3D Human Body Shape and Pose Estimation from Depth Image

Lei Liu, Kangkan Wang(✉), and Jian Yang

PCA Lab, Key Lab of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education, and Jiangsu Key Lab of Image and Video Understanding for Social Security, School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China
{liuleijs,wangkangkan,csjyang}@njust.edu.cn

Abstract. This work addresses the problem of 3D human body shape and pose estimation from a single depth image. Most 3D human pose estimation methods based on deep learning utilize RGB images instead of depth images. Traditional optimization-based methods using depth images aim to establish point correspondences between the depth images and the template model. In this paper, we propose a novel method to estimate the 3D pose and shape of a human body from depth images. Specifically, based on the joints features and original depth features, we propose a spatial attention feature extractor to capture spatial local features of depth images and 3D joints by learning dynamic weights of the features. In addition, we generalize our method to real depth data through a weakly-supervised method. We conduct extensive experiments on SURREAL, Human3.6M, DFAUST, and real depth images of human bodies. The experimental results demonstrate that our 3D human pose estimation method can yield good performance.

Keywords: Human shape and pose estimation · Deep learning · Weak supervision

1 Introduction

Human pose estimation has numerous applications in robotics, augmented reality (AR), and virtual reality (VR). With the rapid development of computer vision, 3D human shape and pose estimation from depth images has gained popularity in the 3D computer vision community. However, estimating 3D human models directly from depth images is still a challenging problem since the human bodies have large deformations and self-occlusions in motion.

Recently, a few methods use deep learning to directly predict 3D human models from depth images. Most 3D human body reconstruction methods based on a sequence of depth images aim to establish the point correspondences between

This work was supported by the Natural Science Foundation of China under Grant Nos. 61602444, U1713208, and Program for Changjiang Scholars.

Y. Peng et al. (Eds.): PRCV 2020, LNCS 12305, pp. 410–421, 2020.
https://doi.org/10.1007/978-3-030-60633-6_34

the consecutive frames, which results in the error accumulation. At the same time, most human pose estimation methods use RGB images as the inputs. However, the RGB images lack depth information naturally, which makes it a classical ill-posed inverse problem. In addition, directly extending these methods to depth images cannot obtain high-precision 3D human models. Another problem with 3D human pose estimation is the lack of a large number of labeled 3D human pose datasets, and labeling 3D human pose datasets is very difficult, which results in great difficulty for the training of deep neural networks.

In this paper, we propose a 3D human body shape and pose estimation method from a single depth image. First, we employ ResNet50 [1] to extract the latent features of depth images. Then, we simultaneously estimate the 3D joints and the 3D human model parameters. Besides, we introduce the attention mechanism to improve the accuracy of 3D human pose estimation. Based on the 3D joints features and the original depth features, we construct a spatial attention feature extractor to learn the dynamical attention weights, which can capture local geometric structure and assess the feature map actively. Similar to [2], we use a cyclic regression network to regress the human model parameters from features. Further, to improve the performance of the network on real depth data, we introduce a weakly-supervised mechanism to fine-tune the network. In addition to using traditional joints information, we also introduce a differentiable rendering layer to render the predicted human models to depth images and silhouettes. Intuitively, the rendered depth images and silhouettes should be consistent with the original inputs. The experimental results on SURREAL [3], Human3.6M [4], DFAUST [5], and the real data of human bodies demonstrate the effectiveness of our proposed method. In summary, the main contributions of our method are as follows:

1. We propose a novel 3D human pose and shape estimation framework based on depth images to predict the 3D human model parameters, which achieves the state-of-the-art performance on human pose and shape recovery.
2. We propose a spatial attention feature extractor for extracting more effective features, which effectively improves the accuracy of the human shape and pose estimation.
3. We fine-tune the network by using joints information and a differentiable render layer in weakly-supervised manner, which improves the performance of the network on real depth data.

2 Related Work

2.1 3D Human Model Estimation from RGB Images

With the simplicity and extensibility of the SMPL model [6], there has been substantial recent work in estimating the parameters of this model. Bogo et al. [7] propose SMPLify, which estimates the position of the corresponding 2D human joints. Then, they recover SMPL parameters by minimizing the 2D projection of model joints and the detected 2D joints. Huang et al. [8] expand the SMPLify

and propose MuVS. MuVS uses multi-view images as inputs and a deep neural network is used to segment the human body, which can eliminate the effects of background pictures. In addition to using joints projection as a constraint, this method also matches the 3D human models with the segmentation results to improve the accuracy of the results. Kanazawa et al. [2] propose an end-to-end framework HMR for recovering a 3D human model from an image, which directly extracts features from the images and recovers the SMPL pose parameters through regression. In addition to directly regressing SMPL model parameters, some recent works also use other methods to predict 3D human models. For example, Kolotouros et al. [9] propose a Graph CNN method, which first attaches the encoded features from an input RGB image to 3D vertex coordinates of a template mesh and then predicts the mesh vertex coordinates using a convolutional mesh regression.

2.2 3D Human Model Estimation from Depth Images

In recent years, there have been some studies on human pose estimation based on depth images. Guo et al. [10] use a L_0 based motion regularizer with an iterative optimization solver to deform the pre-scanned template model to each input depth images. There are some template-less methods which can create the 3D human models without any prior knowledge about the human shape and fuse all depth maps to reconstruct 3D human models with slow motion. These methods aim to build point correspondences for each depth frame by searching the closest 3D point, which will be invalid when the input depth image is very different from the template model. Wei et al. [10] build the point correspondences by matching the learned feature descriptors for depth images of human bodies. Then, the 3D models are generated by fitting the template SMPL model to learned point correspondences using [10]. Kadkhodamohammadi et al. [11] propose a multi-view RGB-D approach for human pose estimation, which proves the advantages of using depth data. Pavlakos et al. [12] use ordinal depth constraints as a form of weak-supervision to train a network which can predict the 3D pose. Li et al. [13] propose a dynamic fusion module that enables training models with RGB-D data to address the ambiguity problem.

3 Approach

For a given depth image, our method can be used to estimate the 3D human models aligned with the depth images. The algorithm flow is shown in Fig. 1. First of all, we use ResNet50 [1] to extract features from depth images. Then, we use a joints estimation network to learn the human pose information. We propose a spatial attention feature extractor to learn the refined features. In this way, the network could suppress irrelevant regions and highlights useful features for the human pose and shape estimation. Therefore, the accuracy of 3D human pose and shape estimation can be effectively improved. In addition, we also introduce a differentiable rendering layer [14], which can render the predicted human

models to silhouettes and depth images. We exploit joints projection, silhouettes, and depth images as constraints to keep our prediction models consistent with the input information. By this means, we could fine-tune the network through a weakly-supervised method and improve its performance on the real data. Therefore, our method does not require real labeled 3D human pose datasets.

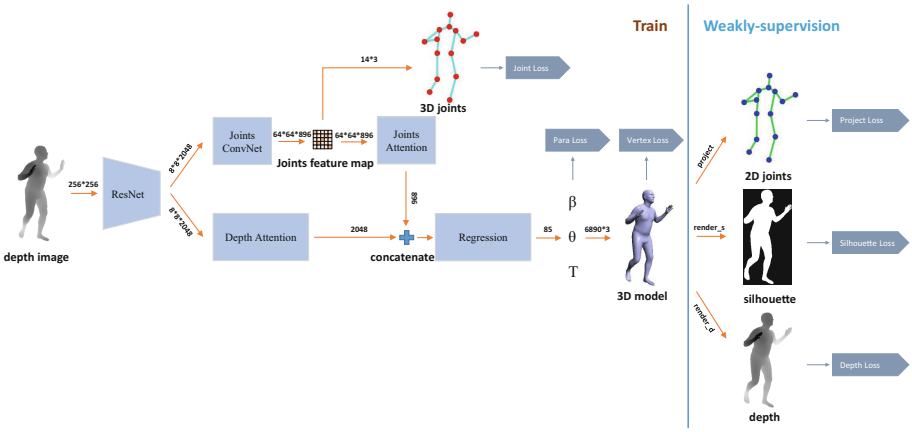

Fig. 1. Overview of the proposed framework. Our framework can predict the 3D human shape and pose from a single depth image. The joints prediction module is used to explicitly extract pose information from the features. The depth attention and joints attention module are used to extract more effective depth features and pose features. We exploit a weakly-supervised method to fine-tune our network on the real depth data. The numbers in the figure show the size of the input of each module.

3.1 3D Body Model

We follow the previous 3D human pose estimation methods, using the Skinned Multi-Person Linear (SMPL) model of Loper et al. [6] to represent the human body. The SMPL model is a statistical parametric differentiable model that can accurately represent various human shapes under natural conditions, and it is compatible with existing graphics pipelines. It uses shape parameter $\beta \in \mathbb{R}^{10}$ to control the appearance of the human body model. Pose parameter $\theta \in \mathbb{R}^{72}$ are used to represent different poses of the human body. For the given shape parameter β and pose parameter θ, the SMPL model provides a function $M(\beta, \theta)$, which can map the shape and pose parameters to 6890 vertexes V of the human body. At the same time, SMPL also provides a $6890 * 24$ joints regression coefficient R, which can map the 3D human model to the corresponding 24 joints through $X(V, R)$.

3.2 3D Human Joints Estimation

As a simplified representation of human pose, 3D joints contain a large amount of pose information. In this paper, the 3D joints estimation module is used to explicitly extract human pose information, which is beneficial to the convergence of the network. We use the same method as [15], which relates and unifies the heat map representation and joints regression with a simple operation. We use the "taking-expectation" to take place the "taking-maximum", which called *integral regression*. The integral function is differentiable and allows end-to-end training. The joints location is estimated as the integration of all locations in the heat map, weighted by their probabilities as:

$$J_k = \sum_{p_z=1}^{D} \sum_{p_y=1}^{H} \sum_{p_x=1}^{W} p * \widetilde{H}_k(p). \tag{1}$$

Here, p is the pixel position in the heat map, $\widetilde{H}_k$ is the normalized heat map. In this way, the heat map can be transformed as joints position by the integral function. We use the L_1 loss as the joints loss:

$$L_{joints} = \sum_i \left\| x_i - \hat{x}_i \right\|_1 . \tag{2}$$

3.3 Spatial Attention Feature Extracting

We introduce the attention mechanism to learn dynamic weights for the 3D joints features and original depth features which improve the effectiveness and generalization of the features and highlights the useful features. As far as we know, this is the first time that the attention mechanism been used in 3D human pose and shape estimation task. The architecture of joints attention and depth attention is shown in Fig. 2. The network builds relationships between spatial regions that are far from each other by providing a dynamic weighted features. Moreover, the network with attention layer can autonomously assess the effectiveness of the feature on the human pose estimation and adjust the weights of different features dynamically. Specifically, the depth attention is used to capture the local structured information in depth features while the joints attention is used to capture the pose information in joints features. We define the operation as: $\psi = W * x$, where W is the weight matrix learned by a MLP. Then, we get the attention map ω_{ij} by computing the softmax of ψ:

$$\omega_{ij} = \frac{exp(\psi_{ij})}{\sum_{i=1}^{N} \psi_{ij}}. \tag{3}$$

To obtain the final attention feature map, we apply the matrix multiplication between the attention map ω_{ij} and the original feature x_j,

$$\hat{x}_i = \sum_{i=1}^{N} \omega_{ij} x_j. \tag{4}$$

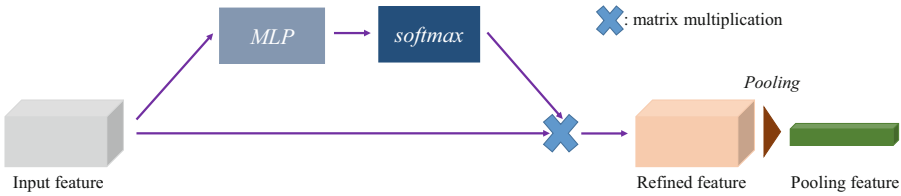

Fig. 2. An illustration of our attention architecture. The depth attention and joints have similar architecture. The input feature map is passed to the attention module which learns the dynamic weights ω. We apply the matrix multiplication between the dynamic weights and original features to obtain the refined features. Finally, the refined features are aggregated by max pooling and average pooling for joints attention and depth attention, respectively.

To obtain their respective global feature descriptors, the new joints features and depth features obtained from the attention layer are aggregated using max pooling and average pooling, respectively.

3.4 Regression the Parameters of Human Model

For the human model parameters regression module, we use a deep neural network, whose architecture has the same design with Kanazawa et al. [2]. We concatenate the joints feature descriptors, depth feature descriptors, and the mean model parameters together as the inputs. Then, we regress the parameters β, θ, and T in an iterative error feedback (IEF) loop, where progressive changes are made recurrently to the current estimation. After that, we map the parameters into 6890 vertexes of 3D human body with the function $M(\beta, \theta)$ provided by SMPL. In this setting, the parameter loss and vertex loss are provided:

$$L_{para} = \left\| [\beta, \theta] - \left[\hat{\beta}, \hat{\theta}\right] \right\|_2^2, \tag{5}$$

$$L_{vertex} = \left\| v_i - \hat{v}_i \right\|_2^2. \tag{6}$$

In summary, the loss function of the entire network is

$$L = \varphi L_{joints} + \gamma L_{para} + \sigma L_{vertex}. \tag{7}$$

Among them, φ, γ, and σ are the hyperparameters of the corresponding loss functions, which are used to balance the value of each loss function.

3.5 Fine-Tuning for Real Depth Data with Weakly-Supervision

Since the training data used in this paper comes from synthetic data and does not contain any real data, the performance of the network on real data is not satisfactory. Therefore, we fine-tune the network on real depth data in a weakly-supervised manner. Similar to other weakly-supervised methods, we use 2D joints

projection to provide weakly-supervised constraints for the network. For the estimated 3D human model, the joints regression coefficients provided by SMPL can be used to calculate the corresponding 3D human joints, whose 2D projection should be consistent with the 2D joints of the depth image. Therefore, we define the joints projection loss function as:

$$L_{project} = \sum_i \left\| \vartheta_i - \hat{\vartheta}_i \right\|_1 . \tag{8}$$

Here, ϑ_i is the 2D projection of predicted 3D human joints, $\hat{\vartheta}_i$ is the ground truth joints of the depth image. However, the joints projection can only constrain the pose of the human body without considering the shape of the 3D human body. Because of that, we introduce a differentiable rendering layer, which can render the 3D models to 2D images. Similar to the joints projection, the 2D rendered silhouettes of the human models should be consistent with the silhouettes of the input depth images. The silhouette loss function is defined as

$$L_{sil} = \| p_{sil}(M(\beta, \theta) + T) - \hat{s} \|_2^2 . \tag{9}$$

Here, p_{sil} is the differentiable rendering silhouette function, $\hat{s}$ is the silhouette corresponding to the input depth image. Whether the joints projection loss or the silhouette loss are essentially 2D constraints on the 3D human models. As a result, there may still be multiple 3D human body models that match the 2D information. Therefore, in addition to the above two constraints, we introduce a depth loss to constrain the network with 3D information. We use the differentiable rendering layer mentioned above to render the 3D human models to depth images. Intuitively, the rendered depth images should be consistent with the input depth images, thus the depth loss of this paper is defined as

$$L_{depth} = \left\| p_d(M(\beta, \theta) + T) - \hat{d} \right\|_2^2 . \tag{10}$$

Here, p_d is the differentiable rendering depth function, $\hat{d}$ is the input depth image. The depth loss function can be used to constrain the depth of the network output model, so that the model is aligned with the input depth image.

Above all, the weakly-supervised loss is

$$L = \lambda L_{project} + \eta L_{sil} + \mu L_{depth}. \tag{11}$$

Among them, λ, η, and μ are the hyperparameters of the corresponding loss functions, which are used to balance the value of each loss function.

4 Empirical Evaluation

4.1 Datasets

Here is a brief description of the training data and test data used in this paper. We conduct extensive experiments on SURREAL [3], Human3.6M [4],

DFAUST [5], and real human depth images. SURREAL is a large synthetic human dataset. The dataset contains more than 55,000 sets of 3D human models, each of which contains 100 frames of different actions. We uniformly sample 200,000 human models from this data for training. The DFAUST dataset contains scan data of more than 40,000 real human bodies and the corresponding SMPL model. We uniformly sample 10,000 human body models from this data. Human3.6M is a RGB indoor dataset, we obtain ground truth SMPL parameters for the training images using MoSh [16] from the raw 3D MoCap markers. We also sample 10,000 human models from it. For all the sampled human models, we render them to depth images and obtain the corresponding 3D joints.

4.2 Implementation Details and Error Metrics

Implementation Details. We use the $256 * 256$ depth images as inputs. The ResNet50 is used for feature extraction. The 3D joints estimation module consists of 3 deconvolution layers with filter $4 * 4$ (stride 2), which channel number is fixed to 256, followed by the $1 * 1$ deconvolution layer. The architectures of joints attention module and the depth attention layer are similar, the set of MLP is (32, 32, 896/2048), the hyperparameters used in the network are set to $\varphi = 1$, $\gamma = 60$, $\sigma = 60$, $\lambda = 100$, $\eta = 10$, and $\mu = 1 * 10^{-4}$. We use Adam [17] optimizer with batch size of 32 and the learning rate is $1 * 10^{-4}$. We train and test our model on a single NVIDIA GTX2080Ti GPU, and the total number of iterations is set to 200.

Error Metrics. We conduct a quantitative and qualitative evaluation of the network. We use the Mean Average Vertex Error (MAVE) [18] over all vertexes of the recovered 3D human models in millimeter (mm) to quantify the reconstruction error:

$$\varepsilon = \frac{1}{N} \sum_{i=1}^{N} \sqrt{\|v_i - \hat{v}_i\|_2^2}. \tag{12}$$

Here, N is the number of vertexes of the 3D model, v is the vertex on the predicted 3d human models, and $\hat{v}$ is the vertex of the corresponding ground truth value.

4.3 Experiment Results

Quantitative Results. In order to verify the performance of the method proposed in this paper, we compare it with other methods on the same test set. Bogo et al. [7] propose SMPLify that first detects 2D joints on the image, and then matches the SMPL model's joints projection to the detected 2D joint points. Wei et al. [10] obtain the correspondences by matching the feature descriptors of the depth image and the template 3D human model. After that, they deform the SMPL models according to the correspondences. We deform the estimated models of [10] and [7] to input depth images by searching the nearest points. The experimental results are shown in Table 1.

Table 1. Reconstruction errors (mm) with different methods on SURREAL, Human3.6M, and DFAUST datasets. Results on the top part are used for comparison, while the results of ablation study about weak-supervision are at the bottom of this table.

Methods	SURREAL	Human3.6M	DFAUST
Bogo et al. [7]	75.3	87.2	91.5
Wei et al. [10]	68.3	80.2	66.5
Kolotouros et al. [19]	59.5	65.1	63.7
Kanazawa et al. [2]	50.1	57.6	56.3
Kolotouros et al. [9]	48.1	52.4	51.8
Zhu et al. [20]	46.8	50.3	49.2
Our method (no joints attention)	27.5	29.2	28.9
Our method (no depth attention)	29.4	31.8	30.1
Our method (no depth or joints attention)	38.8	40.0	39.5
Our method	22.4	24.9	23.6

In addition, We also compare our method with the methods of Kanazawa et al. [2], Kolotouros et al. [9], Kolotouros et al. [19], and Zhu et al. [20]. Since these methods are based on RGB images, we re-train them use depth images to make the comparison more fair. The comparison results are shown in Table 1. Besides, we also provide the visualization results of all the above methods on SURREAL in Fig. 3, all results are presented in the form of heat map.

Qualitative Results. We also evaluate our method qualitatively on real depth data. Among them, "Kungfu" from [21], "crane" from [22] "BUFF" from [23], and "Taiji" is collected in the laboratory with Kinect v2. Because our training data does not contain any real depth data, so we fine-tune our network in a

Fig. 3. The visualization of reconstruction accuracy using different methods on the SURREAL data. (a) The input scan. (b) our method. (c) Zhu et al. [20]. (d) Kolotouros et al. [9]. (e) Kanazawa et al. [2]. (f) Kolotouros et al. [19]. (g) Wei et al. [10]. (h) Bogo et al. [7].

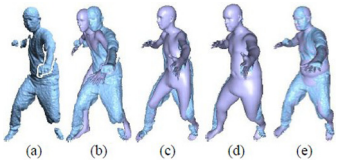

Fig. 4. An example of weakly-supervised fine-tuning on "Kungfu" data [21]. (a) The input scan. (b) The results before fine-tuning. (c, d, e) The results after weakly-supervised fine-tuning with $L_{project}$, $L_{project} + L_{sil}$, and $L_{project} + L_{sil} + L_{depth}$, respectively.

weakly-supervised way to improve the performance of the network on real depth data. The sampled results of "Kungfu" before and after fine-tuning are shown in the Fig. 4. One can see that the network can output a 3D human model aligned with the input depth image after fine-tuning the network. More results using our method are shown in Fig. 5. From the experimental results, one can see that our method can deal well with the occlusion and random noise of real depth data, and generate corresponding 3D human models.

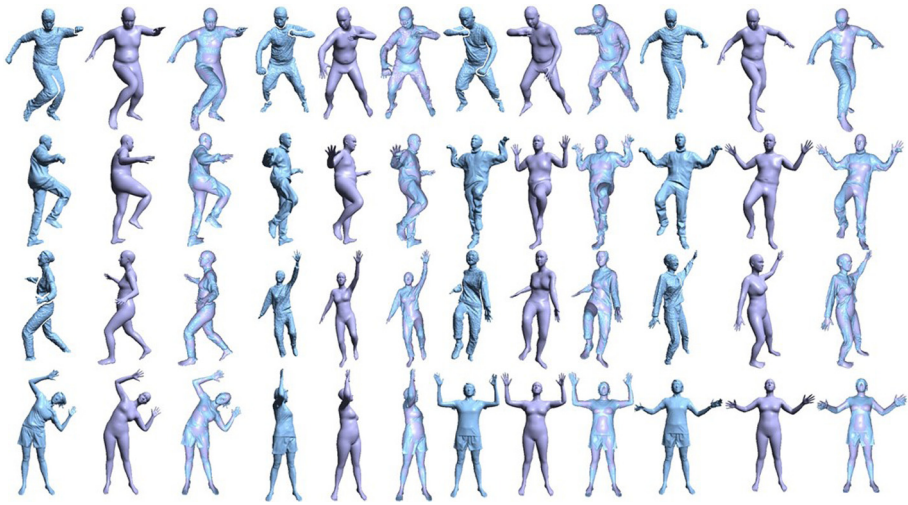

Fig. 5. Some recovered 3D models using our method on real data. For each result, we show the extracted raw depth scan, the reconstructed model, and the overlay with alignment between the reconstructed model and the raw depth scan. From top to bottom: "Kungfu" from [21], "crane" from [22], "Taiji" from [23], and "BUFF".

4.4 Ablation Study

Spatial Attention Layers. We first evaluate the effectiveness of our spatial attention feature layers. In Table 1, we provide the results for four different settings of our approach, one where the network is trained with all the spatial attention layers, a second where the network is trained without the module of depth attention, a third where the network is trained without the module of joints attention, a fourth where the network is trained without the both attention module. As we can see from the Table 1, the error of the network which uses all the spatial attention layers is better than the network without the module. It shows that the spatial attention feature extracting module can extract the spatial local features from the original depth features and 3D joints features, thus leading to higher recovery accuracy of 3D human pose and shape estimation.

Weakly-Supervision for the Real Depth Data. We also evaluate the effectiveness of our weakly-supervised method on real depth data, we compare the alignment results of the 3D human model and the original input depth image before and after weakly-supervised fine-tuning the network with different weakly-supervised loss functions. We provide the results of "Kungfu" in Fig. 4. As shown in Fig. 4, the alignment results of the predicted 3D human model are not satisfactory because the training data does not contain any real labeled data. The $L_{project}$ or $L_{project} + L_{sil}$ can only supervise the network with 2D information. The network can generate more accurate 3D models that have consistent shape and pose with the input real depth images after introducing L_{depth}.

5 Conclusion

In this paper, we propose a novel 3D human pose and shape estimation method from a single depth image, which achieves the state-of-the-art performance on human pose and shape recovery. In order to capture the effective information in joints features and original depth features, we propose a spatial attention layer, which builds relationships between spatial regions and extracts the refined features. Besides, we introduce weakly-supervised mechanism to improve the performance of the proposed method on real data. In addition to the traditional joints projection, we also introduce a differentiable rendering layer, which can render the 3D models to depths and silhouettes. In this way, we can fine-tune the network in both 2D and 3D constraints. The experimental results on SURREAL, Human3.6M, DFAUST, and the real data of human bodies show the effectiveness of our method.

References

1. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
2. Kanazawa, A., Black, M.J., Jacobs, D.W., Malik, J.: End-to-end recovery of human shape and pose. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7122–7131, June 2018
3. Varol, G., et al.: Learning from synthetic humans. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 109–117, July 2017
4. Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: Human3.6M: large scale datasets and predictive methods for 3D human sensing in natural environments. IEEE Trans. Pattern Anal. Mach. Intell. **36**(7), 1325–1339 (2014)
5. Bogo, F., Romero, J., Pons-Moll, G., Black, M.J.: Dynamic faust: registering human bodies in motion. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6233–6242, July 2017
6. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: SMPL: a skinned multi-person linear model. ACM Trans. Graph. **34**, 1–16 (2015)
7. 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

8. Huang, Y., et al.: Towards accurate marker-less human shape and pose estimation over time. In: 2017 International Conference on 3D Vision (3DV), pp. 421–430. IEEE (2017)
9. Kolotouros, N., Pavlakos, G., Daniilidis, K.: Convolutional mesh regression for single-image human shape reconstruction. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4501–4510 (2019)
10. Wei, L., Huang, Q., Ceylan, D., Vouga, E., Li, H.: Dense human body correspondences using convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1544–1553 (2016)
11. Kadkhodamohammadi, A., Gangi, A., de Mathelin, M., Padoy, N.: A multi-view RGB-D approach for human pose estimation in operating rooms. In: 2017 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 363–372. IEEE (2017)
12. Pavlakos, G., Zhou, X., Daniilidis, K.: Ordinal depth supervision for 3D human pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7307–7316 (2018)
13. Li, R., Cai, C., Georgakis, G., Karanam, S., Chen, T., Wu, Z.: Towards robust RGB-D human mesh recovery. arXiv preprint arXiv:1911.07383 (2019)
14. Kato, H., Ushiku, Y., Harada, T.: Neural 3D mesh renderer. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3907–3916 (2018)
15. Sun, X., Xiao, B., Wei, F., Liang, S., Wei, Y.: Integral human pose regression. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11210, pp. 536–553. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01231-1_33
16. Loper, M., Mahmood, N., Black, M.J.: MoSh: motion and shape capture from sparse markers. ACM Trans. Graph. (TOG) **33**(6), 1–13 (2014)
17. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
18. Yu, R., Saito, S., Li, H., Ceylan, D., Li, H.: Learning dense facial correspondences in unconstrained images. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4723–4732 (2017)
19. Kolotouros, N., Pavlakos, G., Black, M.J., Daniilidis, K.: Learning to reconstruct 3D human pose and shape via model-fitting in the loop. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2252–2261 (2019)
20. Zhu, H., Zuo, X., Wang, S., Cao, X., Yang, R.: Detailed human shape estimation from a single image by hierarchical mesh deformation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4491–4500 (2019)
21. Guo, K., Xu, F., Wang, Y., Liu, Y., Dai, Q.: Robust non-rigid motion tracking and surface reconstruction using L0 regularization. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3083–3091 (2015)
22. Vlasic, D., Baran, I., Matusik, W., Popović, J.: Articulated mesh animation from multi-view silhouettes. In: ACM SIGGRAPH 2008 papers, pp. 1–9 (2008)
23. Zhang, C., Pujades, S., Black, M.J., Pons-Moll, G.: Detailed, accurate, human shape estimation from clothed 3D scan sequences. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4191–4200