal with the small differences of poses between the scanned inner-body and clothed-body, we deform the inner-body scan along with its SMPL+O model by SMPL skeleton to align with the clothed-body scan by constraining the projection of inner-body scan to be inside that of the clothed scan. Some registered samples are shown in Fig. 5.

Besides, we also collect a synthetic dataset with CLO3D [45] for the training of SUNet, which contains 528 inner-bodies (300 males and 228 females) and 6168 clothed-bodies with the canonical A-pose. CLO3D can simulate realistic garment deformations on virtual avatars using a rich number of adjustable parameters based on real-world physics such as gravity. Some samples are shown in Fig. 6.

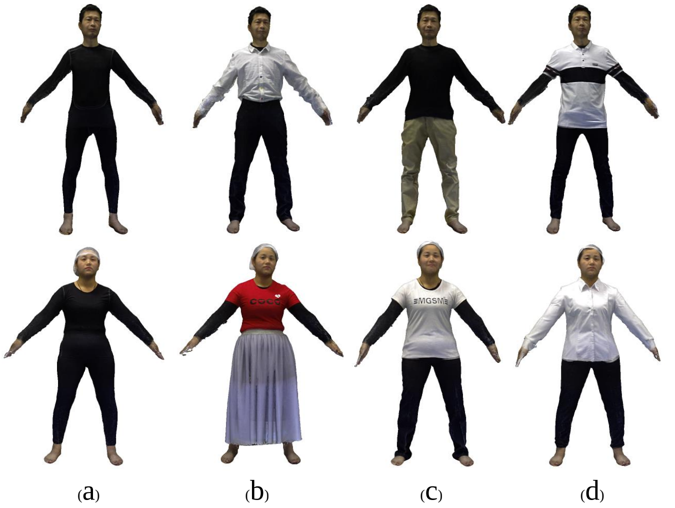


Fig. 5. Some samples in our *IBUC* dataset. The 3D inner-bodies (a) are aligned with the 3D clothed-bodies (b-d) of the same person.

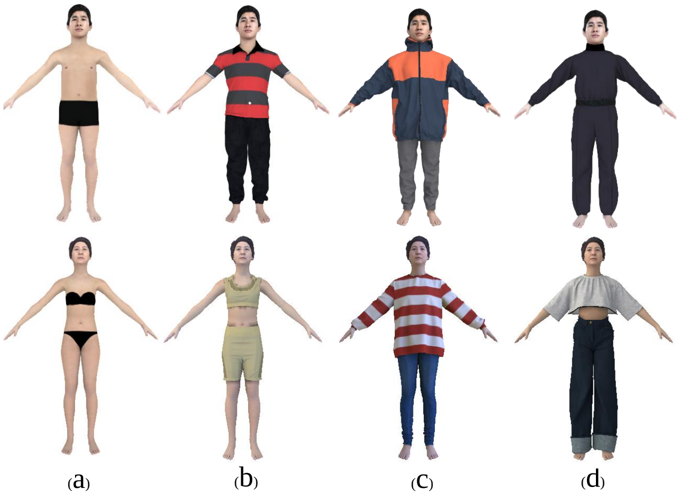


Fig. 6. Some samples in our synthetic dataset with pairwise inner-bodies (a) and clothed-bodies (b-d).

4.2 Comparison

Because our work aims to solve a new problem, there is no exactly matched state-of-the-art methods to compare with. Hence, we compare with the state-of-the-art image-based and video-based human body reconstruction methods: PyMAF [13], Metra [46] and TCMR [47]. We ignore comparisons with some work [10], [11], [20], [21], [26] that have already been compared. We use the official weights of the compared methods, and try our best to fine-tune them on the training and validation sets of our dataset. We automatically extract eight frames as our input, and the first frame with front view is used as the input of the single-image-based methods PyMAF [13] and Metra [46]. The whole video is used as the input of video-based method TCMR [47], and we average the shape parameters of SMPLs estimated by TCMR [47] for its better performance. Note that neither our method nor the compared methods are fine-tuned on *BUFF* dataset.

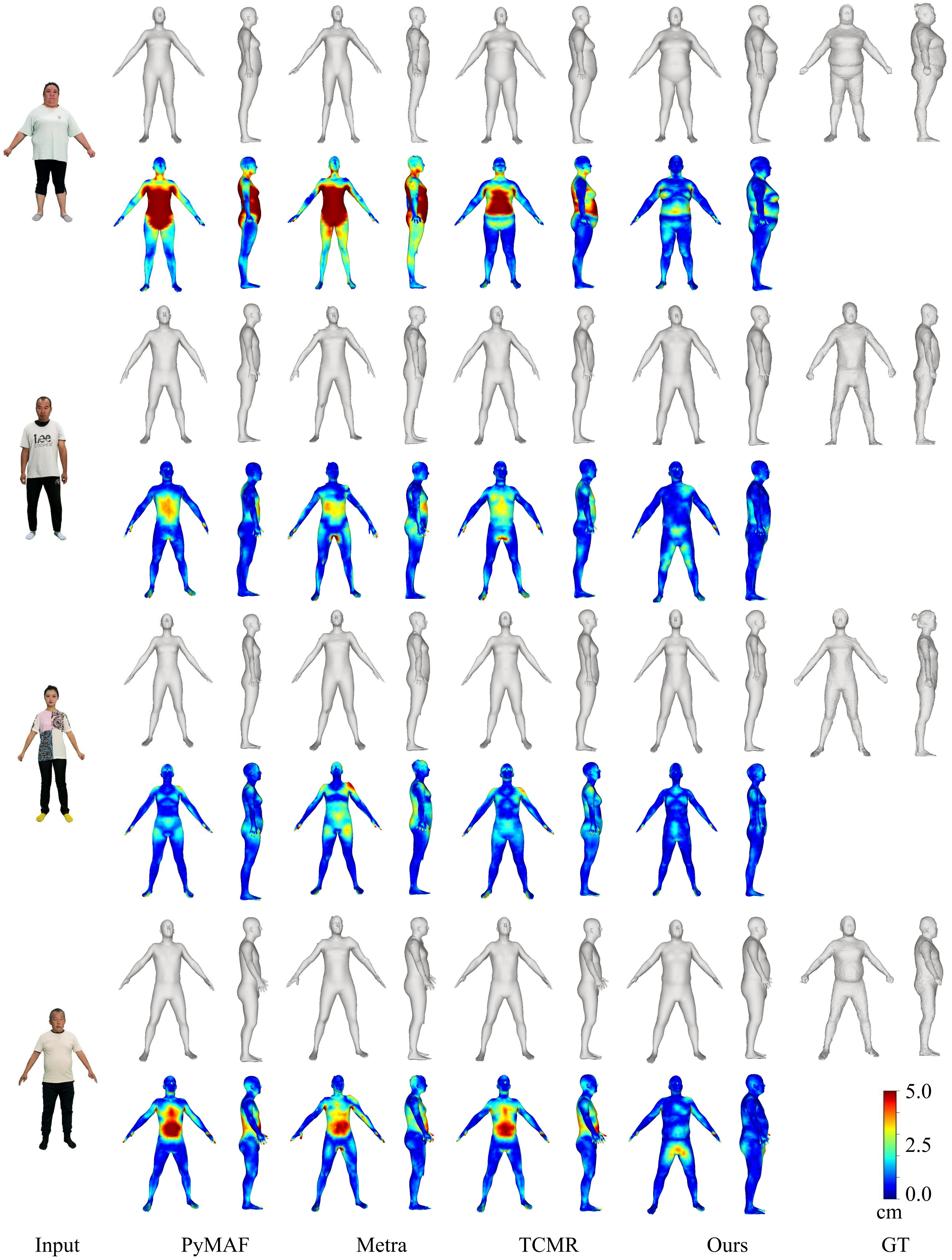


Fig. 7. Reconstructed 3D inner-bodies by PyMAF [13], Metra [46], TCMR [47] and our method on *IBUC* datasets. The errors between the reconstructed models and the ground-truths are color-coded on the reconstructed models for visual inspection.

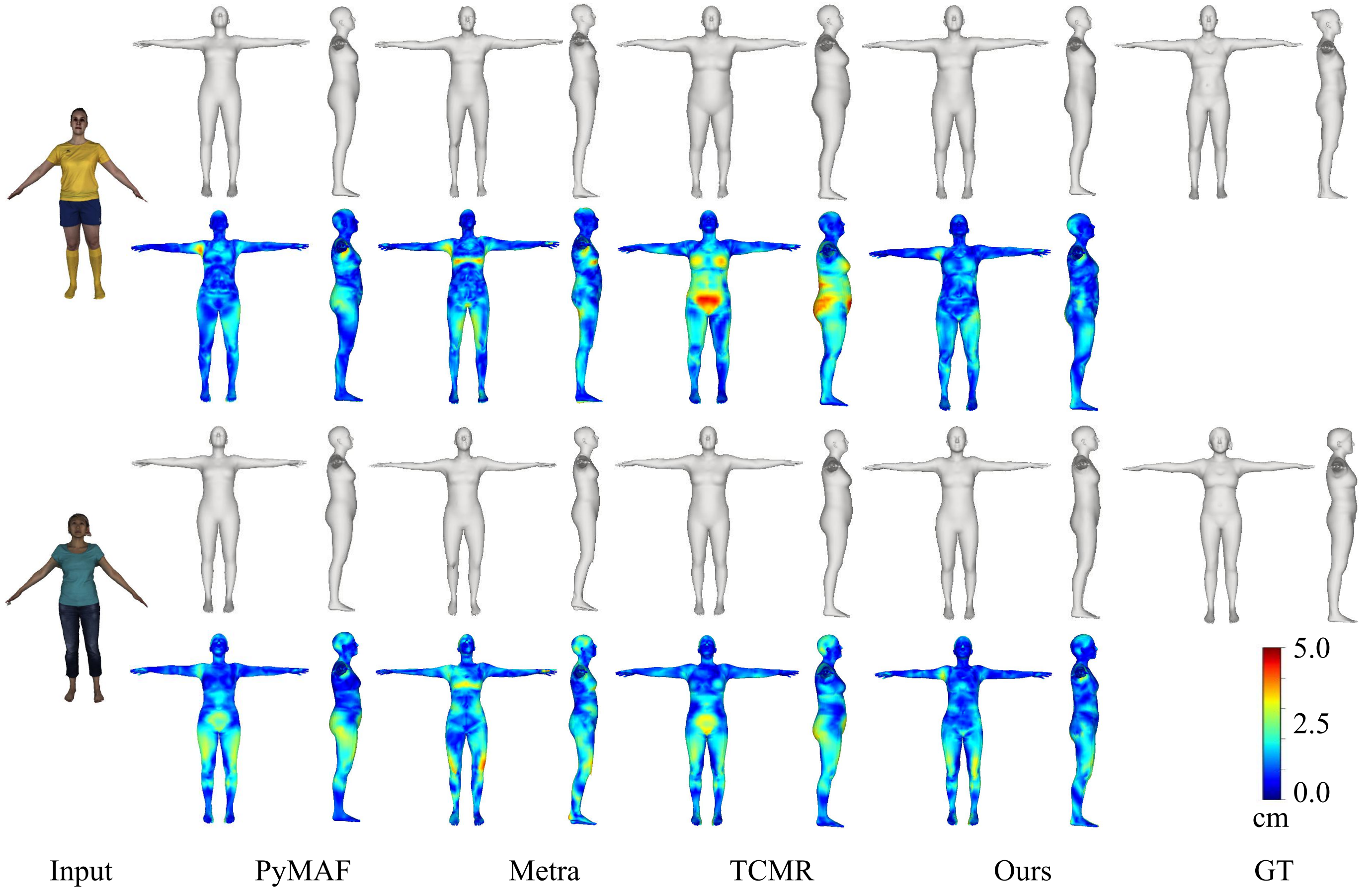


Fig. 8. Reconstructed 3D inner-bodies by PyMAF [13], Metra [46], TCMR [47] and our method on *BUFF* datasets. The errors between the reconstructed models and the ground-truths are color-coded on the reconstructed models for visual inspection.

TABLE 1
Quantitative comparison on two datasets.

Method	*BUFF*		*IBUC*	
	Mean (cm)↓	RMS (cm)↓	Mean (cm)↓	RMS (cm)↓
PyMAF	0.748	0.954	1.067	1.391
TCMR	0.980	1.252	0.945	1.223
Metra[1]	0.972	1.231	1.275	1.632
Ours	**0.732**	**0.949**	**0.730**	**0.956**

Quantitative Comparison. Table 1 gives the quantitative results on *BUFF* [48] and our *IBUC* dataset. Because the public dataset *BUFF* does not provide self-captured A-pose videos of clothed people that correspond to the 3D inner-bodies, we need to animate their scanned clothed models to generate the self-captured videos. Besides, the pose of the ground-truth inner-body scan is slightly different from the pose of the estimated model, and hence, for quantitative evaluation, we deform the output model to have the same pose as the ground-truth inner-body scan by optimizing the pose parameters and the scale of the output model. We calculate Mean and Root-Mean-Square (RMS) errors between the reconstructed model and the ground-truth scan across all the models using the standard Metro tool [49][2]. As shown in Table 1, our method outperforms the other methods in terms of all the metrics, which indicates that our model achieves the best accuracy of inner-body reconstruction from a single RGB camera.

TABLE 2
Quantitative evaluation for SUNet ablation study.

Method	IoU↑	XOR↓	Mean (cm)↓	RMS (cm)↓
w/o Seg.	0.9411	0.06234	1.246	1.621
w/o Pose	0.9398	0.06373	0.890	1.161
Full	**0.9423**	**0.06073**	**0.730**	**0.956**

TABLE 3
Quantitative evaluation for IITNet ablation study.

Method	w/o T	Intra-T	Inter-T	Full
Mean (cm)↓	0.832	0.737	0.744	**0.730**
RMS (cm)↓	1.105	0.980	0.975	**0.956**

[1] Metra is a non-parametric method which has a few wrong pose registrations for quantitative evaluation, and hence 5 samples of *BUFF* and 8 samples of *IBUC* are excluded.

[2] Metro is a popular tool designed to evaluate the difference between two triangular meshes. It adopts an approximated approach based on surface sampling and point-to-surface distance computation. Because the source and target models are two surface meshes instead of point sets, it is more suitable to use this tool for quantitative evaluation.

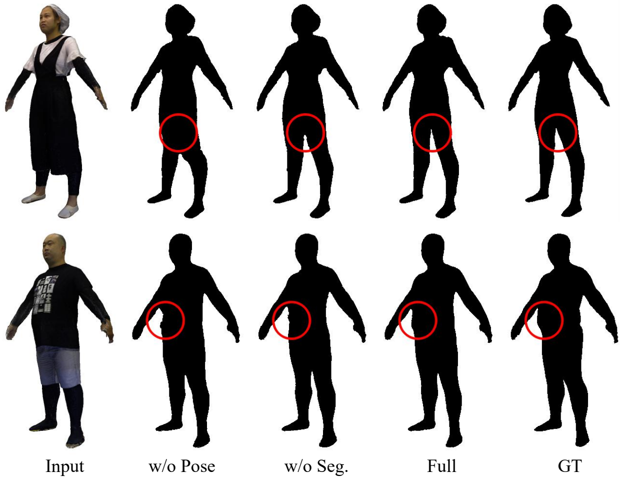


Fig. 9. Qualitative results of SUNet ablation study.

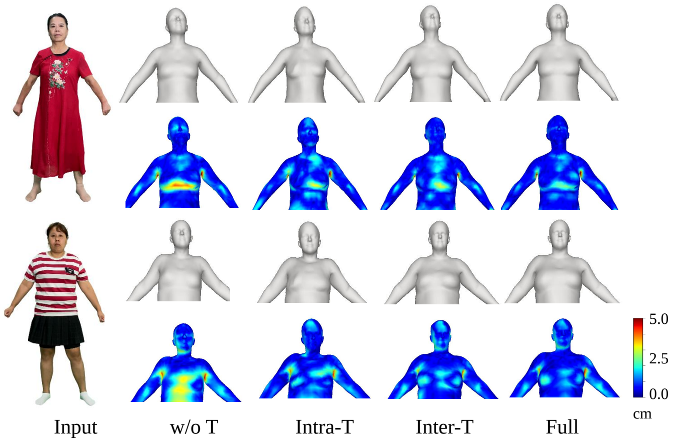


Fig. 10. Qualitative results of IITNet ablation study.

Qualitative Comparison. Fig. 7 and Fig. 8 shows some visual results on our *IBUC* datasets and *BUFF*. It can be seen that PyMAF [13] and Metra [46] cannot estimate accurate inner-body shapes, due to the limited representation ability of parametric models or the ambiguities caused by the single view inference. TCMR [47] recovers temporally consistent and smooth 3D models from the video, but tends to estimate a fatter inner-body due to the influence of clothing. On the contrary, our approach reconstructs the most accurate and reasonable inner-bodies, benefiting from the accurate inference of 2D inner-body masks by our SUNet and the elegant design for intra-frame and inter-frame feature learning in our IITNet. Besides, our network is only pre-trained using our *IBUC* dataset without fine-tuning on *BUFF*, which demonstrates the generalizability of our approach to recover accurate inner-body from only a single RGB camera.

4.3 Ablation Study

• **SUNet Ablation Study.** We train three variants to prove our hypotheses and validate the effect of different inputs:

The SUNet without Segmentation Map (w/o Seg.). The input of this model is the combination of the source image and the heatmap.

The SUNet without Heatmap (w/o Pose). The input of this model is the concatenated source image and the semantic segmentation map in the depth axis.

The SUNet Full Model (Full). The input of this model contains the source image, the heatmap and the semantic segmentation map.

To demonstrate the effectiveness of using heatmap and segmentation map at the same time, we compare our method with the variants of only using heatmap or segmentation map. Table 2 gives the quantitative results on our dataset. We calculate Intersection over Union (IOU) and exclusive OR (XOR) between the predicted inner-body mask and the ground-truth. Mean and RMS errors of the reconstructed 3D models are also given. As shown in the table, our method with both segmentation map and heatmap achieves the most accurate inference. Some visual results are shown in Fig. 9. It can be seen that the results without heatmap (w/o Pose) are unstable and unreasonable, and the results without segmentation map (w/o Seg.) are less accurate than the results of the full model.

• **IITNet Ablation Study.** We train four variants to prove our hypotheses and validate the effect of our improvements:

Without Transformer (w/o T). This model averages the features from all the frames directly.

With Intra-transformer (Intra-T). This model only uses the intra-transformer to pay attention to the effective features within the frame.

With Inter-transformer (Inter-T). This model only uses the inter-transformer to integrate the features of different frames.

With Intra-Inter Transformer (Full). This model uses both intra-transformer and inter-transformer to make better use of pose-inconsistent multi-view information.

Table 3 gives the quantitative results in terms of Mean and RMS errors. The results of Intra-T and Inter-T are better than those of w/o T, which demonstrates the effectiveness of our transformer design. Our full model achieves the best performance, which verifies the importance of exploring both intra-frame and inter-frame information. Some visual results are shown in Fig. 10. The reconstructed inner-body model by the full model is the most accurate. As shown in Fig. 4, the results with intra-inter transformers are better than those without intra-inter transformers, regardless of inaccurate segmentation or overweight people.

To further demonstrate the effectiveness of IITNet, we also compare our method with a video-based method Octopus [35] and its variant Octopus-SUNet that uses mask images from our SUNet on the IBUC dataset. Table 4 gives the quantitative results in terms of Mean and RMS errors. It can be seen that the results of Octopus-SUNet are better than those of Octopus, which demonstrates the effectiveness of our SUNet design. Meanwhile, our method outperforms them in terms of all the metrics, which further verifies the effectiveness of our IITNet design. Fig. 11 shows some visual results on *IBUC* dataset.

• **Neural Optimization.** To evaluate the effectiveness of our neural optimization, we compare the reconstructed inner-body results before and after optimization in Fig. 12. As shown in the figure, our neural optimization can obviously improve the accuracy of reconstructed results.

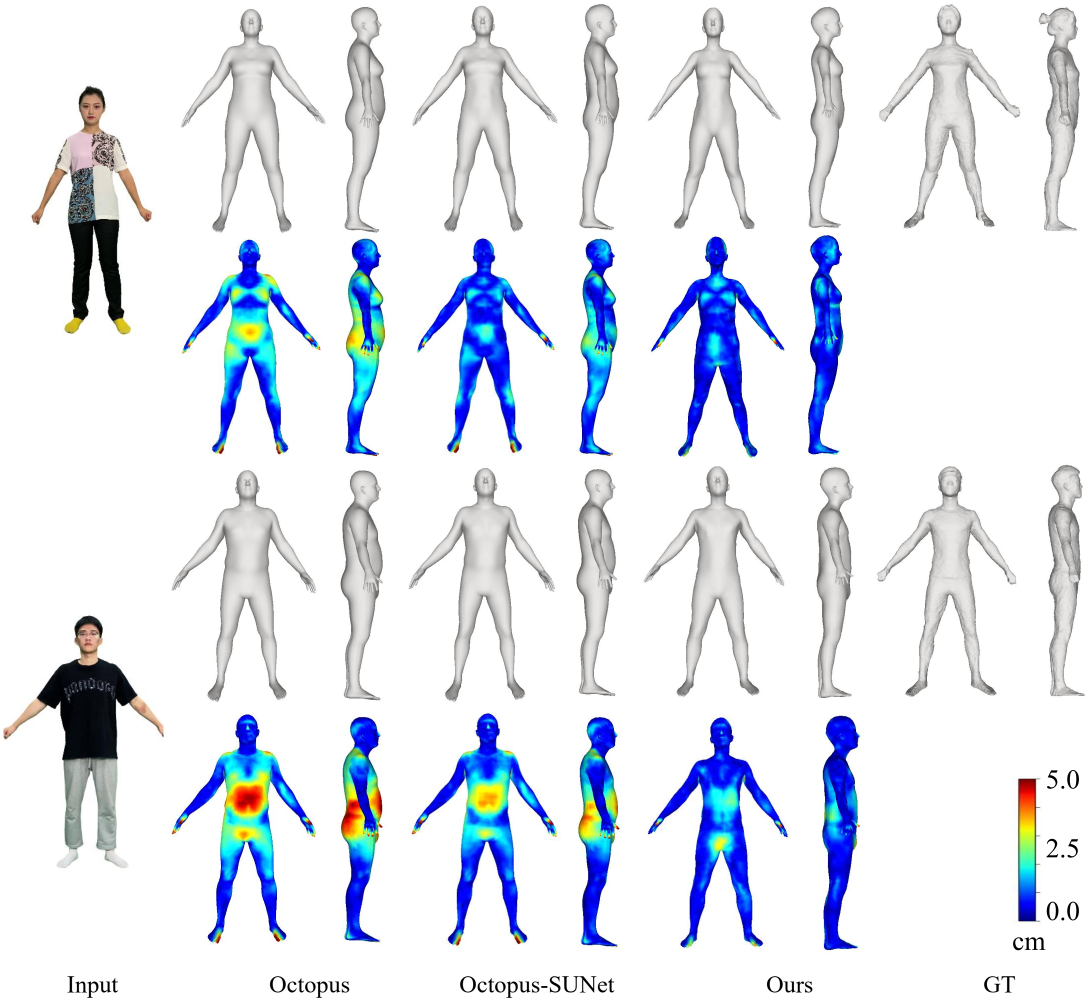


Fig. 11. Reconstructed 3D inner-bodies by Octopus [35], Octopus-SUNet and our method.The errors between the reconstructed models and the ground-truths are color-coded on the reconstructed models for visual inspection.

TABLE 4
Quantitative evaluation for the effectiveness of IITNet.

Method	Octopus	Octopus-SUNet	Ours
Mean (cm)↓	1.115	0.924	**0.730**
RMS (cm)↓	1.382	1.220	**0.956**

TABLE 5
Quantitative evaluation for the effectiveness of Using SMPL+O.

Representation	SMPL	Ours
Mean (cm)↓	0.849	**0.730**
RMS (cm)↓	1.108	**0.956**

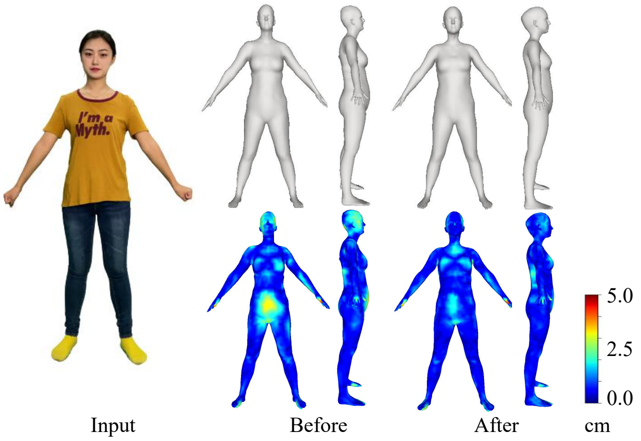


Fig. 12. Comparison results of before and after neural optimization. One of 8 image inputs is shown. The errors between the reconstructed models and the ground-truths are color-coded on the reconstructed models for visual inspection.

• **SMPL+O Ablation Study.** To demonstrate the effectiveness of using SMPL+O instead of SMPL representation, we compare our method with the variant of using SMPL as inner-body with our full pipeline. As shown in Table 5, our SMPL+O can obviously improve the accuracy of reconstructed results. Parametric models are not accurate inner-bodies because the pose and shape parameters have limited capacity to represent the body of an arbitrary person. Thus, we instead estimate the offsets over the parametric model to better represent personalized body shapes, given the estimated inner-body masks.

• **Different Numbers of Views.** We compare the perfor-

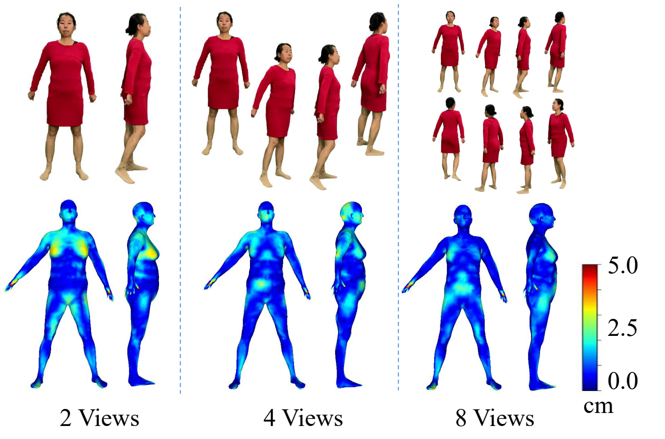


Fig. 13. Reconstructed 3D inner-body models with different number of views (images). The first row shows the input images, and the second row shows the corresponding errors color-coded on the reconstructed models.

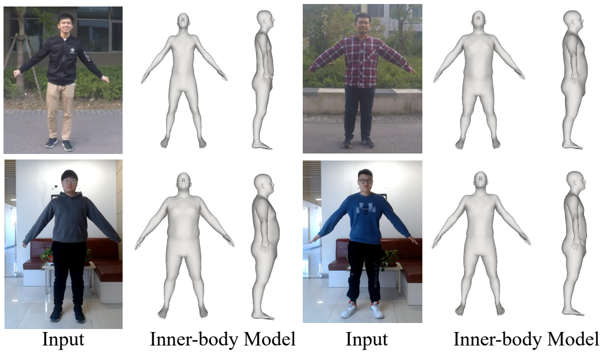


Fig. 14. The results generated by our method on in-the-wild videos. The first row shows the result from IPER dataset, and the second row shows the result captured by ourselves.

mances of using different numbers of input images (views) in Table 6. Mean and RMS errors between the pose-registered reconstructed model and the ground-truth scan across all the models are calculated by the standard Metro tool [49]. Note that the result of our 2-view version is not better than that of PyMAF [13] (Table 1). This is due to large pose variation in the two view images (as shown in Fig. 13), which increases the difficulty of utilizing multi-view information. As the number of views increases, our IITNet can better fuse the misaligned features.

The running times of these variants are also given, and all the experiments are run on a desktop with an RTX 3090 GPU and an i9-10900X CPU. We remove the startup time of the models for fair comparison. Some visual results are shown in Fig. 13. The reconstructed 3D inner-body by the model with 8 views is the most accurate with an acceptable running time, and hence we use this model for the comparison with the state-of-the-art methods.

- **Clothing Style.** We further explore the accuracy of human reconstruction under different clothes. Tab. 7 shows our results on different styles of clothing. Our method is robust to various clothing, including dress, coat, short/long skirts, short/long sleeves and shorts/trousers.

TABLE 6
Quantitative results and running times of using different numbers of views.

Views / Images	2	4	8
Mean (cm)↓	1.196	0.899	**0.730**
RMS (cm)↓	1.567	1.173	**0.956**
Times (s)	4.91	7.92	14.98

TABLE 7
Performance on different clothing styles (cm).

Cloth Type	Mean (cm)↓	RMS (cm)↓
Dress	0.745	0.953
Short/Long sleeves + Skirt	0.755	0.977
Long sleeves + Shorts	0.687	0.906
Long sleeves + Trousers	0.721	0.948
Coat + Trousers	0.735	0.951
Short sleeves + Trousers	0.730	0.959
Short sleeves + Shorts	0.719	0.943

4.4 Test on In-the-wild Videos

To better illustrate that our method can handle the inputs with complex backgrounds, we also test our method on in-the-wild videos. We use both outdoor videos from the IPER public dataset [50] and indoor videos captured with a mobile phone by ourselves as input. As shown in Fig. 14, our method is robust to complex backgrounds, regardless of whether the person is overweight or thin. Our method does not require a green background. The ordinary background can be well removed by MODNet [36].

5 Applications

Our method enables convenient body measurement and virtual try-on applications, as shown in Fig. 15.

Body Measurement. Scale is inherently ambiguous in monocular images. In order to get the correct body measurement, we need to scale the model to the actual height of the user. In case that the height is unknown, we can also directly estimate the height by fixing the camera height and view orientation [51]. For each measurement indicator, we mark several anchors on the template model (SMPL+O) and compute the geodesic distances between the anchors. Please note that we only need to mark once on the template model, which is the same for different persons.

Virtual Try-on. For better visualization, we replace the head of our estimated 3D inner-body model with the reconstructed head from the front image [52], and calculate the skin color of the inner-body from the face of the input image. We generate dynamic results according to the poses of the inner-body by physical-based methods using the clothes collected via CLO3D [45].

6 Conclusion and Discussion

Conclusion. In this paper, we aim to solve a new but valuable problem: reconstructing a personalized inner-body under clothing from a self-captured video. To alleviate the complexity and uncertainty, we propose a two-stage framework with a semantic-guided undressing network and

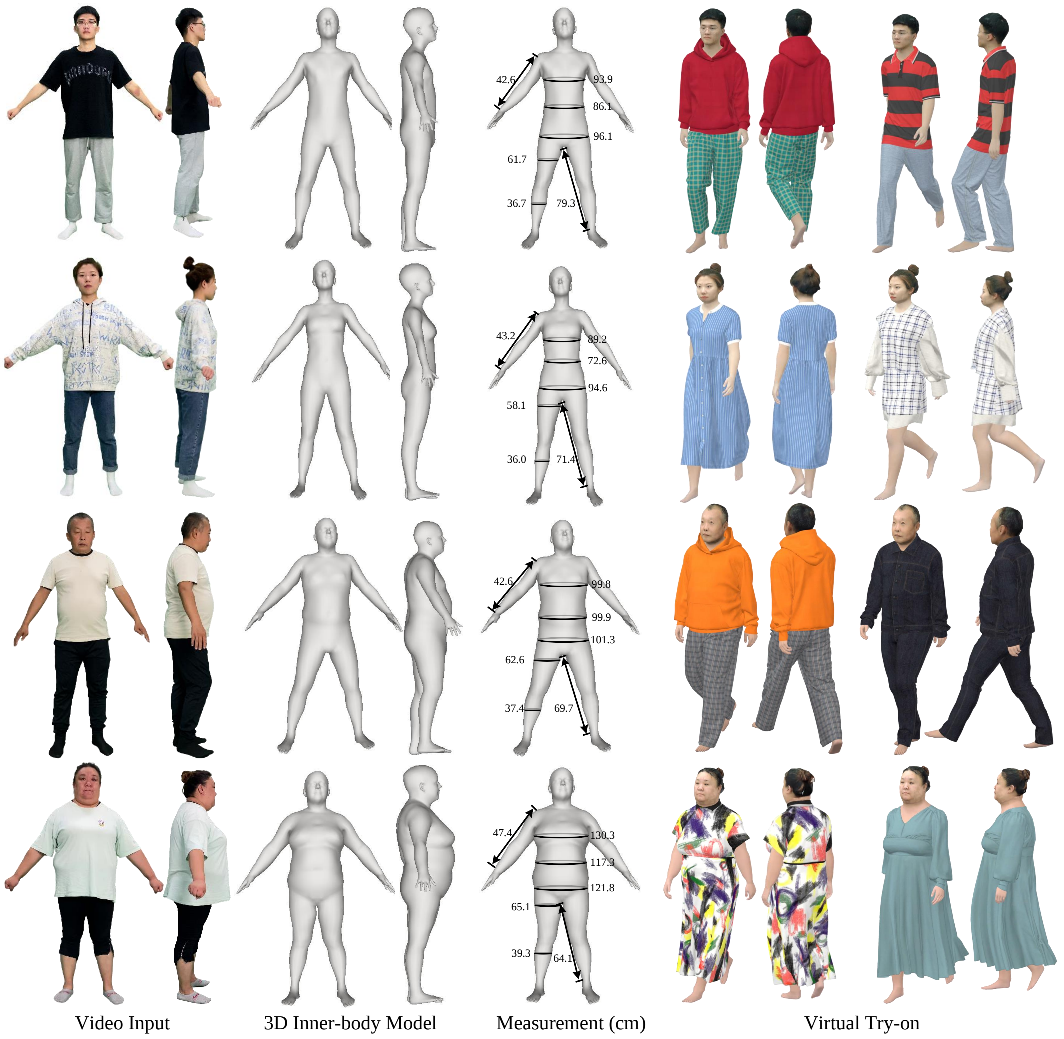


Fig. 15. Applications of body measurement and virtual try-on.

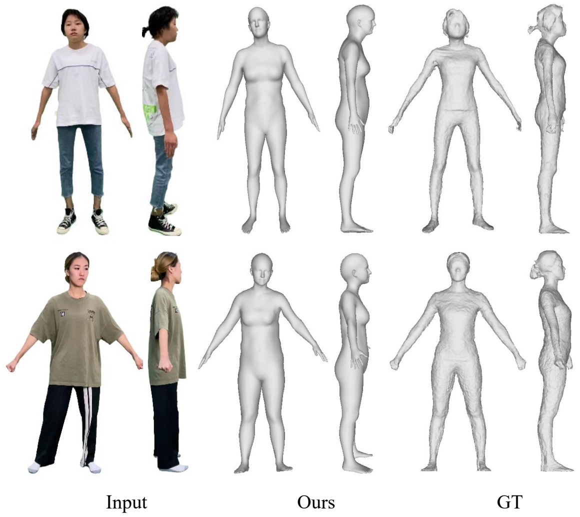


Fig. 16. Some examples of failure cases.

an intra-inter transformer network. To make better use of pose-inconsistent multi-view information, we propose intra-transformer and inter-transformer to learn the importance of intra-frame and inter-frame features. Experimental results on both public dataset and our dataset demonstrate that our method can infer 3D inner-bodies with the mean reconstruction error of 0.73cm within 15s. Our method also enables convenient virtual try-on and body measurement applications.

Limitations. Although our method can reconstruct the personalized inner-body model from a video when the subject is clothed, the results of cases with oversized clothes may not be good, due to less relevance between inner-body and clothing. Oversized clothing refers to the clothing that is larger than the proper size of the person, which is different from loose clothing that does not fit tightly around the body. As demonstrated in Table 7, our method is robust to various styles of clothing, including dress, coat, short/long skirts, short/long sleeves and shorts/trousers. However, our method may fail to estimate the accurate inner-body when the person is wearing oversized clothing, which could be solved by adding oversize data and priors in the future work. Fig. 16 gives some examples of failure cases. When the person is wearing tops and bottoms with inconsistent sizes (top row) or oversized clothes (bottom row), our method may fail to estimate accurate inner-bodies.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (62171317 and 62122058). We are grateful to Associate Editor and anonymous reviewers for their help in improving this paper.

References

[1] S. Pujades, B. Mohler, A. Thaler, J. Tesch, N. Mahmood, N. Hesse, H. H. Bülthoff, and M. J. Black, "The Virtual Caliper: Rapid creation of metrically accurate avatars from 3D measurements," *IEEE Transactions on Visualization and Computer Graphics*, pp. 1887–1897, 2019.

[2] A. C. S. Genay, A. Lécuyer, and M. Hachet, "Being an avatar "for Real": a survey on virtual embodiment in augmented reality," *IEEE Transactions on Visualization and Computer Graphics*, 2021.

[3] Y. Xu, S. Yang, W. Sun, L. Tan, K. Li, and H. Zhou, "3D virtual garment modeling from RGB images," in *Proc. IEEE International Symposium on Mixed and Augmented Reality*, 2019, pp. 37–45.

[4] O. Sarakatsanos, E. Chatzilari, S. Nikolopoulos, I. Kompatsiaris, D. Shin, D. Gavilan, and J. Downing, "A VR application for the virtual fitting of fashion garments on avatars," in *Proc. IEEE International Symposium on Mixed and Augmented Reality Adjunct*, 2021, pp. 40–45.

[5] X. Chen, A. Pang, W. Yang, P. Wang, L. Xu, and J. Yu, "Tightcap: 3D human shape capture with clothing tightness field," *ACM Transactions on Graphics*, vol. 41, pp. 1–17, 2021.

[6] J. Liang and M. C. Lin, "Shape-aware human pose and shape reconstruction using multi-view images," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 4352–4362.

[7] S. Wuhrer, L. Pishchulin, A. Brunton, C. Shu, and J. Lang, "Estimation of human body shape and posture under clothing," *Computer Vision and Image Understanding*, vol. 127, pp. 31–42, 2014.

[8] A. O. Bălan and M. J. Black, "The naked truth: Estimating body shape under clothing," in *Proc. European Conference on Computer Vision*, 2008, pp. 15–29.

[9] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black, "Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image," in *Proc. European Conference on Computer Vision*, 2016, pp. 561–578.

[10] 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/CVF International Conference on Computer Vision*, 2019, pp. 2252–2261.

[11] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik, "End-to-end recovery of human shape and pose," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7122–7131.

[12] J. Li, C. Xu, Z. Chen, S. Bian, L. Yang, and C. Lu, "HybrIK: A hybrid analytical-neural inverse kinematics solution for 3D human pose and shape estimation," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3383–3393.

[13] H. Zhang, Y. Tian, X. Zhou, W. Ouyang, Y. Liu, L. Wang, and Z. Sun, "PyMAF: 3D human pose and shape regression with pyramidal mesh alignment feedback loop," in *Proc. IEEE/CVF International Conference on Computer Vision*, 2021, pp. 11 446–11 456.

[14] N. Kolotouros, G. Pavlakos, D. Jayaraman, and K. Daniilidis, "Probabilistic modeling for human mesh recovery," in *Proc. IEEE/CVF International Conference on Computer Vision*, 2021, pp. 11 605–11 614.

[15] Z. Su, W. Wan, T. Yu, L. Liu, L. Fang, W. Wang, and Y. Liu, "MulayCap: Multi-layer human performance capture using a monoocular video camera," *IEEE Transactions on Visualization and Computer Graphics*, pp. 1862–1879, 2022.

[16] C. Ferrari, L. Casini, S. Berretti, and A. Del Bimbo, "Monocular 3D body shape reconstruction under clothing," *Journal of Imaging*, vol. 7, 2021.

[17] S. Yang, Z. Pan, T. Amert, K. Wang, L. Yu, T. Berg, and M. C. Lin, "Physics-inspired garment recovery from a single-view image," *ACM Transactions on Graphics*, vol. 37, pp. 1–14, 2018.

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

[19] D. Anguelov, P. Srinivasan, D. Koller, S. Thrun, J. Rodgers, and J. Davis, "SCAPE: Shape completion and animation of people," in *Proc. ACM SIGGRAIPH*, 2005, p. 408–416.

[20] N. Kolotouros, G. Pavlakos, and K. Daniilidis, "Convolutional mesh regression for single-image human shape reconstruction," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4501–4510.

[21] G. Moon and K. M. Lee, "I2L-MeshNet: Image-to-lixel prediction network for accurate 3D human pose and mesh estimation from a single RGB image," in *Proc. European Conference on Computer Vision*, 2020, pp. 752–768.

[22] K. Li, H. Wen, Q. Feng, Y. Zhang, X. Li, J. Huang, C. Yuan, Y.-K. Lai, and Y. Liu, "Image-guided human reconstruction via multi-scale graph transformation networks," *IEEE Transactions on Image Processing*, vol. 30, pp. 5239–5251, 2021.

[23] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. Osman, D. Tzionas, and M. J. Black, "Expressive body capture: 3D hands, face, and body from a single image," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10 975–10 985.

[24] P. Guan, A. Weiss, A. O. Balan, and M. J. Black, "Estimating human shape and pose from a single image," in *Proc. IEEE/CVF International Conference on Computer Vision*, 2009, pp. 1381–1388.

[25] T. Luan, Y. Wang, J. Zhang, Z. Wang, Z. Zhou, and Y. Qiao, "PC-HMR: Pose calibration for 3D human mesh recovery from 2D images/videos," *arXiv preprint arXiv:2103.09009*, 2021.

[26] M. Kocabas, N. Athanasiou, and M. J. Black, "VIBE: Video inference for human body pose and shape estimation," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5253–5263.

[27] Z. Zheng, T. Yu, Y. Wei, Q. Dai, and Y. Liu, "DeepHuman: 3D human reconstruction from a single image," in *Proc. IEEE/CVF International Conference on Computer Vision*, 2019, pp. 7739–7749.

[28] A. S. Jackson, C. Manafas, and G. Tzimiropoulos, "3D human body reconstruction from a single image via volumetric regression," in *Proc. European Conference on Computer Vision*, 2018.

[29] G. Varol, D. Ceylan, B. Russell, J. Yang, E. Yumer, I. Laptev, and C. Schmid, "BodyNet: Volumetric inference of 3D human body shapes," in *Proc. European Conference on Computer Vision*, 2018, pp. 20–36.

[30] S. Saito, Z. Huang, R. Natsume, S. Morishima, A. Kanazawa, and H. Li, "PIFu: Pixel-aligned implicit function for high-resolution clothed human digitization," in *Proc. IEEE/CVF International Conference on Computer Vision*, 2019, pp. 2304–2314.

[31] S. Saito, T. Simon, J. Saragih, and H. Joo, "PIFuHD: Multi-level pixel-aligned implicit function for high-resolution 3D human digitization," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 84–93.

[32] Z. Huang, Y. Xu, C. Lassner, H. Li, and T. Tung, "ARCH: Animatable reconstruction of clothed humans," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3093–3102.

[33] T. He, J. Collomosse, H. Jin, and S. Soatto, "Geo-PIFu: Geometry and pixel aligned implicit functions for single-view human reconstruction," *arXiv preprint arXiv:2006.08072*, 2020.

[34] H. Choi, G. Moon, and K. M. Lee, "Pose2Mesh: Graph convolutional network for 3D human pose and mesh recovery from a 2D human pose," in *Proc. European Conference on Computer Vision*, 2020, pp. 769–787.

[35] T. Alldieck, M. Magnor, B. L. Bhatnagar, C. Theobalt, and G. Pons-Moll, "Learning to reconstruct people in clothing from a single RGB camera," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1175–1186.

[36] Z. Ke, K. Li, Y. Zhou, Q. Wu, X. Mao, Q. Yan, and R. W. H. Lau, "Is a green screen really necessary for real-time portrait matting?" *arXiv preprint arXiv:2011.11961*, 2020.

[37] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2015, pp. 234–241.

[38] P. Li, Y. Xu, Y. Wei, and Y. Yang, "Self-correction for human parsing," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[39] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017, pp. 7291–7299.

[40] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words:

Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2021.

[41] A. Ranjan, T. Bolkart, S. Sanyal, and M. J. Black, "Generating 3D faces using convolutional mesh autoencoders," in *Proc. European Conference on Computer Vision*, 2018, pp. 704–720.

[42] T. Alldieck, M. Magnor, W. Xu, C. Theobalt, and G. Pons-Moll, "Video based reconstruction of 3D people models," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8387–8397.

[43] W. Liu, Z. Piao, J. Min, W. Luo, L. Ma, and S. Gao, "Liquid Warping GAN: A unified framework for human motion imitation, appearance transfer and novel view synthesis," in *Proc. IEEE/CVF International Conference on Computer Vision*, 2019, pp. 5904–5913.

[44] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[45] C. V. F. LLC., "CLO3D," https://www.clo3d.com/.

[46] K. Lin, L. Wang, and Z. Liu, "End-to-end human pose and mesh reconstruction with transformers," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 1954–1963.

[47] H. Choi, G. Moon, J. Y. Chang, and K. M. Lee, "Beyond static features for temporally consistent 3D human pose and shape from a video," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 1964–1973.

[48] C. Zhang, S. Pujades, M. J. Black, and G. Pons-Moll, "Detailed, accurate, human shape estimation from clothed 3D scan sequences," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4191–4200.

[49] P. Cignoni, C. Rocchini, and R. Scopigno, "Metro: measuring error on simplified surfaces," *Computer Graphics Forum*, vol. 17, pp. 167–174, 1998.

[50] W. Liu, Z. Piao, J. Min, W. Luo, L. Ma, and S. Gao, "Liquid warping gan: A unified framework for human motion imitation, appearance transfer and novel view synthesis," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 5904–5913.

[51] R. Zhu, X. Yang, Y. Hold-Geoffroy, F. Perazzi, J. Eisenmann, K. Sunkavalli, and M. Chandraker, "Single view metrology in the wild," in *Proc. European Conference on Computer Vision*, 2020.

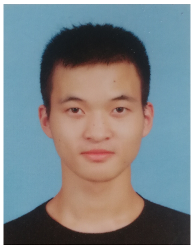
[52] Z. Gao, J. Zhang, Y. Guo, C. Ma, G. Zhai, and X. Yang, "Semi-supervised 3D face representation learning from unconstrained photo collections," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshop*, 2020, pp. 348–349.

Xiongzheng Li received the B.E. degree from East China Jiaotong University, Jiangxi Province, China, in 2019. He is currently pursuing the Ph.D. degree with the College of Intelligence and Computing in Tianjin University, Tianjin, China. His research interests include 3D vision and computer graphics.

Jing Huang received the B.E. degree from Tianjin University, Tianjin, China, in 2021. He is currently pursuing the M.E. degree in computer science in Tianjin University. His research interests include computer vision, 3D and computer graphics.

Jinsong Zhang received the B.E. and M.E. degree from Tianjin University, Tianjin, China, in 2018 and 2021, respectively. He is currently pursuing the Ph.D. degree in computer science from Tianjin University, Tianjin. His research interests include mainly in computer vision and computer graphics.

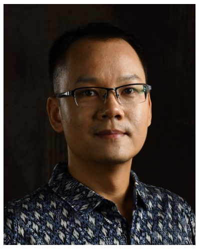
Xiaokun Sun received the B.E. degree in communication engineering from Hefei University of Technology, in 2021. He is currently pursuing the master's degree with the College of Intelligence and Computing in Tianjin University. His research interests include human 3D reconstruction and representation.



Haibiao Xuan received the B.E. degree from Northeastern University at Qinhuangdao in 2021 and he is currently pursuing the master degree with the College of Intelligence and Computing, Tianjin University, Tianjin, China. His research focuses on computer vision and computer graphics, with a current focus on human-scene interaction synthesis. His research focuses on computer vision and computer graphics, with a current focus on human-scene interaction synthesis.

Yu-Kun Lai received his bachelor and Ph.D. degrees in computer science from Tsinghua University in 2003 and 2008, respectively. He is currently a professor in the School of Computer Science & Informatics, Cardiff University. His research interests include computer graphics, geometry processing, image processing and computer vision. He is on the editorial boards of Computer Graphics Forum and The Visual Computer.

Yingdi Xie received Ph.D. Degree from Waseda University, Japan in 2010. He is currently with VRC Inc., a ubiquitous avatar platform company in Tokyo, and visiting researcher of Waseda University.



Jingyu Yang received the B.E. degree from the Beijing University of Posts and Telecommunications, Beijing, China, in 2003, and the Ph.D. degree (Hons.) from Tsinghua University, Beijing, in 2009. He has been a Faculty Member with Tianjin University, Tianjin, China, since 2009, where he is currently a Professor with the School of Electrical and Information Engineering. He was with Microsoft Research Asia (MSRA), Beijing, in 2011, within the MSRAs Young Scholar Supporting Program, and with the Signal Processing Laboratory, EPFL, Lausanne, Switzerland, in 2012 and from 2014 to 2015. His research interests include image processing, 3-D imaging, and computer vision.

Kun Li received the B.E. degree from Beijing University of Posts and Telecommunications, Beijing, China, in 2006, and the master and Ph.D. degrees from Tsinghua University, Beijing, in 2011. She visited École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland, in 2012 and from 2014 to 2015. She is currently an Associate Professor with the School of Computer Science and Technology, Tianjin University, Tianjin, China. Her research interests include dynamic scene 3D reconstruction and image/video processing.