3D BODY POSE AND SHAPE ESTIMATION FROM MULTI-VIEW IMAGES WITH LIMB GEOMETRIC CONSTRAINT

Zixuan Gai, Xu Zhao, Xin Cao*

Department of Automation, Shanghai Jiao Tong University

ABSTRACT

The estimation of 3D body pose and shape has always been a challenging problem due to various reasons, such as the ambiguity in 2D images and complex articulated structure of the human body. In order to solve the ill-conditioned problems, in this paper, we bring up an end-to-end method to estimate 3D human shape and pose from multi-view RGB images. In the proposed framework, we first implement a CNN embedded with attention module to extract the image feature and design the view-pooling layer to combine the features from multiple views. Then we adopt a regression network with a novel geometric constraint of body limbs to estimate 3D human pose and shape. Additionally, during the training process, we employ the idea of adversarial learning in our model to help regress accurate pose and shape parameters. Extensive experiments are conducted on Human3.6M and MPI-INF-3DHP datasets, and our method achieves competitive results in the 3D pose and shape estimation task.

Index Terms— 3D Pose and Shape Estimation, SMPL, Attention Mechanism, Adversarial Learning, CNN

1. INTRODUCTION

The estimation of 3D body pose and shape from RGB images is an important task in computer vision. It is a core technique in various applications, such as computer animation, orthopedic diagnosis, activity surveillance, virtual reality and so forth.

However, this task is difficult due to the inherent ambiguity when estimating 3D information from 2D clues. In order to erase the ambiguity, researchers have used image sequences [1, 2], employed multi-view images [3] or adopted additional information using depth cameras [4] like Kinect as inputs. Most 3D pose and shape estimation methods employ SMPL [5] or SCAPE [6] as basic parametric model, and adopt RGB images [7, 8, 9, 10, 11], depth images [4, 12], or silhouettes [13] as inputs, and outputs 3D meshes. These methods could be divided into two categories roughly: one is traditional optimization-based, which often takes RGB images as inputs and fits SMPL to the detected 2D keyjoints along with various priori assumptions about the joint angles and shape constraints, like [7] and [8]. However, the fitting process of the optimization-based method is time-consuming due to the slow convergence, requiring about several minutes per image to estimate the 3D model; the other is deep-learning based. Since convolutional neural network is widely used recently because of its high efficiency and accuracy, some works have implemented CNN to estimate 3D human pose and shape. Pavlakos *et al*.[10] use hourglass model to predict 2D joints and silhouette from input image and design two CNNs to regress pose and shape parameters. Similarly, Omran *et al*.[11] first employ a CNN to generate the segmentation of the human in the image, then use this segmentation to estimate SMPL related parameters. However, the process of lifting the processed 2D input to 3D estimation may lose plenty of useful information.

In order to solve the aforementioned problems, namely, the low time efficiency and the loss of useful information when encoding the images to estimate 3D body pose and shape, in this paper, we propose a novel approach to estimate 3D human pose and shape from multi-view images. Our method mainly includes three contributions.

First, we propose an encoder with attention mechanism to extract image feature and design a view-pooling layer to combine the features from multiple views. This could harvest elaborate information from multi-view images. In the mean time, the attention module helps the encoder to extract specific image feature, which is essential for predicting accurate human pose and shape parameters.

Second, we propose a 3D regression module to estimate shape and pose parameters from combined image feature, during which we design a novel descriptor of the pose structure. This module explores the constraints of the inner correlations of body and the original 3D space, therefore could achieve satisfying results.

In addition, we implement a discriminator in the training process to act as pose and shape priors. This could boost the performance of our framework to generate accurate 3D human model in a generative-adversarial way.

Our model is trained end-to-end and involves no optimization during test time, which runs in real-time and achieves

* Corresponding author. This research has been supported by the National Key Research and Development Program of China (Grant No. 2017YFC0806501) and NSFC Program (61673269, 61273285).

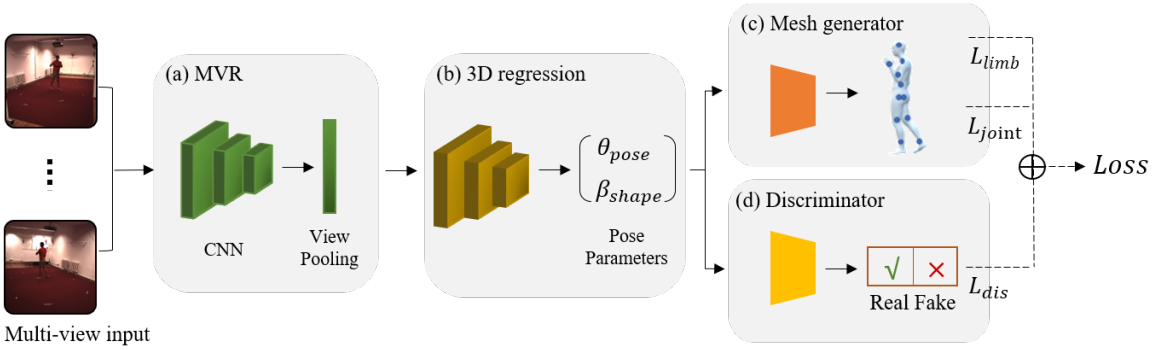

Fig. 1. **Overview of our framework.** Each image in the multi-view input is passed through the (a) MVR (Multi-view ResNet) separately, aggregated at the view-pooling layer, then sent to the (b) 3D regression module to estimate pose θ and shape β. Then we use (c) mesh generator to build 3D human model and get the predicted joints. We calculate losses on the geometric constraints $\mathbf{L}_{limb}$ and $\mathbf{L}_{joint}$ which are back propagated through the entire model. The (d) discriminator acts as the prior, forcing the model to generate accurate 3D parameters.

state-of-the-art performance on benchmark datasets.

2. METHOD

2.1. SMPL Model

Our method builds 3D human model through a skinned vertex based model, SMPL [5]. SMPL can be described as a function $\mathbf{M}(\theta, \beta)$ that takes pose θ and shape β as inputs and produces a weighted triangulated mesh with 6890 vertices and 13776 triangles. The pose $\theta \in \mathbb{R}^{3K+1}$ is the axis angle representation of the relative 3D rotation of the $K = 23$ joints and one global rotation of the human. The shape $\beta \in \mathbb{R}^{10}$ is the first 10 PCA coefficients of the shape space, learned from thousands of registered scans. In order to build a SMPL model, first we apply shape and pose dependent deformations $\mathcal{B}_s(\beta)$ and $\mathcal{B}_p(\theta)$ to the template $\overline{\mathbf{T}}$ with zero pose and shape; then we use the LBS skinning function W to build the final triangulated surface with the deformed vertices $\mathbf{T}$, keyjoints $\mathcal{J}(\beta)$, pose θ and the skinning weights $\mathcal{W}$

$$\mathbf{T}(\beta, \theta) = \overline{\mathbf{T}} + \mathcal{B}_s(\beta) + \mathcal{B}_p(\theta), \tag{1}$$

$$\mathbf{M}(\beta, \theta) = \mathrm{W}(\mathbf{T}(\beta, \theta), \mathcal{J}(\beta), \theta, \mathcal{W}). \tag{2}$$

Through the above functions we could easily generate person-specific 3D mesh given the corresponding pose and shape. Also SMPL is fully differentiable, therefore we can employ SMPL as part of our deep learning framework and regard this mesh generator as a neural network with fixed parameters.

2.2. Multi-view ResNet with Attention Module

In this section, we explain how to encode image features from multi-view images. As illustrated in Fig.1, the input of this module is a series of multi-view images. We design a multi-view CNN based on ResNet, abbreviated as MVR, on top of the 3D regression module in section 2.3. In order to exclude the interference of the background and make the model focus on human-related areas in the image, inspired by [14], we adopt the attention mechanism in our MVR. Given an intermediate feature map $\mathbf{F} \in \mathbb{R}^{C \times H \times W}$ as input, the attention module infers attention maps $\mathbf{M}_c \in \mathbb{R}^{C \times 1 \times 1}$ and $\mathbf{M}_s \in \mathbb{R}^{1 \times H \times W}$ sequentially along the channel and spatial dimensions. Then the output of the attention module are multiplied to the input feature according to the following functions

$$\mathbf{F}' = \mathbf{M}_c(\mathbf{F}) \otimes \mathbf{F}, \tag{3}$$

$$\mathbf{F}'' = \mathbf{M}_s(\mathbf{F}') \otimes \mathbf{F}', \tag{4}$$

where the $\otimes$ means element-wise multiplication. $\mathbf{F}'$ is the intermediate feature after the original image feature multiplies the channel attention map and $\mathbf{F}''$ is the final refined feature with the spatial attention map multiplied with $\mathbf{F}'$. Similar to [14], in the MVR, we put this attention module in every ResBlock in ResNet.

Each image in a multi-view batch is first passed through the MVR separately, then aggregated at the view-pooling layer. All branches in the MVR share the same parameters. View-pooling layer is similar to max-pooling layer but the pooling operations are carried out in different dimensions. We perform element-wise maximum operation across all views at the view-pooling layer and get the combined multi-view image feature Φ_{mv}.

2.3. 3D Regression

The goal of our 3D regression module is to process the combined image feature Φ_{mv} and predict the pose and shape parameters $\Theta \in \mathbb{R}^{82}$, including $\theta \in \mathbb{R}^{72}$ and $\beta \in \mathbb{R}^{10}$. Then we use Θ to build the SMPL model as illustrated in section 2.1. Since θ and β are high dimensional vectors, directly regressing them through CNN is dificult. Our work adopts the iterative error feedback loop [9, 15] to progressively regress the residual $\Delta\Theta_t$ according to the previous estimate Θ_t. Then the current parameter is updated by adding the residual to the previous estimate $\Theta_{t+1} = \Theta_t + \Delta\Theta_t$. The 3D regression module takes the current parameters Θ_{t+1} and the combined

image feature Φ_{mv} as inputs, forming a close loop to iteratively regress the pose and shape parameters. During the training process, we only update the weights of the first iteration and other iterations share the same parameters with the first one.

This module is trained with two kinds of supervision. The first is the joints supervision. As mentioned in section 2.1, given the pose and shape parameters $[\theta, \beta]$, we could obtain the predicted 3D joints $\mathcal{J}_{pred} \in \mathbb{R}^{3N}$ through SMPL. Then we calculate the euclidian distance between the ground truth and predicted result

$$\mathbf{L}_{joint} = \|(\mathcal{J}_{gt} - \mathcal{J}_{pred})\|_2^2. \quad (5)$$

Since human pose has high degree of freedom, estimating 3D pose and shape simply with the joints position constraint is unreliable. Model may predict accurate 3D keyjoints without capturing correct limb orientation. We have learned that human pose is a kinematic skeleton, and in this kinematic structure, vectors directing from parent joints to child joints could represent limbs, such as legs and arms. In order to explore the inner correlations of human joints, we bring up a novel geometric supervision of body pose. For the k_{th} joint, its associated limb is defined as a vector directing from parent $j_{parent(k)}$ to its child j_k

$$\mathcal{B}_k = j_{parent(k)} - j_k, \quad (6)$$

where $parent(k)$ returns the index of parent joint for the k_{th} joint. The joints $j_k \in \mathcal{J}$ are defined in the global coordinate and we define the pelvis as the root joint. Estimating the pose using limb representation could express the geometric structure more sufficiently than joints position. Besides, the limbs are more stable than joints and easier to learn. limbs can be learned by minimizing the loss function

$$\mathbf{L}_{limb} = \sum_{k=1}^{K} \|\tilde{\mathcal{B}}_k^{gt} - \tilde{\mathcal{B}}_k\|_1, \quad (7)$$

where the $\tilde{\mathcal{B}}$ means the normalized limb, $\tilde{\mathcal{B}}_k = \frac{\mathcal{B}_k}{\|\mathcal{B}_k\|_2^2}$.

2.4. Discriminator

The estimation of human shape has always been a tough problem since there is no label for human shape in the existing image datasets. Since human shape is highly non-rigid, the regression module may generate unrealistic parameters, such as extreme body shape and unnatural bending of joints. In real life, given an image and its corresponding predicted pose, people could easily tell whether the estimation is correct or not based on the perception of image and pose correspondence and the knowledge of human structure. This perception ability to distinguish right from wrong could be learned by neural network, using adversarial learning.

Based on the above observation, similar to [9, 16], we propose an adversarial learning paradigm in the training process to help the 3D regression module to generate accurate 3D parameters. The discriminator aims at telling ground-truth 3D pose and shape from predicted ones. We train $K + 2$ discriminators, K for K joints, one for the global rotation and one for shape. The loss function for each discriminator is

$$\mathbf{L}_{D_i} = \mathcal{L}_{cls}(D_i(\Theta), 1) + \mathcal{L}_{cls}(D(G(I)), 0), \quad (8)$$

where $D_i(\Theta)$ and $D(G(I))$ represent the outputs of the discriminators for ground-truth parameters from 3D datasets and the predicted parameters respectively. Each discriminator outputs value between [0, 1], representing the possibility that the parameters come from the real data. $\mathcal{L}_{cls}$ is the binary entropy loss defined as $\mathcal{L}_{cls}(\hat{y}, y) = -(y\log(\hat{y})) + (1-y)\log(1-\hat{y})$. Intuitively, $\mathbf{L}_D$ is trained to enforce the discriminator D to classify the ground-truth parameters as 1 and the predictions as 0.

On the contrary, the 3D regression module, here considered as the generator G, is trained through the following function to predict indistinguishable shape and pose to fool the discriminator,

$$\mathbf{L}_G = \mathcal{L}_{cls}(D(G(I)), 1). \quad (9)$$

During the training process, we train the generator and discriminator jointly.

3. EVALUATION

3.1. Implementation Details and Datasets

Implementation details. For MVR, we adopt the refined ResNet brought up in [14] with view pooling layer mentioned in section 2.2, obtaining the combined feature $\Phi_{mv} \in \mathbb{R}^{2048}$. The 3D regression module consists of two fully-connected layers with 1024 neurons with dropout layer in between, followed by a final layer of 82D neurons. In our experiment, the regression module has three iterations. The discriminator for shape is three fully-connected layers with 10, 5, 1 neurons. For pose, θ is first transformed to 3×3 matrices via Rodriguez formula. Then each rotation matrix is passed through two convolutional layers with kernel size of 1 and output channel of 32, followed by a fully-connected layer with 1 neuron. The learning rates for the generator and discriminator are 10^{-5} and 10^{-4} respectively. We use the Adam optimizer and train the model for 50 epochs in Tensorflow [17].

Datasets. For multi-view image datasets, we adopt two popular datasets, Huamn3.6M [18] and MPI-INF-3DHP [19]. For 3D pose and shape dataset used to train the discriminator, we adopt the one brought up in [9], which is generated by applying MoSh [20] to MoCap datasets.

3.2. Experimental Result on Human3.6M

Human3.6M is one of the largest datasets for 3D human pose estimation. We follow the standard evaluation protocol on Human3.6M, using subjects 1, 5, 6, 7, 8 for training and subjects 9 and 11 for testing. We adopt the Mean Per Joint Position Error (MPJPE) to evaluate the pose accuracy. In order to remove the global misalignments, we also use PMPJPE as the evaluation metric, which is the aligned joint error between the predicted results and ground truth using Procrustes alignment.

Fig. 2. **Qualitative results on Human3.6M and MPI-INF-3DHP.** The first row shows results on Human3.6M. The second row shows results on MPI-INF-3DHP. Our model could generate accurate 3D models even though the MPI-INF-3DHP isn't used for training.

Table 1. Results on Human3.6M

Method	MPJPE	PMPJPE	Runtime
Bogo *et al.*[7]	–	82.33	$\sim$ 1 min
Kanazawa *et al.*[9]	87.97	58.1	0.04 sec
Pavlakos *et al.*[10]	71.9	**51.23**	–
Omran *et al.*[11]	78.99	–	–
Rhodin *et al.*[3]	66.8	**51.6**	–
Yang *et al.*[16]	**58.6**	–	–
Zhou *et al.*[21]	**64.9**	–	–
Ours-1vs	82.33	65.41	$\sim$0.03 sec
Ours-2vs	76.32	58.12	$\sim$0.03 sec
Ours-3vs	68.87	58.57	$\sim$0.03 sec
Ours-4vs	**62.49**	**56.6**	$\sim$0.03 sec

* MPJPE and PMPJPE estimation loss in mm. The results are taken from respective papers.

The results on Human3.6M are reported in table 1. Ours-1vs, Ours-2vs, Ours-3vs, Ours-4vs represent the model trained with 1, 2, 3, 4 views of images respectively. We notice that with the increase of the number of views, the joint error decreases. This proves that our multi-view method could boost the performance of pose and shape estimation. Our method outperforms the work of Bogo *et al.*[7], which is a typical traditional method, in both joint error and runtime. We achieve better results than the state-of-the-art methods which estimate both shape and pose parameters, such as [11, 10, 9]. Besides, our method achieves comparable results with methods which only predict 3D pose. We notice that even trained with only 1 view, our method still achieves better results than [9]. This proves the effectiveness of our framework which adopts attention mechanism and explores the body geometry sufficiently.

3.3. Experimental Result on MPI-INF-3DHP

In order to demonstrate how our method generalizes to outdoor scenes and different viewpoints, we evaluate our method on the recent MPI-INF-3DHP dataset, using model trained on Human3.6M. Since the 3D labels of this dataset contain certain noise, follow the common practice, besides MPJPE, we use the Percentage of Correct Keypoints (PCK-h@0.5) and Area Under the Curve (AUC) [22] as the evaluation metrics.

Table 2. Generalization results on MPI-INF-3DHP

Method	PCK	AUC	MPJPE
Mehta *et al.*[19]	76.6	40.4	124.7
Kanazawa *et al.*[9]	72.9	36.5	124.2
Yang *et al.*[16]	69.0	32.0	–
Rhodin *et al.*[3]	66.9	–	–
Ours-4vs	67.0	31.9	139.48

* Accuracy is higher with higher PCK and AUC and lower MPJPE. The results are taken from respective papers.

The results are shown in table 2. Even though this dataset isn't used for training, our best model Ours-4vs achieves 67.0 and 31.9 in PCK and AUC respectively, which is comparable with the state-of-the-art methods [16, 9]. This shows our method is robust and strong enough towards outdoor scenes and different viewpoints. The reason why our method achieves good generalization to different datasets is that, our model is trained to pay attention towards the human geometry in the images and is robust to the diversification of the background. Fig.2 shows the qualitative results on MPI-INF-3DHP and Human3.6M.

4. CONCLUSION

In this paper, we make three contributions towards a full estimation of 3D human. We adopt a novel view-pooling method with attention mechanism to encode multi-view images, providing rich information for human pose and shape estimation without strict camera calibration parameters. We also bring up a novel geometric constraint in the 3D regression module. During the training, we adopt a generative-adversarial way to guide the 3D regression module to estimate accurate human pose and shape parameters. Our method achieves comparable results with the state-of-the-art methods.

References

[1] Thiemo Alldieck, Marcus A Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll, "Video based reconstruction of 3d people models," *arXiv preprint arXiv:1803.04758*, 2018.

[2] Yinghao Huang, "Towards accurate marker-less human shape and pose estimation over time," in *3DV*. IEEE, 2017, pp. 421–430.

[3] Helge Rhodin, Jörg Spörri, Isinsu Katircioglu, Victor Constantin, Frédéric Meyer, Erich Müller, Mathieu Salzmann, and Pascal Fua, "Learning monocular 3d human pose estimation from multi-view images," in *CVPR*, 2018, number CONF.

[4] Federica Bogo, Michael J Black, Matthew Loper, and Javier Romero, "Detailed full-body reconstructions of moving people from monocular rgb-d sequences," in *ICCV*, 2015, pp. 2300–2308.

[5] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black, "Smpl: A skinned multi-person linear model," *ACM Transactions on Graphics (TOG)*, vol. 34, no. 6, pp. 248, 2015.

[6] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis, "Scape: shape completion and animation of people," in *ACM transactions on graphics (TOG)*. ACM, 2005, vol. 24, pp. 408–416.

[7] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black, "Keep it smpl: Automatic estimation of 3d human pose and shape from a single image," in *ECCV*. Springer, 2016, pp. 561–578.

[8] Christoph Lassner, Javier Romero, Martin Kiefel, Federica Bogo, Michael J Black, and Peter V Gehler, "Unite the people: Closing the loop between 3d and 2d human representations," in *CVPR*, 2017, vol. 2, p. 3.

[9] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik, "End-to-end recovery of human shape and pose," in *CVPR*, 2018.

[10] Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis, "Learning to estimate 3d human pose and shape from a single color image," *arXiv preprint arXiv:1805.04092*, 2018.

[11] Mohamed Omran, Christoph Lassner, Gerard Pons-Moll, Peter Gehler, and Bernt Schiele, "Neural body fitting: Unifying deep learning and model based human pose and shape estimation," in *3DV*. IEEE, 2018, pp. 484–494.

[12] Mei Oyama, Naoshi Kaneko Aoyama, Masaki Hayashi, Kazuhiko Sumi, and Takeshi Yoshida, "Two-stage model fitting approach for human body shape estimation from a single depth image," in *Machine Vision Applications (MVA)*. IEEE, 2017, pp. 234–237.

[13] Alexandru O Balan, Michael J Black, Horst Haussecker, and Leonid Sigal, "Shining a light on human pose: On shadows, shading and the estimation of pose and shape," in *ICCV*. IEEE, 2007, pp. 1–8.

[14] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon, "Cbam: Convolutional block attention module," in *ECCV*, 2018.

[15] Joao Carreira, Pulkit Agrawal, Katerina Fragkiadaki, and Jitendra Malik, "Human pose estimation with iterative error feedback," in *CVPR*, 2016, pp. 4733–4742.

[16] Wei Yang, Wanli Ouyang, Xiaolong Wang, Jimmy Ren, Hongsheng Li, and Xiaogang Wang, "3d human pose estimation in the wild by adversarial learning," in *CVPR*, 2018, vol. 1.

[17] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al., "Tensorflow: a system for large-scale machine learning.," in *OSDI*, 2016, vol. 16, pp. 265–283.

[18] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu, "Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 7, pp. 1325–1339, 2014.

[19] Dushyant Mehta, Srinath Sridhar, Oleksandr Sotnychenko, Helge Rhodin, Mohammad Shafiei, Hans-Peter Seidel, Weipeng Xu, Dan Casas, and Christian Theobalt, "Vnect: Real-time 3d human pose estimation with a single rgb camera," *ACM Transactions on Graphics (TOG)*, vol. 36, no. 4, pp. 44, 2017.

[20] Matthew Loper, Naureen Mahmood, and Michael J Black, "Mosh: Motion and shape capture from sparse markers," *ACM Transactions on Graphics (TOG)*, vol. 33, no. 6, pp. 220, 2014.

[21] Xingyi Zhou, Qixing Huang, Xiao Sun, Xiangyang Xue, and Yichen Wei, "Towards 3d human pose estimation in the wild: a weakly-supervised approach," in *ICCV*, 2017.

[22] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt, "Monocular 3d human pose estimation in the wild using improved cnn supervision," in *3DV*, 2017.