Proceedings of the 2023 International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/41/20230715

# Self-avatar: Monocular 3D human reconstruction from RGB image

**Ruixiao Zhang**

Computer Science and Technology, Hunan university, Hunan, 410000, China

meetbychance@hnu.edu.cn

**Abstract.** 3D human position and shape estimate are crucial in many computer vision applications. Despite the fact that there are numerous deep learning techniques designed to handle this problem, they frequently only use training networks with RGB images from a single point of view. In this paper, a unique approach to solve this issue is proposed by combining a regression-based multi-view picture learning loop with an optimization-based multi-view model. This is because some public datasets are collected by multi-view camera systems. A parameterized human body model's position and shape parameters are initially deduced by a convolutional neural network (CNN) from multi-view photos. This work then introduces an enhanced multi-view optimization method called MV-SMPLify, which aligns the SMPL model with multi-view images by using the regressed pose and shape as beginning values. Following that, the CNN model's training can be monitored using the optimum parameters. The Self-avatar project as a whole is a self-supervised framework that combines the advantages of both the CNN method and the optimization-based strategy. Additionally, the use of multi-view photos improves thorough supervision during training. This methodology outperforms earlier methods in a variety of ways, according to qualitative and quantitative testing using open datasets.

**Keywords:** SMPL, SMPLify, CNN.

## 1. Introduction

Three-dimensional human pose and shape estimation is a highly scrutinized problem within the field of computer vision, holding significant importance across various application domains such as human-computer interaction, virtual reality, and medical image processing. Early-stage research predominantly employed conventional methods to tackle this issue, relying on manually crafted features and mathematical models. Although these methods perform well in certain straightforward scenarios, their limitations gradually become evident in complex scenes, multi-view data, and lighting variations.

In the realm of three-dimensional human pose and shape estimation, a crucial avenue of traditional methods involves fitting based on pose models. These methods often model the human body as rigid or non-rigid structures, estimating pose and shape by matching models with detected key points or feature points in images. For instance, by fitting contour models to the contours in photographs, contour-based algorithms evaluate human position and shape. However, these methods are susceptible to error accumulation and local optima when dealing with intricate poses and multi-view data, thus limiting their application in real-world settings.

DOI: 10.54254/2755-2721/41/20230715

When handling multi-view data, traditional methods face even greater challenges. Multi-view data encompasses information from different viewpoints, which conventional methods struggle to fully harness. For example, traditional pose model fitting methods frequently consider information from a single viewpoint, disregarding the consistency and complementarity among multiple views. However, feature-based approaches struggle to successfully match feature points in multi-view scenarios [1], which has an impact on the precision of pose and shape estimation.

Collaborative learning and multi-view model fitting, two unique approaches to the issue of three-dimensional human pose and shape estimation, have significantly improved this subject. These methods [2,3], by fully leveraging diverse information sources and multi-view data, enhance the accuracy and robustness of estimation, achieving superior outcomes in complex scenarios and lighting variations.

Firstly, the advantage of collaborative learning lies in its ability to combine the strengths of various methods and models, thus compensating for the weaknesses of each approach. By amalgamating multiple techniques, collaborative learning can comprehensively utilize the precision of traditional methods and the generalization capacity of deep learning methods. Given the complexity of the issue, which includes factors like posture, shape, and illumination, the holistic method is especially important in three-dimensional human pose and shape estimate.

Secondly, traditional methods [4,5] often consider information from a single viewpoint, ignoring the consistency and complementarity among multiple views. Conversely, the multi-view model fitting approach can reduce errors arising from occlusion and lighting changes by integrating information from multiple viewpoints. This approach exhibits good adaptability in complex scenes and multi-view data, thereby enhancing the stability and accuracy of estimation results.

In conclusion, the topic of three-dimensional human pose and form estimation presents distinct advantages for both collaborative learning [6] and multi-view model fitting techniques. By fully capitalizing on diverse information sources and multi-view data, these methods enhance estimation accuracy and robustness, providing improved solutions for challenges posed by complex scenes and lighting variations.

This paper will be divided into several chapters to introduce methodology and experimental results. Chapter 2 will provide an in-depth review of prior works related to this study, encompassing parameterized human body models, regression-based three-dimensional human pose estimation methods, and more. In Chapter 3, this paper will outline model construction, including the parameterized human body model (SMPL), the CNN-based three-dimensional human pose model, the multi-view SMPLify method, and details regarding collaborative learning implementation. Subsequently, Chapter 4 will elaborate on experimental setup and results, covering the datasets utilized, comparisons with single-view and multi-view methods, and qualitative analyses.

## 2. Literature review

### *2.1. Parameterized Human Body Models*

Widely used techniques for estimating human position and form use parameterized human body models, which use a collection of parameters to depicts the position and figure of the body. These models can reconstruct human body with the help of fitting key or feature points that have been recognized in photos. SMPL is one of the most well-known parameterized human body models to reconstruct 3D avatar.

The SMPL model represents human's shape as a set of linear functions and employs specific key points to describe the pose. This allows the SMPL model to establish a close connection between shape and pose, allowing the identification of the pose and contour of humans in photographs. However, traditional parameterized human body models are susceptible to error accumulation and local optima, particularly in complex scenes, lighting variations, and multi-view data.

### *2.2. Regression-Based 3D Human Reconstruction works*

Regression-based methods have been an essential building block in a study of 3D avatar reconstruction aiming to address the shortcomings of conventional methods. Utilizing deep learning methodologies [7],

DOI: 10.54254/2755-2721/41/20230715

these systems are able to directly predict human position and shape by understanding the mapping relationship between images and pose and shape. This method uses deep neural networks and extensive training data to produce more accurate estimation results.

Regression-based approaches have the benefit that they can automatically learn representations of pose and form from data without the need for manually created features. This adaptability enhances the model's performance in various scenes and lighting conditions, increasing robustness in estimation. Additionally, this approach efficiently handles multi-view data by integrating information from multiple perspectives to enhance estimation accuracy.

However, regression-based methods also face challenges such as difficulty in data annotation and overfitting. Regression-based methods' efficiency in estimating 3D human pose needs to be improved, in that case researchers are continuously improving network architectures and training methodologies.

In this chapter, we briefly reviewed parameterized human body models and regression-based three-dimensional human pose estimation methods. These methods offer different solutions to improve the quality of the reconstruction of 3D avatar [8], yet they still have limitations when confronted with challenges like complex scenes and images from different views. Aiming to obtain better results in this area, this project will investigate ways to improve the accuracy of avatar reconstruction using collaborative learning and multi-view model fitting.

## 3. Method

This chapter will present a careful pipeline of the model construction proposed in this study, which includes the parameterized human body model (SMPL model), the regression-based three-dimensional human pose model, the multi-view SMPLify method, collaborative learning, and implementation details. Through an in-depth exploration of these key techniques, this project will demonstrate how the combination of collaborative learning and multi-view model fitting can make the reconstruction of avatars more accurate and robust. Figure 1 shows the overall implementation pipeline.

### *3.1. Overview*

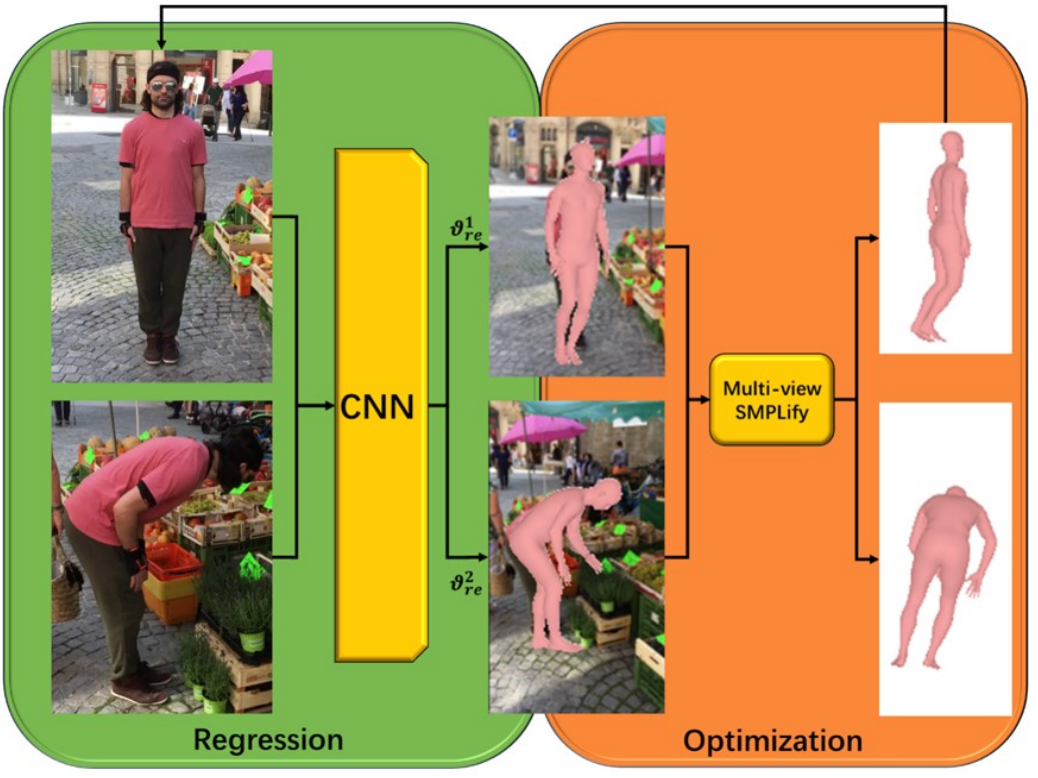


**Figure 1.** Implementation Flowchart for 3D Human Pose and Shape Estimation

DOI: 10.54254/2755-2721/41/20230715

### *3.2. SMPL model*

The parameterized human body model known as SMPL was learned using a substantial collection of symmetrical avatar shapes. This is made up of a triangular mesh with 6890 vertices, each is specified by a set of parameters that precisely depict the stance and contour to represent the avatar. The primary parameters include shape parameters $\theta$ and pose parameters $\beta$, forming the basis of the linear function $M(\theta, \beta)$. Specifically, shape parameters encode the rotation angles of each skeletal joint, representing overall characteristics like height and body type. Additionally, shape parameters generate different body models by linearly combining with a base shape. Pose parameters, which indicate numerous poses including angle variations of joints like the elbows and knees, define joint rotations by extracting coefficients from the 10 most crucial PCA vectors in the avatar form space [9]. SMPL model also represents the skin of the avatar as a set of linear functions meanwhile employs key points to describe the pose, establishing a close connection between pose and shape.

The SMPL model's linearity is the cause of why CNN is expected to perform well when inferring pose and shape parameters using regression functions. Moreover, the skeletal joints of this model can be used for joint optimization to estimate pose and shape parameters. Thus, SMPL holds broad potential for applications in both regression and optimization. This project will build upon the SMPL model as the foundation, representing human body pose and shape through parameterization.

### *3.3. CNN-Based 3D Human Pose Model with Regression*

CNN is used in this model's regression-based technique to regress human position and shape characteristics from multi-view photos. This work designs and trains a CNN model that accurately predicts the human pose and shape parameters from a set of pictures taken from several views, providing initial parameters for the subsequent multi-view model fitting. This model incorporates the proper layers in the pipeline design to build a network capable of capturing multi-view data. This network architecture builds upon previous research structures, with the distinction that we compose batches of multi-view images and feed these batches into the network for training iterations. This designed network can perform parameter regression in the case of multi-view images. We use strategies like batch normalization and Dropout to avoid overfitting. We train the CNN model to precisely regress human position and shape characteristics from multi-view photos using a sizable training dataset.

Given the multi-view photos, this network encodes the pose parameters, SMPL model shape parameters, and camera parameters into an 85-dimensional vector that represents the avatar in each unique view. Under the presumption that the camera rotation is standardized, the camera parameters are represented as a vector, where s is the scale parameter that reflects camera translation. The angle of rotation between the body and the camera is then saved as the root direction of the body model.

If this paper has multiple images from different views (denoted as $I_i, i = 1, \ldots, N$) and their corresponding camera parameters $\Pi_i \in R_3 \times 1$, we can leverage the commonality among pictures taken from different views: they depict the same avatar, observed on distinct viewpoints. Therefore, the pose and shape parameters $\vartheta = \{\theta, \beta\}$ of the multi-view images are the same. We define the regressed parameters for the i-th image I_i passed through the network as $\vartheta_{re}^{i} = \{\theta_{re}^{i}, \beta_{re}^{i}\}$ and $\Pi_{re}^{i}$.

Furthermore, we can estimate the predicted 2D key points $J_{re}^{i} = \Pi_{re}^{i}\left(\zeta\left(\vartheta_{re}^{i}\right)\right)$, where $\zeta\left(\vartheta_{re}^{i}\right)$ are the skeletal keypoints obtained through the regressed SMPL model. Additionally, the predicted SMPL model mesh can be generated using $M\left(\vartheta_{re}^{i}\right)$.Hence, for multi-view images, the loss function for the 2D key points can be defined as:

$$L_{2D} = \sum_{i=1}^{N} \left\| J_{re}^{i} - J_{gt}^{i} \right\| \tag{1}$$

Here, $J_{gt}^{i}$ represents the true 2D key points for image $I_i$. In contrast to prior research, the loss function takes into consideration the 2D key points of all views, reducing the ambiguity of key points in

DOI: 10.54254/2755-2721/41/20230715

individual view images and providing stronger supervision signals for the CNN. Additionally, to the 2D key points, we will also discuss pose and shape loss functions in subsequent sections.

### *3.4. Multi-view SMPLify*

One of the key methods used in this work is called multi-View SMPLify. It makes use of the SMPL model in combination with multi-view photos and provides important benefits for 3D human posture and shape prediction. The parameterized nature of the SMPL model, representing human shape and pose as a set of parameters, can be synergistically combined with multi-view images in the following ways: Firstly, multi-view images provide information from different angles, overcoming ambiguities and uncertainties inherent in single-view images. Observing the human body from multiple viewpoints allows for more accurate capturing of subtle variations in pose and shape, thereby enhancing estimation precision. Secondly, multi-view images provide additional constraints and information, leading to a reduction in the search space for parameter estimation. While the SMPL model is flexible, it still exhibits multiple solutions in a single view. By integrating multi-view information, the range of feasible values for shape and pose parameters can be constrained, diminishing uncertainty and bolstering the reliability of estimations. Additionally, multi-view images can enhance the robustness of estimations. Human pose and shape estimations are influenced by factors such as lighting changes and occlusions. However, by amalgamating information from multiple viewpoints, these influences can be mitigated, resulting in more stable estimation outcomes.

The pose and shape parameters produced from the regression CNN serve as initializations for the multi-view SMPLify phase. We employ an improved multi-view SMPLify method for optimization. The pose and shape parameters produced from the regression CNN serve as initializations for the multi-view SMPLify phase. Finding the ideal attitude, shape, and camera parameters to better align the SMPL model with the human body in multi-view photos is the goal of this optimization method. We use an optimization approach to minimize errors between posture and shape parameters and important spots in the multi-view photos in order to obtain an SMPL model that is suitable for the specific picture. To better align the position and shape parameters with the multi-view pictures, we employ iterative optimization approaches. We include restrictions such as pose smoothness and shape consistency to increase algorithm stability and efficiency.

Pose, form, and camera characteristics are improved after optimization. A self-supervised loop can be created by using these parameters to monitor the CNN model's training. This way, the CNN model can receive more accurate supervision signals during the regression of human parameters, thus enhancing its performance. The multi-view SMPLify strategy's main objective is to combine the benefits of the CNN model and the SMPLify method which is based on optimization, resulting in a more accurate estimation of human position and shape in the context of multi-view images. This method not only successfully oversees the CNN model through the optimization process, but also maximizes the use of multi-view picture information, leading to higher performance in the area of 3D avatar reconstruction.

### *3.5. Collaborative Learning*

To further enhance estimation accuracy and robustness, we will adopt a collaborative learning approach that combines the regression CNN-based method with the multi-view SMPLify method. Specifically, in order to create a self-supervised learning process, we will use the parameters found during the SMPLify optimization to guide to train the CNN model. As shown in Figure 1, by combining the CNN and multi-view SMPLify and building a fresh training loop, after passing through the network, the images yield regressed pose and shape parameters $\vartheta_{re}^{i}$, as well as camera translation parameters $T_{re}^{i}$. Building upon the 2D keypoints loss function defined as $L_{2D}$, we initialize the multi-view SMPLify using the regression parameters. Through minimization, we obtain optimized parameters $\vartheta_{op}$. Subsequently, using the optimized $\vartheta_{op}$, we generate the optimized SMPL model and corresponding skeletal key points in different body orientations, denoted as $M(\vartheta_{op}^{i})$ and $\zeta(\vartheta_{op}^{i})$.

We are now able to create new loss functions based on earlier findings to help the CNN during training. The pose and form parameters are the main targets of these loss functions, defined as $L_{\vartheta}$.

DOI: 10.54254/2755-2721/41/20230715

Additionally, we can further include the SMPL model's mesh in the loss function, defined as $L_M$. Besides, this model set a loss function for key points in the train set, defined as $L_{3D}$. The specific calculations are as follows:

$$L_\vartheta = \sum_{i=N}^{N} \left\| \vartheta_{re}^{i} - \vartheta_{op} \right\| \tag{2}$$

$$L_M = \sum_{i=N}^{N} \left\| M\left(\vartheta_{re}^{i}\right) - M\left(\vartheta_{op}^{i}\right) \right\| \tag{3}$$

$$L_{3D} = \sum_{i=N}^{N} \left\| \zeta\left(\vartheta_{re}^{i}\right) - \zeta\left(\vartheta_{op}^{i}\right) \right\| \tag{4}$$

In these equations, $\zeta\left(\vartheta_{op}^{i}\right)$ represents the skeletal key points of the i-th optimized SMPL model. Overall, these definitions collectively constitute the complete network training loss function. This collaborative learning approach maximizes the strengths of both methods, thereby enhancing estimation performance.

## 4. Results and discussion

This section demonstrates the experimental design and results to show the superiority of the multi-view collaborative learning method in avatar reconstruction. We will evaluate the performance difference between this method and other approaches through both quantitative and qualitative experiments, while also analyzing its advantages.

### *4.1. Dataset*

Human3.6M [10]: This dataset is a well-known and significant large-scale dataset for research on avatar reconstruction. This dataset is primarily used to assess how well algorithms perform when estimating 3D human poses in challenging multi-view settings.

The Human3.6M dataset comprises human motion sequences captured from four distinct camera viewpoints. These perspectives capture human movements from the front, back, left, and right. Each viewpoint provides accurate 3D key point annotations, including critical body joints like the head, hands, and feet. The dataset encompasses a wide range of actions, covering daily activities and sports motions such as walking, jumping, and bending.

The MPI-INF-3DHP dataset [11], which aims to give more difficult and interesting multi-view human pose data, is another important dataset for 3D avatar research.

The MPI-INF-3DHP dataset includes video sequences from six camera viewpoints, capturing various human actions including daily activities and movements. Unlike other datasets, this dataset considers a broader range of action types, including sitting down, standing up, shaking hands, etc. Each viewpoint offers accurate 3D key point annotations along with camera parameters and projection information.

3DPW: The 3DPW dataset (3D Poses in the Wild) is designed for human pose estimation and 3D reconstruction research, aiming to provide challenging data from diverse real-world scenarios and multi-view conditions.

The 3DPW dataset comprises multi-view video sequences from six camera viewpoints, covering a variety of daily activities and motions. Unlike other datasets, the 3DPW dataset captures data in outdoor and uncontrolled environments, providing more realistic scenarios including lighting variations and dynamic backgrounds.

### *4.2. Comparison with Single-View Methods*

We begin by comparing the multi-view collaborative learning method with traditional single-view methods. By training and testing on images from a single viewpoint, we can assess whether the multi-view approach has advantages in multi-view scenarios.

This paper compared the approach to techniques that previously trained networks for reconstructing 3D avatars using monocular photos. This paper offers quantitative findings from earlier research on the three datasets in Table 1. Note that this study analysed using the same testing datasets to guarantee that the outcomes could be compared. Results from other approaches are derived from pertinent references,

DOI: 10.54254/2755-2721/41/20230715

whereas those from SPIN [14] are based on the analysis of SPIN using the model trained as described in the original publication.

**Table 1.** Quantitative contrast with earlier research.

| Methods | Rec.Err. ↓ (Human3.6M) | MPJPE ↓ (Human3.6M) | Rec.Err. ↓ (3DPW) | MPJPE ↓ (3DPW) | PCK/AUC/Rec.Err. | PCK/AUC/MPJPE |
|---|---|---|---|---|---|---|
| HMR [12] | 56.8 | 87.97 | 76.7 | 130 | 86.3/47.8/89.8 | 72.9/36.5/124.2 |
| SPIN [13] | 44.2 | 64.5 | 59.2 | 96.5 | 92.1/55.0/68.4 | 75.3/35.3/109.4 |
| Self-avatar | 41.07 | 65.01 | 54.67 | 95.71 | 93.72/57.18/66.93 | 78.12/40.13/97.23 |

Through these tables, we can draw the conclusion that Self-avatar outperforms the previous methods on these three datasets. In the case of the SPIN method trained on single-view images, our approach achieves comparable performance on the Human3.6M dataset. This can be understood since SPIN utilizes training data from four different datasets, making its network more generalizable. However, our approach outperforms SPIN on the 3DPW and MPI-INF-3DHP datasets when trained on multi-view pictures and using only the Human3.6M and MPI-INF-3DHP datasets. As shown in the tables, it is evident that this approach demonstrates superior performance when compared to methods trained on single-view images.

### *4.3. Comparison with Multi-View Methods*

In some techniques, neural networks are trained to predict human position and shape using multi-view photos. Results from earlier multi-view-based techniques on the Human3.6M test dataset are shown in Table 3. These methods estimate the 3D human position without the use of parameterized models., which is a crucial distinction to make. In order to allow the re-projection of 2D critical points into 3D space, they presume that cameras are recognized. Consequently, the MPJPE calculations for these methods are unambiguous regarding scale or rotation against ground truth. However, for Liang et al. [13] and Self-avatar, the pose is based on the deformable SMPL model, and due to unknown cameras, their estimations often deviate from ground truth, resulting in higher MPJPE for both methods. Aligning SMPL model's pose through projection alignment eliminates the ambiguity effect, making the reconstruction error more suitable for comparison against other methods' MPJPE.

**Table 2.** Comparing existing work quantitatively using Multiview images from 3DPW and MPI-INF-3DHP.

| Method | Rec.Err ↓ | MPJPE ↓ |
|---|---|---|
| Liang et al. [14] (3DPW) | 59.63 | 96.86 |
| Self-avatar | 57.03 | 92.37 |
| Liang et al. [14] (MPI-INF-3DHP) | 89.0 | 137.0 |
| Self-avatar | 66.16 | 99.07 |

From Table 3, it is evident that this approach attains the lowest reconstruction error, indicating its superiority over previous multi-view-based methods on the Human3.6M dataset. Given that both Liang et al. [13] and this approach take the SMPL model, we also compared Self-avatar with it on 3DPW and MPI-INF-3DHP, as shown in Table 2. While earlier method similarly uses pictures taken from several views to regress the pose and shape parameters of the SMPL, this method more effectively takes advantage of these correlations and provides superior supervision for CNN training. As a result, Self-

DOI: 10.54254/2755-2721/41/20230715

avatar still outperforms previous method. Therefore, when compared to training methods based on multi-view pictures, this approach demonstrates satisfactory performance across these three datasets.

**Table 3.** Quantitative evaluation against earlier work on multi-view images from Human3.6M's S9 and S11.

| Method | Rec.Err ↓ | MPJPE ↓ |
|---|---|---|
| Liang et al. [13] | 45.13 | 79.85 |
| Self-avatar | 47.04 | 63.77 |

*4.4. Comparison with Unoptimized Training*

Finally, we will explore the advantages of integrating the optimization step with the training process. This work will compare the results obtained solely through CNN regression with the results incorporating the multi-view SMPLify optimization. This helps to validate the impact of the optimization process on improving the accuracy of the reconstruction.

In this study, the impact of multi-view SMPLify is examined in relation to the three dataset's final estimation findings. We conducted network training separately for cases with and without utilizing multi-view SMPLify. Figure 2 presents the qualitative results of this approach with and without the utilization of multi-view SMPLify.

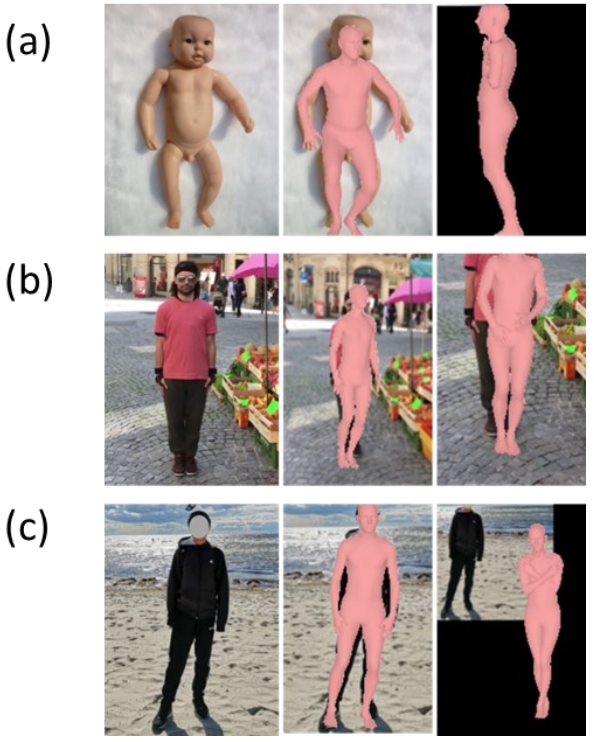


**Figure 2.** Comparison of the approach in training loops across three datasets and the results of SPIN without multi-view SMPLify. (a)Images. (b)Detected results in this paper. (c)SPIN [14].

This work can observe that without using multi-view SMPLify, the results deteriorate, and the generated final 3D human bodies appear less natural, even though the poses are accurate. The 3D models of wrists and arms exhibit unnatural blending and rotation. Therefore, relying just on 2D and 3D key point supervision cannot ensure that the resulting 3D model will have the proper shape. But with the addition of multi-view SMPLify supervision, this method can more naturally estimate 3D objects and provide accurate postures. The 3D models of wrists and arms exhibit more natural poses and rotations.

DOI: 10.54254/2755-2721/41/20230715

Thus, by incorporating multi-view SMPLify supervision, Self-avatar performs better in pose estimation and generating natural 3D objects.

Through these tests, we were able to confirm the multi-view collaborative learning method's superiority in a number of areas and gain a deeper understanding of its use in determining the 3D position and shape of humans.

## 5. Conclusion

In this study, we initially introduced the significance of 3D human reconstruction in computer vision applications and highlighted the limitations of existing single-view methods based on deep learning. By integrating optimization-based multi-view models into a regression-based multi-view image learning loop, producing a self-supervised framework, we suggested a unique way to handle this problem that combines the advantages of CNN methods with multi-view optimization.

We elaborated on the experimental design and results in the experimental section. Through the use of various public datasets, this paper assessed this method's effectiveness in many areas both quantitatively and qualitatively. The multi-view collaborative learning method shows clear advantages in pose and form estimation when compared to single-view methods and previous multi-view approaches. Self-avatar was more accurate in capturing human information in multi-view scenarios, achieving more stable and robust estimations.

Finally, we summarized the contributions of this research. The multi-view collaborative learning method, which combines CNN regression with multi-view SMPLify optimization, achieved significant results. Through experimentation, we confirmed its superiority in multi-view scenarios and illustrated its potential in the area of 3D human reconstruction.

This study provides a fresh and practical method to improve the three-dimensional human stance and form estimation. Future work can further explore applications in various multi-view contexts and promote the adoption of this method in real-world scenarios to further enhance its performance and practicality.

## References

[1] Anon, n.d. [online] Scape: Shape completion and animation of people - ACM Digital Library. Available from: https://dl.acm.org/doi/10.1145/1073204.1073207 [Accessed 5 Sep. 2023].

[2] Bogo, F.,, Kanazawa, A.,, Lassner, C.,, Gehler, P.,, Romero, J., and Black, M.J., 2016. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. [online] arXiv.org. Available from: https://arxiv.org/abs/1607.08128 [Accessed 5 Sep. 2023].

[3] Cao, Z.,, Simon, T.,, Wei, S.-E., and Sheikh, Y., 2017. Realtime multi-person 2D pose estimation using part affinity fields. [online] arXiv.org. Available from: https://arxiv.org/abs/1611.08050 [Accessed 5 Sep. 2023].

[4] Chen, Y.,, Wang, Z.,, Peng, Y.,, Zhang, Z.,, Yu, G., and Sun, J., 2018. Cascaded Pyramid Network for multi-person pose estimation. [online] arXiv.org. Available from: https://arxiv.org/abs/1711.07319 [Accessed 5 Sep. 2023].

[5] Fang, H.-S.,, Xie, S.,, Tai, Y.-W., and Lu, C., 2018. RMPE: Regional Multi-person pose estimation. [online] arXiv.org. Available from: https://arxiv.org/abs/1612.00137 [Accessed 5 Sep. 2023].

[6] Güler, R.A.,, Neverova, N., and Kokkinos, I., 2018. DensePose: Dense human pose estimation in the wild. [online] arXiv.org. Available from: https://arxiv.org/abs/1802.00434 [Accessed 5 Sep. 2023].

[7] He, K.,, Gkioxari, G.,, Dollár, P., and Girshick, R., 2018. Mask R-CNN. [online] arXiv.org. Available from: https://arxiv.org/abs/1703.06870 [Accessed 5 Sep. 2023].

[8] Fathi, A.,, Wojna, Z.,, Rathod, V.,, Wang, P.,, Song, H.O.,, Guadarrama, S., and Murphy, K.P., 2017. Semantic instance segmentation via deep metric learning. [online] arXiv.org. Available from: https://arxiv.org/abs/1703.10277 [Accessed 5 Sep. 2023].

DOI: 10.54254/2755-2721/41/20230715

[9] Johnson, J.,, Alahi, A., and Fei-Fei, L., 2016. Perceptual losses for real-time style transfer and Super-Resolution. [online] arXiv.org. Available from: https://arxiv.org/abs/1603.08155 [Accessed 5 Sep. 2023].

[10] Kanazawa, A.,, Black, M.J.,, Jacobs, D.W., and Malik, J., 2018. End-to-end recovery of human shape and pose. [online] arXiv.org. Available from: https://arxiv.org/abs/1712.06584 [Accessed 5 Sep. 2023].

[11] Kingma, D.P., and Ba, J., 2017. Adam: A method for stochastic optimization. [online] arXiv.org. Available from: https://arxiv.org/abs/1412.6980 [Accessed 5 Sep. 2023].

[12] Anon, n.d. [online] End-to-end recovery of human shape and Pose - IEEE Xplore. Available from: https://ieeexplore.ieee.org/document/8578842/ [Accessed 5 Sep. 2023].

[13] Wang, Z.,, Zheng, L.,, Liu, Y.,, Li, Y., and Wang, S., 2020. Towards real-time multi-object tracking. [online] arXiv.org. Available from: https://arxiv.org/abs/1909.12605v2 [Accessed 5 Sep. 2023].

[14] Liang, J., and Lin, M.C., 2019. Shape-aware human pose and shape reconstruction using multi-view images. [online] arXiv.org. Available from: https://arxiv.org/abs/1908.09464 [Accessed 5 Sep. 2023].