3D Human Pose and Shape Estimation Through Collaborative Learning and Multi-view Model-fitting

Zhongguo Li, Magnus Oskarsson, Anders Heyden
Centre for Mathematical Sciences, Lund University
Sölvegatan 18A, Lund, Sweden
zhongguo.li@math.lth.se, magnuso@maths.lth.se, anders.heyden@math.lu.se

Abstract

3D human pose and shape estimation plays a vital role in many computer vision applications. There are many deep learning based methods attempting to solve the problem only relying on single-view RGB images for training the network. However, since some public datasets are captured from multi-view cameras system, we propose a novel method to tackle the problem by putting optimization-based multi-view model-fitting into a regression-based learning loop from multi-view images. Firstly, a convolutional neural network (CNN) regresses the pose and shape of a parametric human body model (SMPL) from multi-view images. Then, utilizing the regressed pose and shape as initialization, we propose an improved multi-view optimization method based on the SMPLify method (MV-SMPLify) to fit the SMPL model to the multi-view images simultaneously. Subsequently, the optimized parameters can be adopted to supervise the training of the CNN model. This whole process forms a self-supervising framework which can combine the advantages of the CNN approach and the optimization-based approach through a collaborative process. In addition, the multi-view images can provide more comprehensive supervision for the training. Experiments on public datasets qualitatively and quantitatively demonstrate that our method outperforms previous approaches in a number of ways.

1. Introduction

Human pose and shape estimation has many applications in virtual/augmented reality and computer games. However, this is a challenging problem since human bodies typically exhibit various motions and shapes in real scenes. Aiming at the problem, there are usually two routes to estimate 3D human pose and shape: optimization-based methods and regression-based methods [18]. Both of the approaches have achieved some success for the problem recently.

Traditionally, through defining a parametric human body model [3, 32, 24, 6, 5] or pre-scanning a 3D model as template [21, 10, 45, 44, 46], optimization-based approaches use some prior information including joint points [6], skeleton[32], silhouettes [5] and RGB-D images [43] to build an energy function. Some work adopt more than one cues in order to achieve better results [12, 1, 44] or propose novel optimization algorithms [21, 10]. By minimizing the energy function, the pre-defined human body model will fit to the prior information, and then, the estimated human pose and shape can be obtained. Although optimization based methods can be used to estimate 3D human body models in many different situations, it is often difficult to automatically extract accurate prior information due to the complexity of real human bodies. In addition, the optimization is often time-consuming.

On the other hand, regression-based methods for human pose and shape estimation have attracted much research with the significant achievements of deep neural networks in many image processing problems [36, 42, 26, 38, 30, 35, 37]. Regression-based methods [15, 39, 31, 27, 4] use deep neural networks that take all or subsets of pixels in the images and regresses the human body and shape parameters based on training on large datasets. Many novel frameworks have been proposed to improve on accuracy of 3D human body estimation [13, 23, 29]. A dataset containing a large number of images and corresponding annotations is required for the methods to train the networks. Both the development of datasets and the time for training are serious drawbacks of regression-based methods. Recently, Kolotouros *et al.* [18] put an optimization-based method into the loop of the regression-based framework and achieved good performance. However, they only used one single-view image during the training.

Considering that some public datasets are captured from multi-view cameras, we propose a novel method for 3D human pose and shape estimation through a collaboration between learning and multi-view model fitting based on multi-view images in this paper. Firstly, a convolutional neural

network (CNN) is advocated to regress the pose and shape parameters of a skinned multi-person linear model (SMPL) from multi-view images. Then, we fit the regressed SMPL model to all the multi-view images simultaneously through optimizing an energy function which is defined according to the joint points of the SMPL model and the ground-truth joint points of the human body in the multi-view images. During the optimization, unlike the single view case in which only pose and shape parameters are optimized, we also optimize the orientation (i.e. the camera view) of the SMPL model for different views to reflect the relation of multi-view images. Finally, in addition to the typical 2D joint points supervision for training, the optimized pose and shape parameters as well as the optimized SMPL model are also adopted to supervise the training of the CNN. Therefore, the CNN can provide initialization of the SMPL model for optimization, while the optimized results can supervise the training process of the CNN, which builds a tight collaboration between the two parts. In addition to this, the multi-view optimization considers the inner relations of the given multi-view images, which can supply more accurate and complete information for the estimation. An overview of our method is shown in Figure 1.

The main contributions of our work have three parts. Firstly, a novel multi-view image based training strategy is used for the training of network, which better exploits the information of the multi-view datasets. Besides, we propose a multi-view model-fitting, merged into a multi-view learning loop to form a novel framework for 3D human pose and shape estimation. Since multi-view model-fitting has better performance than single-view fitting, this provides reliable supervision for training the CNN and the results of our method surpass several recent methods. Finally, our framework can be used for 3D human pose and shape estimation from both single-view images and multi-view images after training the network with multi-view images. The experiments on some public datasets show that our method can better estimate 3D human pose and shape than some previous methods. The code is available at: https://github.com/leezhongguo/MVSPIN_NEW.

2. Related work

There are many previous studies on the problem of human pose and shape estimation aiming at different tasks like joint points estimation, silhouette segmentation, part segmentation and so on. Here we mainly describe those relevant approaches for 3D human pose and shape estimation.

Parametric human body models have been widely used in the estimation of pose and shape. Anguelov *et al*. proposed a data-driven method called SCAPE to generate a deformable human body model [3]. It contained two models which were functions of pose and shape, respectively. They could be combined to create a 3D mesh with realistic

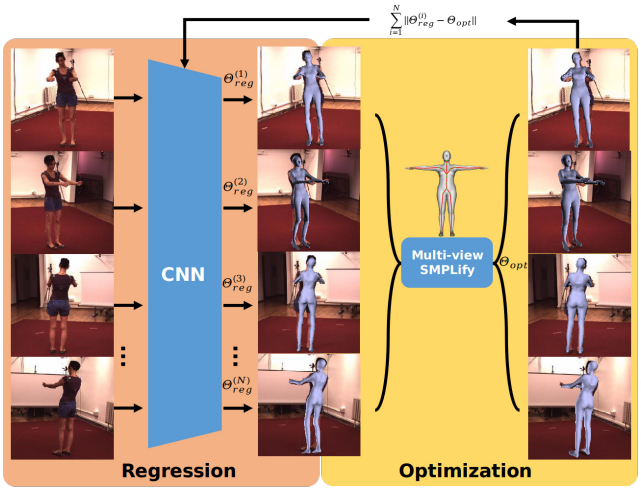


Figure 1. Overview of the proposed method. The CNN regresses the parameters Θ_{reg} from multi-view images. Then, using Θ_{reg} as initialization, multi-view SMPLify optimizes the parameters to obtain Θ_{opt}. The optimized parameters Θ_{opt} of the multi-view images are used to supervise the training of the CNN.

muscle deformation. Some improvements based on SCAPE were proposed over the next several years [43, 41]. A new parametric human body was proposed by Loper *et al*. [24], skinned multi-person linear model (SMPL). It can model various body shapes with natural human poses by defining a function of pose and shape parameters, which made the model be widely used in human pose and shape estimation tasks. Pavlakos *et al*. extended SMPL to SMPL-X by adding more key points on the face, hands and feet [28]. In [33], a dynamic human body model was proposed for modeling human body motion. The above human body models were all learned from a large human body dataset.

Optimization-based methods have traditionally been used to estimate human pose and shape parameters. In [32], a 3D human body model was estimated by fitting SCAPE to the manually acquired joint points and silhouettes. With the development of depth sensors, range data acquired by Kinect was used as prior information and an improved SCAPE model was fitted to the range data in [5, 43]. In addition to the use of prior cues, novel optimization methods were also explored by many researchers [21, 10, 45, 46]. It was also popular to use several different cues to estimates the 3D human body [44, 11]. With the success of human pose estimation by deep neural networks, an automatic approach called SMPLify was proposed to estimate the parameters of the SMPL by using 2D joint points predicted by deep neural networks [6]. Inspired by the method, some approaches based on multi-view images [12, 22] and video [1] were proposed to improve the estimation.

Regression-based methods have also been developed and achieved significant success on 2D [36, 42, 7] and 3D hu-

man pose estimation [30, 38, 9, 37, 35]. Most regression-based work use deep neural networks as encoders to estimate the pose and shape parameters directly from images. The training of the networks often relies on the annotation of 2D/3D joint points [18, 15], dense pose [20], multi-view images [23], silhouettes [8, 31], texture [29, 2] and part segmentation [27]. In [8], silhouettes were used to train a network to estimate the shape of a human body in a simple pose. For human bodies with complicated poses, Kanazawa *et al*. proposed an end-to-end framework using 2D joint locations [15]. In this method, the pose and shape parameters of the SMPL model were learned by the deep neural networks, using the reprojection loss which was defined by ground truth of 2D joint points and the projection of skeleton joints from the SMPL model. Inspired by this framework, many approaches were proposed by designing new routes to acquire various information to better supervise the network. Even for multiple people in images, Zanfir *et al*. [47] proposed a regression-based method to solve the problem. In addition to 2D CNN, some papers use 3D CNN to regress a volume and use a signed distance function to represent a detailed 3D model [13, 34]. In the above methods, Kolotouros *et al*. incorporated SMPLify into the training loop of the CNN, which was the first attempt to combine optimized-based method and regression-based method [18]. This made the training of the CNN self-supervised and achieved competitive performance.

3. Method

The details of our method are presented in this section. We will first introduce the learning-based parametric human body model used in our method. Then, the regression part and the optimization part of our approach are presented, respectively. Based on these two parts, we define the collaboration of them to complete our whole method. Finally, we present the implementation details of our method.

3.1. The SMPL model

The SMPL model is a parametric human body model learned from a very large number of aligned human body shapes. It is a triangulated mesh with $N = 6890$ vertices and the position of each vertex is a linear function $M(\theta, \beta)$ of the pose parameters $\theta \in \mathbb{R}^{72}$ and the shape parameters $\beta \in \mathbb{R}^{10}$. The pose θ encodes the rotation angle of each skeleton joint point in terms of the root point. The shape β contains the coefficients of the ten most significant PCA vectors of the human body models extracted from the human body shape space. In addition, the skeleton joint points $\mathcal{J}$ of the SMPL model are also a linear function of pose θ and shape β. Since it is a linear model, a CNN is expected to perform well, when estimating a regression function to infer the pose and shape parameters. The skeleton joint points of the SMPL model can also be used for the optimization on joint points in order to estimate the pose and shape parameters. Therefore, the SMPL model can be used for both regression and optimization.

3.2. The architecture of our regression CNN

In this section the architecture of the CNN to regress the human body parameters from images is introduced. The design of the network is based on the structure in [18]. Instead of using single view image for one training loop as in [18], we propose to form the multi-view images as a small batch and fed the small batch into the network for one training loop. Given the multi-view images, the network encodes the body in each single view image as a $\mathbb{R}^{85}$ vector containing the pose θ, shape β of the SMPL model and the camera Π as shown in Figure 1. The camera Π is a weak perspective model and is represented by a 3×1 vector (s, t_x, t_y) where s denotes the scale parameter and it can be converted to camera translation. This can be done because the rotation of the camera is assumed to be the identity. Then, the relative rotation between the human body and the camera is coded in the root orientation of the body model. Suppose we have several images from different view-points, denoted $I_i, i = 1, ..., N$ along with the corresponding camera parameters $\Pi_i \in \mathbb{R}^{3\times1}$. Since the multi-view images are from the same human body (pose and shape) from different view-points, the multi-view images have the same ground truth for the pose and shape parameters $\Theta = \{\theta, \beta\}$. For the i-$th$ image I_i passing through the networks, the regressed parameters are defined as $\Theta^{(i)}_{reg} = \{\theta^{(i)}_{reg}, \beta^{(i)}_{reg}\}$ and $\Pi^{(i)}_{reg}$. Then, the predicted 2D joint points can be obtained by projecting the skeleton joint points of the SMPL model through the estimated cameras, i.e., $J^{(i)}_{reg} = \Pi^{(i)}_{reg}(\mathcal{J}(\Theta^{(i)}_{reg}))$, where $\mathcal{J}(\Theta^{(i)}_{reg})$ are the skeleton joint points of the regressed SMPL model. In addition, the predicted mesh of the SMPL model can also be generated by $M^{(i)}_{reg}(\Theta^{(i)}_{reg})$. Therefore, the loss function of the 2D joint points on the multi-view images can be defined as:

$$L_{2D} = \sum_{i=1}^{N} ||J^{(i)}_{reg} - J^{(i)}_{gt}|| \ , \tag{1}$$

where $J^{(i)}_{gt}$ denotes the ground truth of 2D joint points of the i-th input image I_i. Compared to [18], this loss function considers the 2D joint points from all of the views, which can reduce the ambiguity of 2D joint points from a single-view image and provide stronger supervision of the CNN model. In addition to the loss function on 2D joint points, loss function for pose and shape will be discussed in the following sections.

3.3. Multi-view SMPLify

In this section we apply an improved SMPLify method based on multi-view images in order to perform the opti-

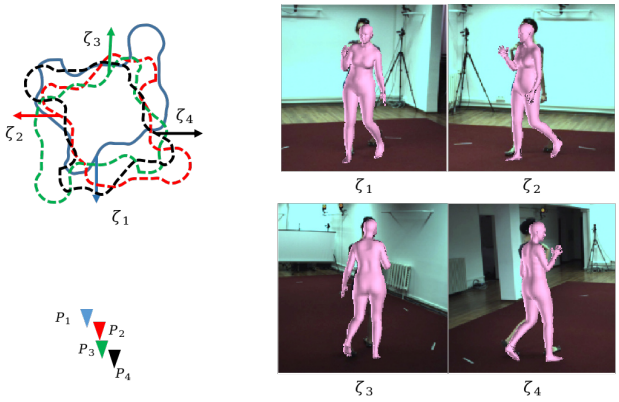


Figure 2. Illustration of the cameras, body orientations and projected SMPL models on the image planes. The four images share the same pose and shape parameters, while the camera translations and body orientations are different.

mization. SMPLify was proposed in [6] and it fitted the SMPL model to a set of 2D joint points predicted by a deep neural network from a single image. In order to extend SMPLify from a single-view image to multi-view images, an improved method was described in [22]. However, the results of [22] are often not robust enough since they initialize the camera rotation as the identity matrix, which may result in the optimization process ending in local optima. In our method, we optimize the body orientation instead of camera rotation because we have assumed that the camera is oriented to human body. According to the definition of the pose θ of the SMPL, the first three elements represent the body orientation denoted by $\zeta \in \mathbb{R}^3$. Then, we define $\tilde{\theta} = \theta \setminus \zeta$ as the pose of the rest joint points. Since the multi-view images share the same pose and shape, we initialize $\tilde{\theta}$ as the mean of $\tilde{\theta}^{(i)}_{reg}$ and β as the mean of $\beta^{(i)}_{reg}$, over all images $i = 1, \ldots, N$. The body orientations $\zeta^{(i)}$ for different views are initialized as $\zeta^{(i)}_{reg}$. We convert the weakly perspective camera $\Pi^{(i)}_{reg}$ to the camera translation $T^{(i)}_{reg}$ and define the camera rotation as the identity matrix. Then, the camera matrix for the projection can be represented as $P^{(i)}_{reg} = \{I, T^{(i)}_{reg}\}$. Using this camera matrix, the reprojected 2D joint points of the regressed SMPL model can be obtained as $P^{(i)}(\mathcal{J}^{(i)})$. Figure 2 illustrates an example of the cameras, body orientations and the corresponding projected regressed SMPL models on the image planes. Based on the above definition, the energy function of the multi-view SMPLify is defined as:

$$\begin{aligned} E(\tilde{\theta}, \beta, \zeta^{(i)}) =& E_J(J^{(i)}_{gt}, P^{(i)}(\mathcal{J}^{(i)})) + \\ & \lambda_\theta E_{\tilde{\theta}}(\tilde{\theta}) + \lambda_\beta E_\beta(\beta) \ , \end{aligned} \tag{2}$$

where E_J measures the errors between J_{gt} and $P^{(i)}(\mathcal{J}^{(i)})$ on all views. $E_\theta(\theta)$ and $E_\beta(\beta)$ are the regularization terms for pose and shape parameters, respectively. For a detailed description of these regularization terms, see [22]. For the energy function above, the minimization is an important step to get the optimized parameters. Similar to [18], fixing the pose and shape parameters, the camera translations of all the images and the orientation of the SMPL model were estimated first. This is implemented by using similar triangles defined by the torso length of regressed SMPL and the ground truth. The initialization of the camera translation T and the body model orientation ζ were obtained from the output of the CNN model. Then, fixing the camera translation, we minimize (2) to obtain the optimized pose $\tilde{\theta}_{opt}$, shape β_{opt} and multi-view body orientation $\zeta^{(i)}_{opt}$. Adam [17] with 0.01 learning rate is used for the optimization and the maximum number of iterations is 100 in our experiments. Therefore, the complete optimized pose for the i-th image is $\theta^{(i)}_{opt} = \{\zeta^{(i)}_{opt}, \tilde{\theta}_{opt}\}$.

3.4. Collaborative learning

In this section we combine the CNN and the multi-view SMPLify into one route in a new training loop. As shown in Figure 1, the regressed pose and shape parameters $\Theta^{(i)}_{reg}$ and camera translation $T^{(i)}_{reg}$ are obtained after the images have passed through the networks. The loss function based on the 2D joint points is defined as in (1), and we use the regressed parameters to initialize the multi-view SMPLify. Through minimizing (2), the optimized parameters can be obtained as $\Theta^{(i)}_{opt} = \{\theta^{(i)}_{opt}, \beta_{opt}\}$. Then, using the optimized $\Theta^{(i)}_{opt}$, the optimized SMPL models and the corresponding skeleton joint points with different body orientations can be generated as $M^{(i)}_{opt}$ and $\mathcal{J}^{(i)}_{opt}$.

Now we can define additional contributing losses to train the CNN by using the above results. The loss for the pose and shape parameters is defined as

$$L_\Theta = \sum_{i=1}^{N} ||\Theta^{(i)}_{reg} - \Theta_{opt}|| \ . \tag{3}$$

Further, the loss function for the mesh of the SMPL model is defined as

$$L_M = \sum_{i=1}^{N} ||M^{(i)}_{reg} - M_{opt}|| \ . \tag{4}$$

In the training dataset, we can also define the loss function of the 3D joint points as

$$L_{3D} = \sum_{i=1}^{N} ||J^{(i)}_{3D} - \mathcal{J}^{(i)}_{opt}|| \ , \tag{5}$$

where $\mathcal{J}^{(i)}_{opt}$ denotes the skeleton joint points of the i-th optimized SMPL model. Therefore, the complete loss function for training the network is defined as:

$$L = \omega_1 L_{2D} + \omega_2 L_{3D} + \omega_3 L_\Theta + \omega_4 L_M \ , \tag{6}$$

where $(\omega_1, \ldots, \omega_4)$ is the weighting of the terms. The loss is defined by mean squared loss function.

Intuitively, our proposed approach has some advantages compared to other methods. Firstly, multi-view images reduce the ambiguity of inferring 3D human pose from 2D joint points. Both in the regression and optimization process, multi-view images can obtain better results than a single view image. Besides, the CNN and the multi-view SMPLify form a tight collaboration during the training loop. The output of the CNN model can initialize the optimization problem, while the optimized results could supervise the training of the CNN model through the loss function defined by optimized parameters.

3.5. Implementation details

Training. In terms of the number of view points, we use four views in our experiments because the training public datasets that we used were acquired from four or eight views. For each training batch, the real number of images is $4 \times N$ where N is the batch-size used in the code. The CNN in our model is trained by Adam with 3×10^{-5} learning rate for 20 epochs. In the total loss function of Equation 6, the weights of each sub-loss $(\omega_1, \omega_2, \omega_3, \omega_4)$ are $(5.0, 5.0, 1.0, 0.001)$. We train our model on two datasets: Human3.6M [14] and MPI-INF-3DHP [25]. In each batch, we use 90% images from Human3.6M and 10% images from MPI-INF-3DHP. All of the images are cropped to 224×224. The network is trained on an NVIDIA TITAN X (Pascal) GPU with 12 GB. The batch-size is set to 16 and each batch takes about 5.5 seconds for one iteration. The total number of iteration is 2441 for one epoch and the whole training takes about 3 days.

Inference. For the inference, we use only a single-view image to evaluate our method. Note that the optimization part is not used in the inference because 2D joint points should be unknown for the inference in practice. More specifically, three datasets are used for inference including the S1 and S9 of Human3.6M, the validation dataset of MPI-INF-3DHP and the test set of 3DPW [40]. These testing images contain various poses and shapes in both indoor and outdoor scenarios.

4. Experiments

In this section some experiments are described to evaluate the performance of our method. We will briefly introduce the datasets used in the experiments for training and evaluation. Then, quantitative and qualitative results are demonstrated to compare the previous methods based on both single-view image and multi-view images, respectively. Finally, an ablation study is given to show the advantage of our method comparing to a method only relying on deep learning.

The metric for quantitative comparison in our experiments contains the reconstruction error, Mean Per Joint Position Error (MPJPE), Percentage of Correct Keypoints with threshold 150 mm (PCK@150 mm) and Area Under Curve (AUC) of 3D joint points. Lower values of the first two metrics means better results, while the higher values of the last two metrics means better results. The reconstruction error is the MPJPE after Procrustes post-processing to remove scale ambiguity. For PCK and AUC, we use the same definition as [26].

4.1. Dataset

Human3.6M. The first dataset in our experiments is the Human3.6M [14]. It contains 11 different subjects and each subject performs 15 different actions indoors. All of the data is acquired from four views and the corresponding 2D/3D joint points and part segmentation are also captured. Similar to previous work [15] which used the protocol 1, the video of the S1, S5, S6, S7 and S8 are used as training dataset, while video S9 and S11 are used for evaluation. For the training set, we extract images from the video every ten frames, while evaluation images are extracted from S9 and S11 every five frames as in [18]. The training set contains 39066×4 images and the evaluation set has 109867 images.

MPI-INF-3DHP. The second dataset is the MPI-INF-3DHP [25]. It contains eight subjects for training and two subjects for testing. For each subjects, eight videos from different views are captured and we choose *video_0*, *video_2*, *video_7* and *video_8* as training data. Only those images with a complete human body in all views are extracted from the videos every ten frames. The testing dataset can be used directly. Totally, the training set has 9452×4 images and the testing set has 2929 images.

3DPW. Since the above datasets are indoor scenarios, we use the test set of 3DPW to evaluate our method on the outdoor scenario case. 3DPW is captured mostly in outdoor conditions using IMU which can provide ground truth 3D pose in the wild. There are 25 test image sequences in 3DPW. After removing some invaild frames, we can obtain totally 35515 images which are used for evaluation.

4.2. Comparison to single-view methods

We compare to the some previous approaches which train the network using single-view images to estimate 3D pose and shape of the human body. Table 1, Table 2 and Table 3 show the quantitative results of some previous work on the Human3.6M, 3DPW and MPI-INF-3DHP, respectively. Note that we use the same testing dataset as the previous methods so that they are comparable. The results of SPIN [18] are obtained through performing the SPIN using the trained model from the original paper, while the results of the other methods come from the corresponding references. We can see from the two tables that our

method outperforms most previous approaches on the three datasets. For the SPIN method which trains the network using single-view images, our method achieved almost the same performance on Human3.6M. This is because SPIN uses four different datasets to train the network, which makes their network more general. However, since we train the network based on multi-view images, the results of our method outperforms SPIN on 3DPW and MPI-INF-3DHP even though we only use Human3.6M and MPI-INF-3DHP to train the network. Therefore, the two tables demonstrate that our method achieves better performance than approaches trained from single-view images.

Methods	Rec.Err. ↓	MPJPE ↓
Pavlakos *et al.* [31]	75.9	-
Omran *et al.* [27]	59.9	-
HMR [15]	56.8	87.97
Kolotouros *et al.* [19]	51.9	74.7
SPIN [18]	44.2	**64.5**
Our	**43.8**	64.8

Table 1. Quantitative comparison to previous work trained by single-view images on Human3.6M.

Methods	Rec.Err. ↓	MPJPE ↓
HMR [15]	76.7	130.0
Kanazawa *et al.* [16]	72.6	116.5
Arnab *et al.* [4]	72.2	-
Kolotouros *et al.* [19]	70.2	-
SPIN [18]	59.2	96.5
Our	**58.6**	**93.4**

Table 2. Quantitative comparison to previous work trained by single-views image on 3DPW.

Methods	PCK/AUC/Rec.Err.	PCK/AUC/MPJPE
VNect [26]	83.9/47.3/98.0	76.6/40.4/124.7
HMR [15]	86.3/47.8/89.8	72.9/36.5/124.2
SPIN [18]	92.1/55.0/68.4	75.3/35.3/109.4
Our	**92.9/56.1/65.6**	**79.2/39.3/98.7**

Table 3. Quantitative comparison to previous work trained by single-view images of MPI-INF-3DHP.

4.3. Comparison to multi-view methods

There are some approaches which also use multi-view images to train the network to regress human pose and shape. Table 6 gives the results of some previous methods based on multi-view images on the test data of Human3.6M. Note that the first three methods did not rely on parametric model to estimate the 3D human pose. They assumed that the cameras were known so that the 2D joint points can be reprojected to 3D space. Therefore, the MPJPE of the three methods was calculated without any ambiguity with the ground truth on the scale or rotation. However, for Liang *et al.* and our method, the 3D poses are the deformed SMPL model and they generally have different scale than the ground truth due to the unknown cameras, so the MPJPE of the two methods are worse. After Procrustes Alignment on the 3D pose of the deformed SMPL model, the effects of ambiguity can be removed and the reconstruction error is more suitable to compare with the MPJPE of the other methods. We can see from the Table 6 that our method achieved the smallest reconstruction error, which demonstrates that our method outperforms the previous methods based on multi-view images on the Human3.6M. Since both Liang *et al.* and our method rely on the SMPL model, we also compare to Liang *et al.* on the 3DPW and MPI-INF-3DHP which contain the images in the outdoor scene in Table 4 and Table 5. Although the method in [23] also uses multi-view images to regress the pose and shape parameters of SMPL, our method still outperforms the method because the MV-SMPLify fully explores the relations between the multi-view images and provides better supervision on the training of the CNN. Therefore, our method achieves satisfying performance on the three datasets even comparing to methods based on multi-view images for training.

Methods	Rec.Err. ↓	MPJPE ↓
Liang *et al.* [23]	-	96.86
Our	**58.6**	**93.4**

Table 4. Quantitative comparison to previous work based on multi-view images on 3DPW.

Methods	PCK/AUC/Rec.Err.	PCK/AUC/MPJPE
Liang *et al.* [23]	86.0/49.0/89.0	66.0/29.0/137.0
Our	**92.9/56.1/65.6**	**79.2/39.3/98.7**

Table 5. Quantitative comparison to previous work based on multi-view images on MPI-INF-3DHP.

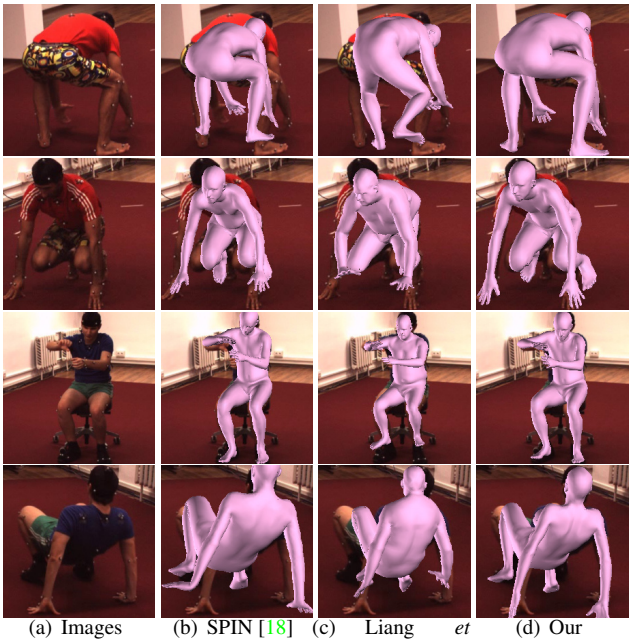
4.4. Qualitative results

In this section, we give some qualitative results of SPIN [18], Liang *et al.* [23] and our method on the datasets of Human3.6M, MPI-INF-3DHP and 3DPW. SPIN combines optimization and regression, but it is a method based on single-view images for training, while Liang *et al.* [23] is the method based on multi-view images for training. Figure 3, Figure 4 and Figure 5 demonstrate several examples from Human3.6M, MPI-INF-3DHP and 3DPW, respectively. In each figure the results of SPIN [18], Liang

Methods	Rec.Err. ↓	MPJPE ↓	Known Camera?	Parametric Model?
PVH-TSP [38]	-	87.3	Yes	No
Trumble *et al.* [37]	-	62.5	Yes	No
Pavlakos *et al.* [30]	-	56.89	Yes	No
Tome *et al.* [35]	-	52.8	Yes	No
Liang *et al.* [23]	45.13	79.85	No	Yes
Our	**43.8**	64.8	No	Yes

Table 6. Quantitative comparison to previous work based on multi-view images on S9 and S11 of Human3.6M.

et al. [23] and our method are shown from the second column to fourth column. The examples shown in the three figures contain various human poses and are captured both in indoor and outdoor scenes.



(a) Images (b) SPIN [18] (c) Liang *et al.* [23] (d) Our

Figure 3. The qualitative results from Human3.6M. From left to right: The original images, the results of SPIN [18], Liang *et al.* [23] and the results of our method.

We can see that the human bodies in the images shown in the Figure 3–5 have complicated poses and shapes with different backgrounds. The figures demonstrate that our method can recover the 3D human body models with better pose and shape estimation than the other two methods. The results of SPIN [18] are also better than the results of Liang *et al.* [23], which shows that putting optimization in the training loop is more useful for the estimation. For the images in the indoor condition, our method achieved almost the same performance as the SPIN [18] on the most examples, especially for the Human3.6M. However, for the images with outdoor condition, our method clearly outperforms SPIN. For example, the last row in Figure 5, SPIN [18] has the errors on the left and right of the body estimation and the results of [23] are also false. For some complicated scenes and poses in 3DPW, for example, the third row in Figure 5, our method also has errors but it still looks better than the two other methods. Since our method uses multi-view images and optimization in the training loop, the results on the fencing of our method are correct. The figures are also consistent with the quantitative results.

(a) Images (b) SPIN [18] (c) Liang *et al.* [23] (d) Our

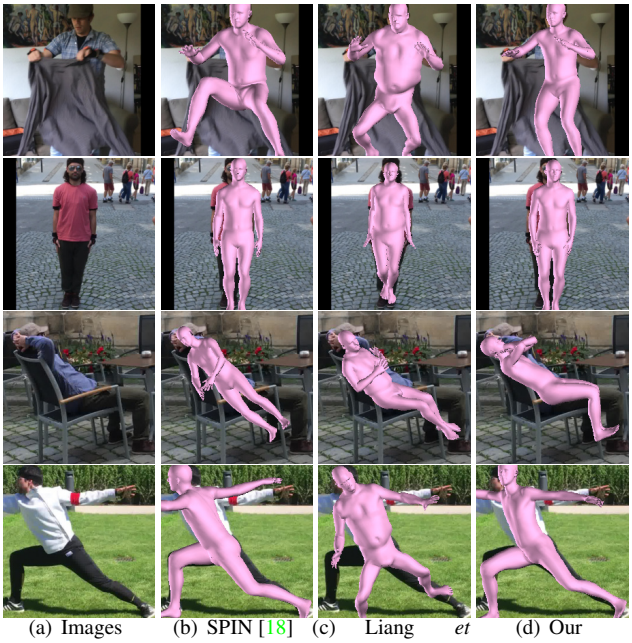
Figure 4. The qualitative results from MPI-INF-3DHP. From left to right: The original images, the results of SPIN [18], Liang *et al.* [23] and the results of our method.

4.5. Comparison to training without optimization

We discuss the effect of multi-view SMPLify on the final estimation on the three datasets. The network was trained with multi-view SMPLify and without multi-view SMPLify, respectively. Table 7 shows the reconstruction error and MPJPE of the two cases. Our method with $L_{2D} + L_{3D}$ stands for the results without multi-view SM-

	Human3.6M		MPI-INF-3DHP		3DPW	
	Rec.Err. ↓	MPJPE ↓	Rec.Err. ↓	MPJPE ↓	Rec.Err. ↓	MPJPE ↓
Our($L_{2D} + L_{3D}$)	46.4	65.8	66.8	100.8	61.7	99.0
Our(Full)	**43.8**	**64.8**	**65.1**	**97.6**	**58.6**	**93.4**

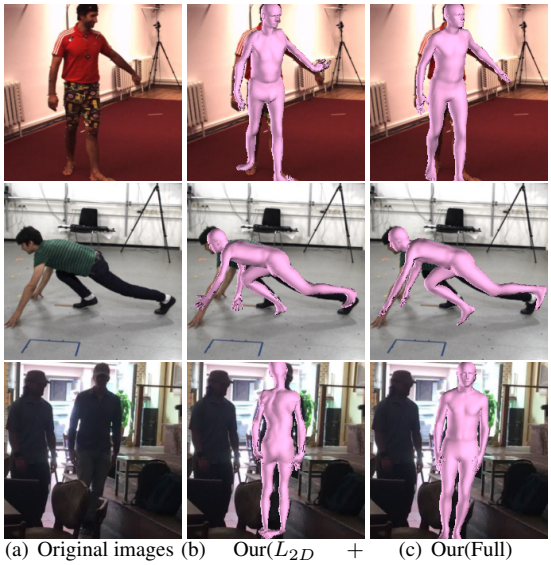
Table 7. The evaluation of the effect of multi-view SMPLify on our method for the three datasets.



(a) Images (b) SPIN [18] (c) Liang *et al.* [23] (d) Our

Figure 5. The qualitative results from 3DPW. From left to right: The original images, the results of SPIN [18], Liang *et al.* [23] and the results of our method.

PLify in the training loop. They only rely on the 2D and 3D joint points losses for training the network. It shows that the accuracy is improved after multi-view SMPLify is used in our training loop. Since the training datasets in our method are Human3.6M and MPI-INF-3DHP, the improvements are not significant. By contrast, the results on 3DPW shows that our full method achieves more clear improvements. Figure 6 shows the qualitative results of our method without and with multi-view SMPLify from the three datasets, respectively. We can see that the results without multi-view SMPLify are worse, especially for the example from 3DPW (the last row in Figure 6). From the results from Human3.6M (the first row in Figure 6), we can see that the final 3D human body is not natural even though the pose is accurate. The wrist and the arm of the 3D model have unnatural blend and rotation. Therefore, only using the 2D and 3D joint points supervision cannot ensure the correct shape of the 3D model. After adding the supervision of multi-view SMPLify, our method can achieve better estimation on the poses and the natural 3D bodies.



(a) Original images (b) Our($L_{2D} + L_{3D}$) (c) Our(Full)

Figure 6. Qualitative results of our method without and with multi-view SMPLify in training loop from the three datasets.

5. Conclusion

In this paper we propose a method to estimate 3D human pose and shape from multi-view images by collaboration between a regression model, a CNN, and an optimization model, multi-view SMPLify. Instead of training the network only from single-view images, multi-view images from some public datasets are utilized for training. The multi-view images are firstly processed by a CNN to regress the pose and shape parameters of the SMPL model as well as the camera parameters. Then, the multi-view SMPLify takes the output of the CNN as initialization to fit the SMPL model to the multi-view images. Multi-view SMPLify achieves better optimized results than SMPLify, which provides stronger supervision of the training. On one hand, our approach sufficiently explores the relations of multi-view images for the network training. On the other hand, the CNN and multi-view SMPLify form a tight self-supervised framework. We validate our method on public datasets and the results of our method indicates the advantage of using multiple views in the training process.

Acknowledgements. We would like to thank the support from ELLIIT, eSSENCE and the China Scholarship Council (CSC).

References

[1] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Video based reconstruction of 3D people models. In *2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[2] Thiemo Alldieck, Gerard Pons-Moll, Christian Theobalt, and Marcus Magnor. Tex2Shape: Detailed full human body geometry from a single image. In *2019 International Conference on Computer Vision (ICCV)*, 2019.

[3] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. SCAPE: Shape completion and animation of people. *ACM Trans. Graph.*, 24(3):408–416, July 2005.

[4] Anurag Arnab, Carl Doersch, and Andrew Zisserman. Exploiting temporal context for 3d human pose estimation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3395–3404, 2019.

[5] Federica Bogo, Michael J. Black, Matthew Loper, and Javier Romero. Detailed full-body reconstructions of moving people from monocular RGB-D sequences. In *2015 International Conference on Computer Vision (ICCV)*, pages 2300–2308, 2015.

[6] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter V. Gehler, Javier Romero, and Michael J. Black. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *2014 European Conference on Computer Vision (ECCV)*, 2014.

[7] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. OpenPose: Realtime multi-person 2D pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2019.

[8] E. Dibra, H. Jain, C. Öztireli, R. Ziegler, and M. Gross. Hs-nets: Estimating human body shape from silhouettes with convolutional neural networks. In *2016 International Conference on 3D Vision (3DV)*, pages 108–117, 2016.

[9] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7297–7306, 2018.

[10] Kaiwen Guo, Feng Xu, Yangang Wang, Yebin Liu, and Qionghai Dai. Robust non-rigid motion tracking and surface reconstruction using L0 regularization. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 3083–3091, 2015.

[11] Marc Habermann, Weipeng Xu, Michael Zollhöfer, Gerard Pons-Moll, and Christian Theobalt. LiveCap: Real-time human performance capture from monocular video. *ACM Trans. Graph.*, 38(2):14:1–14:17, Mar. 2019.

[12] Yinghao Huang, Federica Bogo, Christoph Lassner, Angjoo Kanazawa, Peter V. Gehler, Javier Romero, Ijaz Akhter, and Michael J. Black. Towards accurate marker-less human shape and pose estimation over time. In *2017 International Conference on 3D Vision (3DV)*, 2017.

[13] Zeng Huang, Tianye Li, Weikai Chen, Yajie Zhao, Jun Xing, Chloe LeGendre, Linjie Luo, Chongyang Ma, and Hao Li. Deep volumetric video from very sparse multi-view performance capture. In *2018 European Conference on Computer Vision (ECCV)*, pages 351–369. Springer International Publishing, 2018.

[14] 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 Trans. Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, jul 2014.

[15] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[16] Angjoo Kanazawa, Jason Y Zhang, Panna Felsen, and Jitendra Malik. Learning 3d human dynamics from video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5614–5623, 2019.

[17] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR 2015*, 2015.

[18] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3D human pose and shape via model-fitting in the loop. In *2019 International Conference on Computer Vision (ICCV)*, 2019.

[19] Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4501–4510, 2019.

[20] Verica Lazova, Eldar Insafutdinov, and Gerard Pons-Moll. 360-degree textures of people in clothing from a single image. In *2019 International Conference on 3D Vision (3DV)*, sep 2019.

[21] Hao Li, Bart Adams, Leonidas J. Guibas, and Mark Pauly. Robust single-view geometry and motion reconstruction. *ACM Trans. Graph.*, pages 175:1–175:10, Dec. 2009.

[22] Zhongguo Li, Anders Heyden, and Magnus Oskarsson. Parametric model-based 3D human shape and pose estimation from multiple views. In *21st Scandinavian Conference on Image Analysis (SCIA)*, 2019.

[23] Junbang Liang and Ming C. Lin. Shape-Aware human pose and shape reconstruction using multi-view images. In *2019 International Conference on Computer Vision (ICCV)*, 2019.

[24] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graph.*, 34(6):248:1–248:16, Oct. 2015.

[25] 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 *2017 International Conference on 3D Vision (3DV)*. IEEE, 2017.

[26] 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*, 36(4), July 2017.

[27] M. Omran, C. Lassner, G. Pons-Moll, P. Gehler, and B. Schiele. Neural body fitting: Unifying deep learning and

model based human pose and shape estimation. In *2018 International Conference on 3D Vision (3DV)*, pages 484–494, 2018.

[28] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive Body Capture: 3D hands, face, and body from a single image. In *2019 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[29] Georgios Pavlakos, Nikos Kolotouros, and Kostas Daniilidis. TexturePose: Supervising human mesh estimation with texture consistency. In *2019 International Conference on Computer Vision (ICCV)*, 2019.

[30] Georgios Pavlakos, Xiaowei Zhou, Konstantinos G Derpanis, and Kostas Daniilidis. Harvesting multiple views for marker-less 3d human pose annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6988–6997, 2017.

[31] Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. Learning to estimate 3D human pose and shape from a single color image. In *2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[32] Peng Guan, A. Weiss, A. O. Bălan, and M. J. Black. Estimating human shape and pose from a single image. In *2009 International Conference on Computer Vision (ICCV)*, pages 1381–1388, 2009.

[33] Gerard Pons-Moll, Javier Romero, Naureen Mahmood, and Michael J. Black. Dyna: A model of dynamic human shape in motion. *ACM Trans. Graph.*, 34(4):120:1–120:14, July 2015.

[34] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *2019 IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[35] Denis Tome, Matteo Toso, Lourdes Agapito, and Chris Russell. Rethinking pose in 3d: Multi-stage refinement and recovery for markerless motion capture. In *2018 international conference on 3D vision (3DV)*, pages 474–483. IEEE, 2018.

[36] A. Toshev and C. Szegedy. DeepPose: Human pose estimation via deep neural networks. In *2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1653–1660, 2014.

[37] Matthew Trumble, Andrew Gilbert, Adrian Hilton, and John Collomosse. Deep autoencoder for combined human pose estimation and body model upscaling. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 784–800, 2018.

[38] Matthew Trumble, Andrew Gilbert, Charles Malleson, Adrian Hilton, and John P Collomosse. Total capture: 3d human pose estimation fusing video and inertial sensors. In *BMVC*, volume 2, pages 1–13, 2017.

[39] Gul Varol, Duygu Ceylan, Bryan Russell, Jimei Yang, Ersin Yumer, Ivan Laptev, and Cordelia Schmid. BodyNet: Volumetric inference of 3D human body shapes. In *2018 European Conference on Computer Vision (ECCV)*, September 2018.

[40] Timo von Marcard, Roberto Henschel, Michael J Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3D human pose in the wild using IMUs and a moving camera. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 601–617, 2018.

[41] Ofir Weber, Olga Sorkine, Yaron Lipman, and Craig Gotsman. Context-Aware skeletal shape deformation. *Computer Graphics Forum*, 2007.

[42] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[43] A. Weiss, D. Hirshberg, and M. J. Black. Home 3D body scans from noisy image and range data. In *2011 International Conference on Computer Vision (ICCV)*, 2011.

[44] Weipeng Xu, Avishek Chatterjee, Michael Zollhöfer, Helge Rhodin, Dushyant Mehta, Hans-Peter Seidel, and Christian Theobalt. MonoPerfCap: Human performance capture from monocular video. *ACM Trans. Graph.*, 37(2):27:1–27:15, May 2018.

[45] Tao Yu, Kaiwen Guo, Feng Xu, Yuan Dong, Zhaoqi Su, Jianhui Zhao, Jianguo Li, Qionghai Dai, and Yebin Liu. BodyFusion: Real-time capture of human motion and surface geometry using a single depth camera. In *2017 IEEE International Conference on Computer Vision (ICCV)*, October 2017.

[46] T. Yu, J. Zhao, Z. Zheng, K. Guo, Q. Dai, H. Li, G. Pons-Moll, and Y. Liu. DoubleFusion: Real-time capture of human performances with inner body shapes from a single depth sensor. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2019.

[47] Andrei Zanfir, Elisabeta Marinoiu, Mihai Zanfir, Alin-Ionut Popa, and Cristian Sminchisescu. Deep network for the integrated 3D sensing of multiple people in natural images. In *Advances in Neural Information Processing Systems*, pages 8410–8419, 2018.