Single-Image 3D Human Pose and Shape Estimation Enhanced by Clothed 3D Human Reconstruction

Leyuan Liu[1,2], Yunqi Gao[1], Jianchi Sun[1], and Jingying Chen[1,2(✉)]

[1] National Engineering Research Center for E-Learning, Central China Normal University, Wuhan, China
{gaoyunqi,sunjc0306}@mails.ccun.edu.cn

[2] National Engineering Laboratory for Educational Big Data, Central China Normal University, Wuhan, China
{lyliu,chenjy}@mail.ccnu.edu.cn

Abstract. 3D human pose and shape estimation and clothed 3D human reconstruction are two hot topics in the community of computer vision. 3D human pose and shape estimation aims to estimate the 3D poses and body shapes of "naked" humans under clothes, while clothed 3D human reconstruction refers to reconstructing the surfaces of humans wearing clothes. These two topics are closely related, but researchers usually study them separately. In this paper, we enhance the accuracy of the 3D human pose and body shape estimation by the reconstructed clothed 3D human models. Our method consists of two main components: the 3D body mesh recovery module and the clothed 3D human reconstruction module. In the 3D body mesh recovery module, an intermediate 3D body mesh is first recovered from the input image by a graph convolutional network (GCN), and then the 3D body pose and shape parameters are estimated by a regressor. In the clothed human reconstruction module, two clothed human surface models are respectively reconstructed under the guidance of the recovered 3D body mesh and the ground-truth 3D body mesh. At the training phase, losses which are described by the residuals among the two reconstructed clothed human models and ground truth are passed back into the 3D body mesh recovery module and used for boosting the body mesh recovery module. The quantitative and qualitative experimental results on THuman2.0, and LSP show that our method outperforms the current state-of-the-art 3D human pose and shape estimation methods.

Keywords: 3D Human Pose and Shape Estimation · Clothed 3D Human Reconstruction · Graph Convolutional Network · SMPL Parameter Regression

This work was supported by the National Natural Science Foundation of China under grant No. 62077026 and the Fundamental Research Funds for the Central Universities under grant No. CCNU22QN012.

H. Lu and J. Cai (Eds.): ISAIR 2023, CCIS 1998, pp. 33–44, 2024.
https://doi.org/10.1007/978-981-99-9109-9_4

1 Introduction

In order to describe 3D human poses and body shapes with a finer granularity while reducing the difficulty of algorithms, the majority of methods [3,8,11,24,25] usually represent human bodies by parametric models such as SMPL [16]. In this way, algorithms only need to output low-dimensional pose and shape parameters, which are then used to recover the corresponding 3D body meshes via the parametric models. Another hot topic is called clothed 3D human reconstruction [15,22,27], which refers to reconstructing 3D surface meshes of humans with clothes. Although 3D human pose and shape estimation and clothed 3D human reconstruction have different goals, representations, methodologies, and outputs, they are two closely related topics.

Fig. 1. Our method can recover accurate 3D poses and body shapes of humans wearing both tight-fitting and loose-fitting clothes.

Estimating 3D human pose and shape from monocular images is quite challenging, not only because it is an inherently ill-posed problem, but also due to complex body kinematic structures, various body shapes, and clothing occlusions. To address these challenges, two different paradigms have been investigated: optimization-based methods and regression-based methods. Optimization-based methods [1,21] usually suffer from local minima due to poor initialization. Thus, the recent mainstream methods [2,3,8,12,18,24,25] have focused on the regression-based paradigm, which usually employs deep learning techniques to regress 3D human poses and shapes directly from the input image information in an end-to-end manner. However, these methods often fail to produce satisfactory results on humans wearing complex and loose-fitting clothing, as the issue caused by clothing occlusions is not given sufficient consideration.

Recently, clothed 3D human reconstruction [15,22,27] has developed rapidly, and most of these methods employ estimated 3D body models as a geometrical prior. Although the estimated 3D body models can help clothed 3D human

reconstruction methods recover more plausible global topologies, inaccurate estimation of 3D human poses and shapes usually leads to poor clothed human reconstructions [15,27]. The accuracy of 3D human pose and shape estimation determines the quality of clothed 3D human models reconstructed by these methods. Conversely, the results of clothed 3D human reconstruction effectively indicate the accuracy of 3D human pose and shape estimation. Based on this insight, we argue that clothed 3D human reconstruction can be employed to enhance the accuracy of 3D human pose and shape estimation.

In this paper, we propose a 3D human pose and shape estimation method enhanced by clothed 3D human reconstruction. Our method consists of two main components: the 3D body mesh recovery module and the clothed human reconstruction module. In the 3D body mesh recovery module, an intermediate 3D body mesh is first recovered from an initial SMPL model by a GCN, and then the 3D body pose and shape parameters are estimated by a regressor. In the clothed human reconstruction module, two clothed human surface models are respectively reconstructed under the guidance of the recovered 3D body mesh and the ground-truth 3D body mesh. At the training phase, losses which are described by the residuals among the two reconstructed clothed human models and ground truth are passed back into the 3D body mesh recovery module and used for optimizing the body mesh recovery module. As illustrated in Fig. 1, our method can recover accurate 3D poses and body shapes of humans with both tight-fitting and loose-fitting clothes. Quantitative and qualitative experimental results show that our method achieves state-of-the-art performance on the THuman2.0 and LSP datasets.

In summary, the main contributions of this paper are three-fold:

- We propose a 3D human pose and shape estimation method enhanced by clothed 3D human reconstruction. Our method is the first method that employs reconstructed clothed 3D human models to enhance the accuracy of 3D human pose and shape estimation.
- We propose to use both the absolute and relative clothed 3D human reconstruction errors as losses pass them back into the 3D body mesh recovery module and use them for optimizing the body mesh recovery module.
- Our method recovers accurate 3D poses and body shapes of humans wearing both tight-fitting and loose-fitting clothes.

2 Related Work

In recent years, 3D technology has found extensive applications in various industries, such as transportation [14,26]. In addition, the field of 3D digital human body technology has also made great development and progress.

2.1 3D Human Pose and Shape Estimation

3D human pose and shape estimation methods can be roughly divided into two categories: optimization-based methods and regression-based methods.

Optimization-Based Methods. Optimization-based methods [1,6,13] attempt to fit a 3D body meshes to the image observation in an explicit iterative manner. Although these optimization-based methods have demonstrated high accuracy in many cases, they may be susceptible to local optima due to poor initialization. Additionally, the iterative optimization processes involved in these methods are time-consuming. **Regression-based methods** [2,3,8,12,18,24,25] directly regress the parameters from the input image. With the prosperity of deep learning, many researchers have recently shifted their focus from optimization-based methods to regression-based methods. Kanazawa et al. [8] proposed an end-to-end 3D body mesh recovery framework called HMR, which uses rich and useful mesh representation parameterized by shape and 3D joint angles and utilizes a generative adversary network to constrain body poses. Zeng et al. [24] recovered 3D body meshes by establishing a dense correspondence between the mesh and local image features in UV space. In contrast to most methods that regress SMPL parameters, works proposed by Choi et al. [2] and Kolotouros et al. [12] employ graph convolutional neural networks to estimate 3D locations of vertices on the 3D body models. Despite these regression-based methods yielding promising results on people with tight-fitting clothes, they often fail to produce satisfactory results on humans wearing complex and loose-fitting clothing.

2.2 Clothed 3D Human Reconstruction

In recent years, clothed 3D human reconstruction tends to be guided by the use of parametric human models. Parametric model guided clothed 3D human reconstruction methods [5,15,22,27,28] employ a 3D body model (e.g., SMPL [16]) to guide the reconstruction of the human surface. DeepHuman [28] uses the SMPL model to constrain the degrees of freedom in the output space. HEI-Human [15] and PaMIR [27] also employ the SMPL model as a geometrical prior when using implicit functions to reconstruct surface details of clothed humans. Although the estimated SMPL models can help clothed 3D human reconstruction methods recover more plausible global topology, inaccurate estimation of 3D human poses and shapes often results in poor clothed human reconstructions (Fig. 2).

3 Method

3.1 3D Body Mesh Recovery

Inspired by GCMR [12], we first employ a GCN to predict the vertex coordinates of the intermediate 3D body mesh and then use a regressor to estimate the SMPL model parameters. Given a single input image, visual features are extracted by a CNN-based encoder, e.g., ResNet-50 [4]. The visual features are then embedded in the GCN for predicting an intermediate 3D body mesh. Finally, the vertices of the intermediate 3D body mesh are input into the regressor for estimating the SPML model parameters.

For the GCN, we start from a non-parametric 3D deformable graph, which is initiated by a T-pose SMPL template. As the original SMPL model has as many

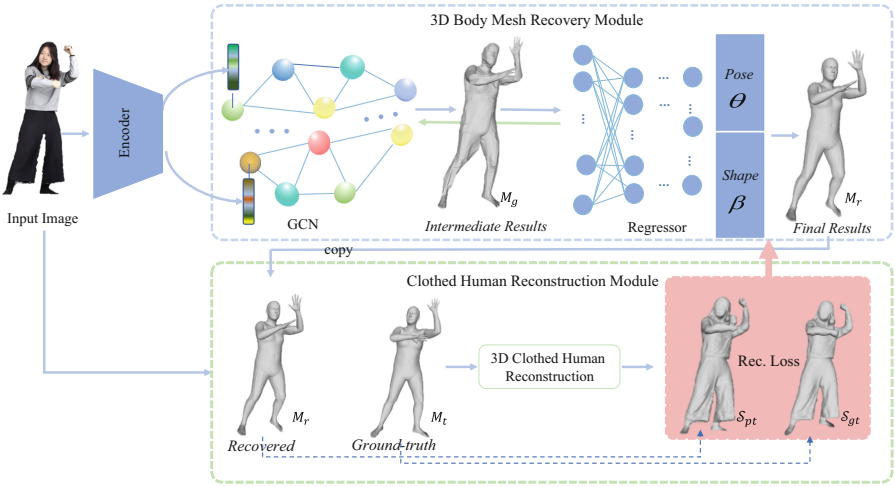

Fig. 2. Overview of our method. Our method consists of two main components: the 3D body mesh recovery module and the clothed human reconstruction module. Given an input image, visual features are first extracted by an image encoder, the image features are then embedded into a GCN for recovering an intermediate 3D body mesh (M_g). The intermediate 3D body mesh is fed into a parameter regressor for regressing the pose and shape parameters of the SMPL model (M_r). Two clothed human models ($\mathcal{S}_{pt}$ and $\mathcal{S}_{gt}$) are respectively reconstructed under the guidance of the recovered 3D body mesh (M_r) and the ground-truth 3D body mesh (M_t). Losses which are described by the residuals among the reconstructed clothed human models $\mathcal{S}_{pt}$ and ground-truth clothed human models $\mathcal{S}_{gt}$ are employed for optimizing the body mesh recovery module.

as 6890 vertices, we use a down-sampling strategy to simplify it to $N(N < 6890)$ vertices. Driven by the visual features embedded in each vertex of the graph, GCN is employed to shift the vertices. Following the work of Kipf et al. [10], our GCN is formulated as:

$$\hat{V} = \tilde{A}(DT \oplus F)W \tag{1}$$

where $T \in \mathbb{R}^{K\times 3}$ and $D \in \mathbb{R}^{N\times K}$ respectively denote the SMPL template and the down-sampling matrix, $\tilde{A} \in \mathbb{R}^{N\times N}$ denotes the row-normalized adjacency matrix of the graph, $F \in \mathbb{R}^{N\times f}$ is the visual feature vector, $W \in \mathbb{R}^{(3+f)\times 3}$ denotes the weight matrix, and $\hat{V} \in \mathbb{R}^{N\times 3}$ is the predicted coordinate vector. To facilitate regressing the SMPL model parameters, we up-sample $\hat{V}$ to 6890 vertices using the bilateral interpolation algorithm and obtain the intermediate 3D body mesh $M_g \in \mathbb{R}^{6890\times 3}$. Essentially, our GCN is equivalent to performing a full join operation for each vertex with visual features and then performing a neighborhood averaging operation. Neighborhood averaging is essential for generating high-quality shapes since it forces neighboring vertices to have similar features so that the output shape is smooth. In addition to regressing the coordinates of each vertex on the intermediate 3D body mesh, our GCN also estimates

camera parameters of the weakly perspective camera model, i.e., the scale and translation parameters $[s, t], t \in \mathbb{R}^2$.

To obtain more smooth and regular 3D body meshes, we employ a regressor to estimate the pose and shape parameters of the parametric SMPL model given the intermediate 3D body mesh as input. A specific 3D human body mesh ($M_r(\beta, \theta) \in \mathbb{R}^{6890X3}$) is described by SMPL [16] using a set of pose parameters ($\theta \in \mathbb{R}^{24X3}$) and a set of shape parameters ($\beta \in \mathbb{R}^{10}$):

$$M_r(\beta, \theta) = W(T_P(\beta, \theta), J(\beta), \theta, \omega) \tag{2}$$

where $T_P(\beta, \theta) = \bar{T} + B_S(\beta) + B_P(\theta)$, $J(\beta; \mathcal{J}, \bar{T}, S) = \mathcal{J}(\bar{T} + B_S(\beta; S))$, $\bar{T}$ is the standard human body model, $W(\cdot)$ is the fusion mask function, $J(\cdot)$ describes the displacement of joint points due to body size change, ω is the fusion weight matrix, $B_P(\cdot)$ is the pose fusion function, $B_S(\cdot)$ is the shape fusion function, $\mathcal{J}$ is a function that transforms rest vertices into rest joints.

Our parameter regressor is simply implemented by a three-layer multi-layer perceptron, which takes the 3D vertex coordinates of the intermediate 3D body mesh as input and outputs the pose (θ) and shape (β) parameters of the SMPL model. The estimated pose and shape parameters are converted to 3D body mesh using the SMPL model described by Eq. 2. So that we can compare whether the 3D body mesh described by the estimated parameters is consistent with the intermediate 3D body mesh.

3.2 Clothed Human Reconstruction

3D body meshes are employed to guide clothed human reconstruction. Similar to PIFu [20], we define the surface of a clothed 3D human model as a level set of an occupancy prediction function $\mathcal{F}(\cdot)$. For each 3D point p in the occupancy field, the occupancy prediction function predicts whether it is on the surface of the clothed 3D human model. To leverage image features and 3D body meshes for predicting the occupancy probability of p, our occupancy probability function $\mathcal{F}$ also takes the input image (I) and the 3D body mesh (M_*) as condition variables and thus is formulated as:

$$\mathcal{S}(p\,|I, M_*) = \mathcal{F}(\ddot{U}(\ddot{f}(I), \pi(p)), \dddot{U}(\dddot{f}(M_*), p)) \tag{3}$$

where $\ddot{f}$ and $\dddot{f}$ are two encoders that respectively extract features from the input image and the 3D body model, $\pi(\cdot)$ denotes the weak perspective transformation that maps the 3D coordinates of point p in the 2D feature plane, and $\ddot{U}(\cdot)$ and $\dddot{U}(\cdot)$ are two sampling functions that respectively take features from the feature maps extracted from the input image and the 3D body mesh. In practice, the two encoders ($\ddot{f}$ and $\dddot{f}$) and the occupancy prediction function ($\mathcal{F}$) are implemented by deep neural networks. In our method, $S(p) < 0.5$ indicates that point p is inside the surface while $S(p) > 0.5$ denotes point p is outside the surface. Hence, the surface of a clothed human model can be denoted as a set of points:

$$\mathcal{S}_* = \{p; S(p|I, M_*) := 0.5\} \tag{4}$$

where M_* can be expressed as M_r or M_t, $\mathcal{S}_*$ can be expressed as $\mathcal{S}_{pt}$ or $\mathcal{S}_{gt}$. Specifically, the recovered 3D body model M_r to guide clothed human reconstruction and obtain $\mathcal{S}_{pt}$, and the ground-truth 3D body model M_t to guide clothed human reconstruction and obtain $\mathcal{S}_{gt}$. To facilitate voxelized the reconstructed results, we converted this point set into a mesh using the Marching Cubes algorithm [17].

Then, the residuals among the reconstructed clothed human models $\mathcal{S}_{pt}$, $\mathcal{S}_{gt}$ and ground-truth clothed human model ($\mathcal{S}^*$) can be passed back into the 3D body mesh recovery module for boosting the accuracy of the 3D body pose and shape parameters.

3.3 Loss Functions

We employ a 3-step training scheme. (S1) We train the GCN to recover the intermediate body mesh. (S2) We fix the trained GCN and then train the regressor to estimate the pose and shape parameters of the 3D body. (S3) We unfix the GCN and then retrain the whole network. The loss functions used in these three steps are respectively represented by $\mathcal{L}_{gcn}$, $\mathcal{L}_{reg}$, and $\mathcal{L}_R$.

The GCN used for recovering the intermediate body mesh is trained using two kinds of supervision, that is, the mesh vertices alignment loss ($\mathcal{L}_v$) and the joints alignment loss ($\mathcal{L}_J$). Hence, $\mathcal{L}_{gcn}$ can be formulated as $\mathcal{L}_{gcn} = \lambda_v \mathcal{L}_v + \lambda_j \mathcal{L}_J$.

Besides the losses used to train the GCN, an additional parameter loss ($\mathcal{L}_p$) is also employed to train the regressor. So, $\mathcal{L}_{reg}$ can be formulated as $\mathcal{L}_{reg} = \lambda_v \mathcal{L}_v + \lambda_j \mathcal{L}_J + \lambda_p \mathcal{L}_p$.

As mentioned before, the results of clothed 3D human reconstruction indicate the accuracy of the 3D body pose and shape estimation. Hence, we use the residuals among the reconstructed clothed human models $\mathcal{S}_{pt}$, $\mathcal{S}_{gt}$ and ground-truth clothed human model $\mathcal{S}^*$ to describe the losses and pass them back into the 3D body mesh recovery module and used for optimizing the body mesh recovery module:

$$\mathcal{L}_{R1} = \frac{1}{n_p} \sum_{i=1}^{n_p} |\mathcal{S}_{pt}(p_i) - \mathcal{S}^*(p_i)|^2 \tag{5}$$

$$\mathcal{L}_{R2} = \mathcal{L}_{R1} - \frac{1}{n_p} \sum_{i=1}^{n_p} |\mathcal{S}_{gt}(p_i) - \mathcal{S}^*(p_i)|^2 \tag{6}$$

where n_p is the number of sampled points, and p_i is the sampled point. The $\mathcal{L}_{R1}$ loss represents the absolute reconstruction errors, while $\mathcal{L}_{R2}$ describes the relative reconstruction loss which removes reconstruction errors due to factors other than 3D human pose and shape estimation. Our clothed human reconstruction loss considers both the absolute reconstruction loss and the relative reconstruction loss:

$$\mathcal{L}_R = \lambda_{r1} \mathcal{L}_{R1} + \lambda_{r2} \mathcal{L}_{R2} \tag{7}$$

where λ_r is a weight to balance these two losses. Finally, the total loss for training the whole network is defined as:

$$\mathcal{L}_{tol} = \mathcal{L}_{gcn} + \mathcal{L}_{reg} + \lambda_r \mathcal{L}_R \tag{8}$$

4 Experimental Results

4.1 Datasets

Our method is trained on the training set of THuman2.0 [23] and tested on the testing set of THuman2.0 as well as LSP datasets [7]. The THuman2.0 dataset is composed of 526 high-resolution 3D scans of 526 subjects with various body shapes and poses. The data in THuman2.0 is randomly split into a training set and a testing set at a ratio of 4:1. For each 3D scan, we render it from 360 views and obtain 360 ⟨RGB image, 3D body mesh⟩ pairs. As a result, the training set is extended and contains 151,200 training data in total. The LSP [7] dataset consists of 2,000 in-the-wild images of sportsmen with difficult poses. Since LSP doesn't provide any ground-truth SMPL annotation, it is only used for qualitative evaluation in our experiments.

4.2 Implementation Details

In our implementation, ResNet-50 [4] is employed as the image encoder, the network architecture proposed in [10] is adopted to implement our GCN, the multiple layer perceptron is employed to construct the parameter regressor and the network of PaMIR [27] is adopted to reconstruct the surface of clothed humans. All our networks are implemented based on PyTorch [19]. Adam [9] is employed for optimizing our networks. In the whole training phase, the learning rate is fixed to 3×10^{-4}, and the batch size is set as 16. Our networks are totally trained for 20 epochs, and it takes about 5 days on a computer with a single NVIDIA GeForce RTX 3080 GPU.

4.3 Comparisons

(1) **Quantitative Comparisons**

Table 1. Quantitative comparisons on the THuman2.0 dataset.

Method	Publication	MPJPE	PA-MPJPE	MVPE
SPIN [11]	ICCV'2019	64.2	48.9	80.5
GCMR [12]	CVPR'2019	93.7	67.3	111.0
DecoMR [24]	CVPR'2020	112.5	84.5	126.1
PyMAF [25]	ICCV'2021	66.9	49.6	83.8
3DCrowdNet [3]	CVPR'2022	101.3	93.9	118.3
Ours	–	**42.9**	**34.6**	**46.5**

Same as other methods [11,12,25], we use three quantitative metrics (MPJPE, PA-MPJPE, and MPVE) to calculate the experimental results. Table 1 shows

the quantitative comparisons on THuman2.0. Our method achieves an MPJPE of 42.9 mm, a PA-MPJPE of 34.6 mm, and an MVPE of 46.5 mm, and outperforms all the other methods involved in the comparison. In terms of pose estimation, our method outperforms the second-best method (i.e., PyMAF [25]) by an MPJPE of 24.0 mm and a PA-MPJPE of 15.0 mm. In terms of shape estimation, our method also achieves the lowest MVPE of 37.3 mm. It can be seen that due to the influence of loose clothing obscuration, most advanced methods are unable to accurately identify not only the human postural motion under the clothing but also the human shape.

(2) **Qualitative Comparisons**

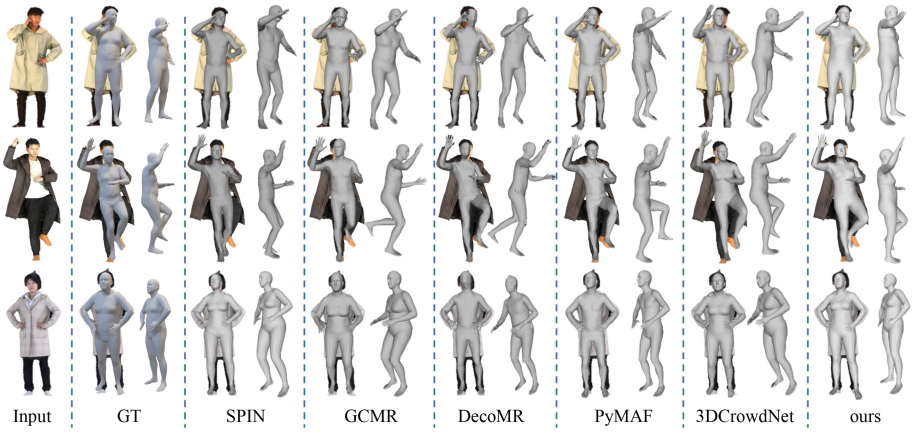

Fig. 3. Qualitative comparisons on the THuman 2.0 dataset.

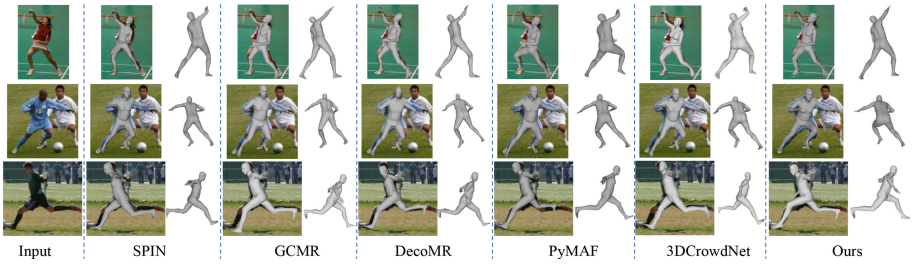

Fig. 4. Qualitative comparisons on the LSP dataset.

We compare our method qualitatively with SPIN [11], GCMR [12], DecoMR [24], PyMAF [25] and 3DCrowdNet [3]. We first test these methods on the testing set of THuman2.0 and then test them on LSP for cross-dataset evaluation. Figure 3

shows the results produced by these four methods and our methods on THuman2.0. It can be seen that our method produces the most accurate poses and body shapes that are similar to ground-truth models, whether examined from the visible view or invisible view. Figure 4 illustrates the cross-dataset results on the LSP datasets. Our method still outputs 3D body models with accurate poses and body shapes, while other methods do not estimate the postures of legs and arms precisely in most of their results.

4.4 Ablation Study

To verify that the accuracy of estimated 3D human poses and shapes can be boosted by clothed 3D human reconstruction, we train four different models (m1∼m4) of our method using different loss functions and test these models on the Thuman2.0 dataset. The quantitative results achieved by these models are shown in Table 2. Without all the two reconstruction loss functions (i.e., the m1 model), our method only yields an MPJPE of 65.6mm, a PA-MPJPE of 44.2 mm, and an MVPE of 65.1 mm. By adding the $\mathcal{L}_{R1}$ loss (i.e., the m2 model), the MPJPE, PA-MPJPE, and MVPE are respectively decreased by 12.7 mm, 9.2 mm, and 16.5 mm. By adding the $\mathcal{L}_{R2}$ loss (i.e., the m3 model), the MPJPE, PA-MPJPE, and MVPE are respectively decreased by 10.4 mm, 6.5 mm, and 12.7 mm. These results indicate that both the absolute reconstruction errors $\mathcal{L}_{R1}$ and the relative reconstruction loss $\mathcal{L}_{R2}$ are beneficial for the accuracy of estimated 3D human poses and shapes. When using both these two reconstruction losses, the MPJPE, PA-MPJPE, and MVPE respectively drop to 42.9 mm, 34.6 mm, and 46.5 mm.

Table 2. Comparisons of our method trained with different loss functions.

Models	Loss Functions	MPJPE	PA-MPJPE	MVPE
m1	$\mathcal{L}_{reg}$	65.6	44.2	65.1
m2	$\mathcal{L}_{reg} + \mathcal{L}_{R1}$	52.9	35.0	48.6
m3	$\mathcal{L}_{reg} + \mathcal{L}_{R2}$	55.2	37.7	52.4
m4	$\mathcal{L}_{reg} + \mathcal{L}_{R}(\mathcal{L}_{tol})$	**42.9**	**34.6**	**46.5**

5 Conclusion

Estimating 3D human pose and body shape from a single image is challenging. In this paper, we have proposed a 3D human pose and shape estimation method enhanced by clothed 3D human reconstruction. Two clothed 3D human models are respectively reconstructed under the guidance of the recovered 3D body mesh and the ground-truth 3D body mesh. The ablation study has validated that the accuracy of the estimated 3D human poses and shapes is significantly improved by our method. Experimental results show that our method achieves state-of-the-art performance.

References

1. 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: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9909, pp. 561–578. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46454-1_34
2. Choi, H., Moon, G., Lee, K.M.: Pose2Mesh: graph convolutional network for 3D human pose and mesh recovery from a 2D human pose. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12352, pp. 769–787. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58571-6_45
3. Choi, H., Moon, G., Park, J., Lee, K.M.: Learning to estimate robust 3D human mesh from in-the-wild crowded scenes. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022). https://doi.org/10.1109/CVPR52688.2022.00153
4. He, K., Zhang, X., Ren, S., Sun, J.: Identity mappings in deep residual networks. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9908, pp. 630–645. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46493-0_38
5. He, T., Xu, Y., Saito, S., Soatto, S., Tung, T.: ARCH++: animation-ready clothed human reconstruction revisited. In: IEEE/CVF International Conference on Computer Vision (ICCV), pp. 11046–11056 (2021). https://doi.org/10.1109/ICCV48922.2021.01086
6. Huang, Y., et al.: Towards accurate marker-less human shape and pose estimation over time. In: International Conference on 3D Vision (3DV), pp. 421–430 (2017). https://doi.org/10.1109/3DV.2017.00055
7. Johnson, S., Everingham, M.: Clustered pose and nonlinear appearance models for human pose estimation. In: British Machine Vision Conference (BMVC), vol. 2, p. 5 (2010). https://doi.org/10.5244/C.24.12
8. Kanazawa, A., Black, M.J., Jacobs, D.W., Malik, J.: End-to-end recovery of human shape and pose. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7122–7131 (2018). https://doi.org/10.1109/CVPR.2018.00744
9. Kingma, D., Ba, J.: Adam: a method for stochastic optimization. In: International Conference on Learning Representations (ICLR), pp. 1–15 (2015)
10. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. In: International Conference on Learning Representations (2016)
11. Kolotouros, N., Pavlakos, G., Black, M.J., Daniilidis, K.: Learning to reconstruct 3D human pose and shape via model-fitting in the loop. In: IEEE/CVF International Conference on Computer Vision (ICCV), pp. 2252–2261 (2019). https://doi.org/10.1109/ICCV.2019.00234
12. Kolotouros, N., Pavlakos, G., Daniilidis, K.: Convolutional mesh regression for single-image human shape reconstruction. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4501–4510 (2019). https://doi.org/10.1109/CVPR.2019.00463
13. Lassner, C., Romero, J., Kiefel, M., Bogo, F., Black, M.J., Gehler, P.V.: Unite the people: closing the loop between 3D and 2D human representations. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6050–6059 (2017). https://doi.org/10.1109/CVPR.2017.500
14. Li, Y., Cai, J., Zhou, Q., Lu, H.: Joint semantic-instance segmentation method for intelligent transportation system. IEEE Trans. Intell. Transp. Syst. 1–8 (2022). https://doi.org/10.1109/TITS.2022.3190369

15. Liu, L., Sun, J., Gao, Y., Chen, J.: HEI-human: a hybrid explicit and implicit method for single-view 3D clothed human reconstruction. In: Ma, H., et al. (eds.) PRCV 2021. LNCS, vol. 13020, pp. 251–262. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-88007-1_21
16. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: SMPL: a skinned multi-person linear model. ACM Trans. Graph. **34**(6), 1–16 (2015). https://doi.org/10.1145/2816795.2818013
17. Lorensen, W.E., Cline, H.E.: Marching cubes: a high resolution 3D surface construction algorithm. In: Conference on Computer Graphics and Interactive Techniques, pp. 163–169 (1987). https://doi.org/10.1145/37401.37422
18. 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: International Conference on 3D Vision (3DV), pp. 484–494 (2018). https://doi.org/10.1109/3DV.2018.00062
19. Paszke, A., et al.: PyTorch: an imperative style, high-performance deep learning library. In: International Conference on Neural Information Processing Systems (NIPS) (2019)
20. Saito, S., Huang, Z., Natsume, R., Morishima, S., Kanazawa, A., Li, H.: PIFu: pixel-aligned implicit function for high-resolution clothed human digitization. In: IEEE/CVF International Conference on Computer Vision (ICCV), pp. 2304–2314 (2019). https://doi.org/10.1109/ICCV.2019.00239
21. Xiang, D., Joo, H., Sheikh, Y.: Monocular total capture: posing face, body, and hands in the wild. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10965–10974 (2019). https://doi.org/10.1109/CVPR.2019.01122
22. Xiu, Y., Yang, J., Tzionas, D., Black, M.J.: ICON: implicit clothed humans obtained from normals. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 13296–13306 (2022). https://doi.org/10.1109/TPAMI.2021.3050505
23. Yu, T., Zheng, Z., Guo, K., Liu, P., Dai, Q., Liu, Y.: Function4D: real-time human volumetric capture from very sparse consumer RGBD sensors. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5746–5756 (2021). https://doi.org/10.1109/CVPR46437.2021.00569
24. Zeng, W., Ouyang, W., Luo, P., Liu, W., Wang, X.: 3D human mesh regression with dense correspondence. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7054–7063 (2020). https://doi.org/10.1109/CVPR42600.2020.00708
25. Zhang, H., et al.: PyMAF: 3D human pose and shape regression with pyramidal mesh alignment feedback loop. In: IEEE/CVF International Conference on Computer Vision (ICCV) (2021). https://doi.org/10.1109/ICCV48922.2021.01125
26. Zheng, Y., Li, Y., Yang, S., Lu, H.: Global-PBNet: a novel point cloud registration for autonomous driving. IEEE Trans. Intell. Transp. Syst. **23**(11), 22312–22319 (2022). https://doi.org/10.1109/TITS.2022.3153133
27. Zheng, Z., Yu, T., Liu, Y., Dai, Q.: PaMIR: parametric model-conditioned implicit representation for image-based human reconstruction. IEEE Tans. Pattern Anal. Mach. Intell. (TPAMI) **44**(6), 3170–3184 (2021)
28. Zheng, Z., Yu, T., Wei, Y., Dai, Q., Liu, Y.: DeepHuman: 3D human reconstruction from a single image. In: IEEE/CVF International Conference on Computer Vision (ICCV), pp. 7739–7749 (2019). https://doi.org/10.1109/ICCV.2019.00783