Everybody Is Unique: Towards Unbiased Human Mesh Recovery

Ren Li
https://liren2515.github.io/page/
Srikrishna Karanam
https://karanams.github.io/
Meng Zheng
https://www.linkedin.com/in/meng-zheng-27ab35144
Terrence Chen
https://www.linkedin.com/in/terrencechen
Ziyan Wu
http://wuziyan.com/

United Imaging Intelligence
Cambridge MA, USA
{first.last}@uii-ai.com

Abstract

We consider the problem of obese human mesh recovery, i.e., fitting a parametric human mesh to images of obese people. Despite obese person mesh fitting being an important problem with numerous applications (e.g., healthcare), much recent progress in mesh recovery has been restricted to images of non-obese people. In this work, we identify this crucial gap in the current literature by presenting and discussing limitations of existing algorithms. Next, we present a simple baseline to address this problem that is scalable and can be easily used in conjunction with existing algorithms to improve their performance. Finally, we present a generalized human mesh optimization algorithm that substantially improves the performance of existing methods on both obese person images as well as community-standard benchmark datasets. A key innovation of this technique is that it does not rely on supervision from expensive-to-create mesh parameters. Instead, starting from widely and cheaply available 2D annotations, our method automatically generates mesh parameters that can in turn be used to re-train and fine-tune any existing mesh estimation algorithm. This way, we show our method acts as a drop-in to improve the performance of a wide variety of contemporary mesh estimation methods. We conduct extensive experiments on multiple datasets comprising both standard and obese person images and demonstrate the efficacy of our proposed techniques.

1 Introduction

We consider the problem of human mesh estimation. Given a person image and a functionally-known parametric human mesh, the problem is to fit the mesh (i.e., estimate its parameters) so as to best explain the 3D pose and shape of the person. With many important real-world applications, including in healthcare for the ongoing COVID-19 pandemic, [6, 8, 18, 39],

This work was done during Ren Li's internship with UII. Corresponding author: Srikrishna Karanam.

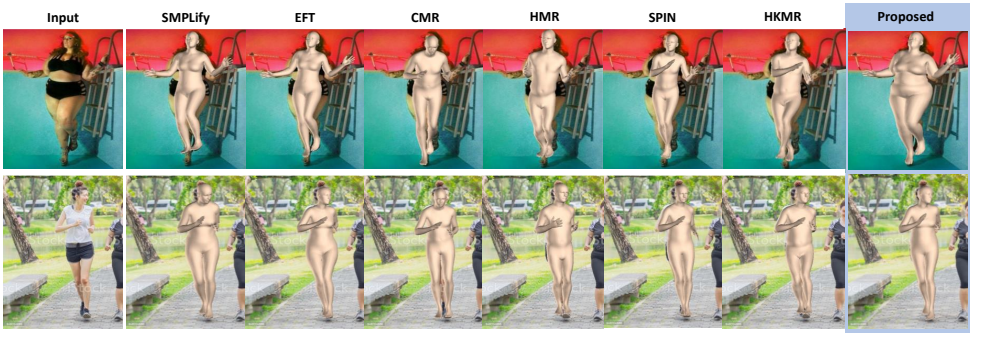

Figure 1: Our method produces improved 3D mesh fits without needing 3D annotations across standard benchmark data and application-specific data such as obese person images.

there has been much recent progress in this field [9, 16, 21]. These applications, e.g., healthcare, demand systems that are robust on a wide variety of data. One such dimension of diversity is the physical size of the person of interest. With obesity prevailing in a considerable section of the world population [19], it is critical that the underlying mesh estimation algorithms work well on images of such people. However, this problem has not attracted much attention in the research community, with Fig. 1 showing some unsatisfactory results with the current state of the art, leading to biased estimates for obese person images.

In this paper, we identify this crucial gap in the literature, and present and discuss the problem of obese human mesh recovery. As noted in prior work [16], training reliable convolutional neural network (CNN) models requires large amounts of data annotated with mesh parameters. However, obtaining these annotations for non-obese person images is not trivial, let alone obese person images (e.g., for SMPL [26], this is an elaborate process involving ground-truth MoCap data and custom marker-based algorithms like MoSH [25]). On the other hand, obtaining 2D annotations, e.g., image keypoints and segmentation, is relatively inexpensive as this can be accomplished using crowdsourcing platforms, e.g., Mechanical Turk. Consequently, while datasets with 2D keypoint annotations can be found in abundance [1, 13, 14, 24], those with full mesh annotations are substantially fewer [10], with obese mesh annotations even more difficult to obtain. Given these considerations, we ask two key questions- (a) *given abundant 2D annotations, can we automatically generate mesh parameters?*, and (b) *can we develop an algorithm that is flexible to address the question above for both standard/non-obese person images in general and obese person images in particular?*

There has been some recent work [4, 15] that propose optimization-based strategies for generating mesh parameters from 2D keypoints. However, a number of issues preclude their use for both general as well as obese mesh fitting. First, the work of Bogo *et al*. [4] does not use the full context information provided by an input image, instead optimizing only a 2D-keypoint-based reprojection error objective, leading to the classic issue of depth ambiguity where multiple 3D configurations may correspond to the same 2D projection. While 3D pose and shape priors may alleviate this issue to a certain extent, these constraints can only ensure the resulting fits belong to a pre-defined distribution. Furthermore, the current community-standard priors used for this purpose [9, 16], however, are not sufficiently representative of obese person images, leading to the unsatisfactory results of Fig. 1 discussed earlier. Next, while the work of Joo *et al*. [15] addresses some of the aforementioned issues by optimizing

the reprojection error cost function for the CNN model parameters (instead of SMPL mesh parameters as in Bogo *et al*. [4]), its performance is also impacted by the priors issue noted above (see Fig. 1 for results) since it starts the optimization from a CNN model that has been pre-trained on the same kind of data. Furthermore, there is a trade-off between performance and number of iterations, increasing which can result in overfitting the reprojection objective.

We take a structured approach to address the questions and issues noted above. First, we present a simple baseline approach that focuses on improving shape fits for obese person images. We achieve this by proposing a loss term that penalizes incorrect shape predictions by means of explicit 2D shape constraints. Next, we propose a generalization of this baseline that inherits the benefits of each strategy as part of an alternating directions scheme that optimizes for *both* mesh- and CNN-parameters jointly (instead of separately as above). Our key insight is that such an alternating iterative framework leads to a virtuous cycle (see Fig 2) where the limitations discussed above can be addressed in a principled manner. Specifically, the issue of depth ambiguity with Bogo *et al*. [4] can be alleviated with a pre-trained data-driven CNN model like in Joo *et al*. [15], whereas the overfitting problem of Joo *et al*. [15] can be addressed by using the pose and shape fits generated by Bogo *et al*. [4] as an explicit regularization term. Our shape constraints can be optionally added to the resulting cost function, leading to both standard/non-obese as well as specific obese mesh fitting (see Fig 1 for improvements with our method).

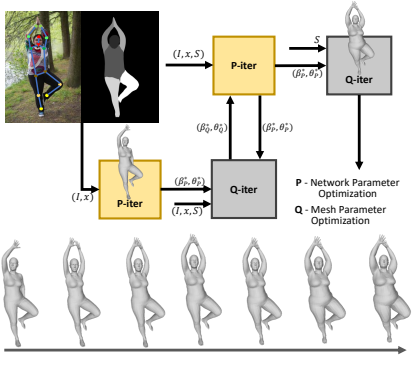

Figure 2: Proposed method; bottom row: iterations

We conduct numerous experiments to evaluate our proposed techniques. First, we discuss limitations of the standard MPJPE metric in capturing shape deviations (from ground truth) in obese images, leading to a new metric. Next, we show that the mesh fits produced by our shape-constraint baseline as well as its generalization outperform both Bogo *et al*. [4] and Joo *et al*. [15] individually on both obese as well as standard benchmark data. Starting from existing pre-trained mesh fitting methods, we then generate inference-time mesh fits with our generalization that result in substantial improvements over the baseline pre-trained models. Specifically, given a test image, we first generate mesh parameters with baseline/pretrained models. We then compute the mesh parameters with our proposed method and compare to the baseline's predictions, showing our method's results substantially outperform the baseline. We call this "inference-time" since our method does not involve any training, requiring only image-specific optimization steps during testing. Finally, since we are able to generate mesh parameters for datasets with only 2D annotations, we can retrain CNN models previously trained using only 2D ground truth, showing substantial performance improvements over the corresponding baselines.

2 Related Work

There is much recent work in human mesh recovery [4, 9, 12, 16, 21, 22, 23, 30, 32, 33, 35, 37, 38, 40, 41]. Here, we discuss a few closely related methods.

Direct mesh regression. Following the end-to-end regression design of Kanazawa *et al.* [16], much effort has been expended in novel architectures, with graph-based [22], structure-based [9], and even video-based [2, 20] approaches being some notable examples. However, as noted in Section 1, these and other related [21] methods produce biased (see Section 3.2) results on obese data while also requiring 3D annotations. In contrast, our method addresses this with a generic framework while only requiring 2D annotations.

Lifting 2D keypoints. There has been some prior work in "lifting" 2D keypoints to 3D data, with approaches based on a direct learning of 2D-3D mapping [5, 17, 27, 31] and a nearest neighbor search in a database of 2D projections [5, 11, 34] being representative examples. However, a key difference is our method is able to exploit the context provided by a full image (as opposed to only 2D keypoints) and recover a complete body mesh. An early method to fit the full body mesh given 2D keypoints was presented in Bogo *et al.* [4] where a cost function based on the 2D reprojection loss and pose/shape priors was optimized, whereas Joo *et al.* [15], optimized the parameters of a pre-trained CNN at test time given 2D keypoints. While these methods present alternative views, we take a more holistic view, arguing that optimizing for both parameter sets jointly leads to substantially improved fits.

3 Method

3.1 Parametric Human Body Representation

We use the Skinned Multi-Person Linear (SMPL) model [26] to represent the 3D human mesh. Given the shape parameters $\boldsymbol{\beta} \in \mathbb{R}^{10}$, pose parameters $\boldsymbol{\theta} \in \mathbb{R}^{72}$, and a fixed pre-trained parameter set $\boldsymbol{\psi}$, SMPL defines the mapping $M(\boldsymbol{\beta}, \boldsymbol{\theta}) : \mathbb{R}^{82} \rightarrow \mathbb{R}^{3N}$ to compute $N = 6890$ 3D body mesh vertices. Given these N vertices $\boldsymbol{J} \in \mathbb{R}^{3N}$, the $K = 24$ joints $\boldsymbol{X} \in \mathbb{R}^{3K}$ defined by the model are obtained as $\boldsymbol{X} = \boldsymbol{W}\boldsymbol{J}$, where $\boldsymbol{W}$ is a learned joint regression matrix. Finally, the 2D image points $\boldsymbol{x} \in \mathbb{R}^{2K}$ can be determined with a known camera model, e.g., a weak-perspective model [16] by defining a function $P(\boldsymbol{X})$ that operates on the 3D joints $\boldsymbol{X}$ as:

$$\boldsymbol{x} = s\Pi(\boldsymbol{X}) + \boldsymbol{t} \tag{1}$$

where $\boldsymbol{t} \in \mathbb{R}^2$ and $s \in \mathbb{R}$ are translation and scale, and Π is an orthographic projection. Therefore, the complete recovery of the 3D mesh corresponding to a person image involves estimating the set of 85 dimensional parameters $\boldsymbol{\Theta} \in \mathbb{R}^{85}$, i.e., $\boldsymbol{\Theta} = \{\boldsymbol{\theta}, \boldsymbol{\beta}, s, \boldsymbol{t}\}$. Note that for notational simplicity in the subsequent sections, we define a function f that takes the $\boldsymbol{\Theta}$ parameters as input and produces the K 3D joints $\boldsymbol{X}$. Given this, the representation $P(f(\boldsymbol{\Theta})) : \mathbb{R}^{85} \rightarrow \mathbb{R}^{2K}$ represents a mapping from the 85-dimensional $\boldsymbol{\Theta}$ parameters to the K 2D image points $\boldsymbol{x}$.

3.2 Current Open Problems and Biased Estimators

Given a training set of n images $\boldsymbol{I}_i, i = 1, \ldots, n$ and their associated parametric annotations $\Theta_i, i = 1, \ldots, n$, the currently dominant paradigm is to train a model to regress the parameters (e.g., with a Euclidean loss between ground-truth and predicted parameters [9, 16, 22]). However, as noted in Section 1, these mesh annotations are either scarcely available or are very expensive to create (e.g., see Loper *et al.* [25]). These issues are only exacerbated for obese person images with little-to-no data available for training obese mesh estimators. On the other hand, 2D annotations can be obtained rather inexpensively. If we can generate

reliable 3D mesh estimates from them, we automatically create annotations for retraining existing models, helping take a step towards unbiased mesh estimators (see Figure 1).

An early approach to address the issue was proposed in Bogo *et al*. [4], where a cost function comprising the 2D reprojection error and associated pose and shape priors was optimized for the mesh parameters $\boldsymbol{\Theta}$. Concretely, this optimization problem is:

$$\boldsymbol{\Theta}^* = \arg\min_{\boldsymbol{\Theta}} L_{2D}(\boldsymbol{x}, \hat{\boldsymbol{x}}), \tag{2}$$

where $L_{2D}(\boldsymbol{x}, \hat{\boldsymbol{x}})$ measures the deviation of the predicted r 2D keypoints $\hat{\boldsymbol{x}} \in \mathbb{R}^{r\times 2}$ from the ground truth $\boldsymbol{x} \in \mathbb{R}^{r\times 2}$. There are a number of issues with this formulation for both obese as well as general mesh fitting. First, the prior terms used in this cost function are not representative of obese person data [16] and the effective shape constrains are missing. Next, as established in recent work [21], the optimization results depend on good initialization, which is not trivial to determine particularly for obese or in-the-wild images. Finally, while minimizing the reprojection error can lead to perfect 2D fits (i.e., 2D loss near zero), the resulting $\boldsymbol{\Theta}$ can still be off due to the classic depth ambiguity problem.

While recent follow-up optimization approaches, e.g., EFT [15], attempt to address some of these issues (e.g., depth ambiguity), the crucial problem of biased estimation remains (see EFT results in Figure 1). Specifically, given a pre-trained mesh regressor (e.g., HMR [16]) $\boldsymbol{\Phi} : \mathbb{R}^{M\times N\times 3} \rightarrow \mathbb{R}^{85}$ trained to predict $\boldsymbol{\Theta} \in \mathbb{R}^{85}$, typically realized as a CNN, this method optimizes a similar 2D reprojection objective with the difference being parameters of optimization are the CNN parameters (instead of the mesh parameters $\boldsymbol{\Theta}$ above). Representing all parameters of the CNN $\boldsymbol{\Phi}$ as the vector $\boldsymbol{\alpha}$, the optimization problem is:

$$\boldsymbol{\alpha}^* = \arg\min_{\boldsymbol{\alpha}} L_{2D}(P(f(\boldsymbol{\Phi}(\boldsymbol{I}))), \boldsymbol{x}), \tag{3}$$

where $\boldsymbol{\Phi}(\boldsymbol{I})$, as noted above, computes the parameters $\boldsymbol{\Theta} = [\boldsymbol{\theta}, \boldsymbol{\beta}, s, \boldsymbol{t}] \in \mathbb{R}^{85}$ given the image $\boldsymbol{I}$. As noted in Section 3.1, given the parameters $\boldsymbol{\Theta}$, the function f then computes the 3D joints $\boldsymbol{X} \in \mathbb{R}^{3\times K}$, which are then projected to 2D image points using the function P from Equation 1. Note that this projection uses the camera parameters $s \in \mathbb{R}$ and $\boldsymbol{t} \in \mathbb{R}^2$ estimated as part of $\boldsymbol{\Theta}$ above. Given $\boldsymbol{\alpha}^*$, and hence the CNN $\boldsymbol{\Phi}^*$, the mesh parameters for the image $\boldsymbol{I}$ are then obtained as $\boldsymbol{\Theta}^* = \boldsymbol{\Phi}^*(\boldsymbol{I})$.

Since this formulation relies on a pre-trained mesh estimator (e.g., HMR [16]), the problem of depth ambiguity can be alleviated to a certain extent with such a data-driven model. Crucially, however, these pre-trained models also rely on the same priors and loss items as above, resulting in the same issue of biased estimation for obese data as above. Further, given that one needs a large number of iterations to obtain good fits, and that the objective comprises only a 2D term, this leads to the risk of overfitting the 2D cost function (i.e., 2D loss can be zero while resulting in an incorrect shape).

3.3 Proposed Method

We take a structured approach to address the issues discussed above. We first propose learnable 2D shape constraints that can be easily integrated into the objective functions of the approaches discussed above, leading immediately to unbiased mesh estimation for obese person images. We then propose a generalized algorithm that optimizes for both mesh and CNN parameters, leading to a holistic optimization-based mesh fitting technique and improved (w.r.t. corresponding base methods) mesh fits for standard benchmark images.

3.3.1 Differentiable 2D shape constraints

To address the issue of bias discussed above, we propose a 2D shape loss term that can be easily added to the training objectives of equations 2 or 3. Specifically, given a 2D binary part segmentation mask $\boldsymbol{S}_i$ ($i = 1, \cdots, 6$ for six parts following LSP's definition [13, 14]) for an image $\boldsymbol{I}$, we define a 2D shape loss as:

$$L_{\text{shape}} = \sum_{i=1}^{6} 1 - \frac{\sum_{m,n} \hat{\boldsymbol{S}}_i^{m,n} \cdot \boldsymbol{S}_i^{m,n}}{\sum_{m,n} \hat{\boldsymbol{S}}_i^{m,n} + \boldsymbol{S}_i^{m,n} - \hat{\boldsymbol{S}}_i^{m,n} \cdot \boldsymbol{S}_i^{m,n}} \tag{4}$$

where $\boldsymbol{S}_i^{m,n}$ is the (m,n) pixel of the ground-truth mask and $\hat{\boldsymbol{S}}_i^{m,n}$ is the (m,n) pixel of the mask estimated during the course of the optimization process. Note this can be easily obtained after the mesh vertices are computed based on the estimated $\boldsymbol{\Theta}$. Given L_{shape}, we can easily adapt Eqs 2 and 3 as $\boldsymbol{\Theta}^* = \arg\min_{\boldsymbol{\Theta}} L_{2D}(\boldsymbol{x}, \hat{\boldsymbol{x}}) + L_{\text{shape}}$ and $\boldsymbol{\alpha}^* = \arg\min_{\boldsymbol{\alpha}} L_{2D}(\pi f(\boldsymbol{\Phi}(\boldsymbol{I})), x) + L_{\text{shape}}$ respectively.

3.3.2 Generalizing mesh and CNN optimization

While our shape constraints help alleviate the bias issues, they do not help tackle the aforementioned problems of depth ambiguity and overfitting. In order to alleviate these issues while also being able to generate unbiased meshes for both obese and non-obese data, we propose *optimization for mesh recovery* (**OMR**), a generalized mesh fitting algorithm that considers both mesh parameters $\boldsymbol{\Theta}$ and model parameters $\boldsymbol{\alpha}$ as an explicit part of the optimization problem. Our core argument is two-fold: (a) the issue of depth ambiguity can be alleviated by using a data-driven predictor, and (b) the issue of overfitting can be tackled by using explicit pose and shape regularizers in the cost function. This leads to our proposed formulation that employs the classical alternating directions scheme as part of a multi-step optimization strategy. Crucially, our proposed L_{shape} can also be optionally integrated in this pipeline, resulting in a generic framework for unbiased mesh estimation.

Given $\boldsymbol{\Phi}$, $\boldsymbol{I}$, and $\boldsymbol{x}$, we first optimize the reprojection loss for $\boldsymbol{\alpha}$, giving an updated CNN:

$$\boldsymbol{\alpha}^* = \arg\min_{\boldsymbol{\alpha}} L_{2D}(P(f(\boldsymbol{\Phi}(I))), \boldsymbol{x}). \tag{5}$$

Note that this first step represents the same process, i.e., optimizing for CNN parameters, we previously explained in Equation 3 above. Given $\boldsymbol{\Phi}^*$, we compute the $\boldsymbol{\Theta}$ prediction for the image $\boldsymbol{I}$ as: $\boldsymbol{\Theta}_0^* = [\boldsymbol{\theta}^*, \boldsymbol{\beta}^*, s^*, \boldsymbol{t}^*] = \boldsymbol{\Phi}^*(I)$. $\boldsymbol{\Theta}_0^*$ is then used to initialize a new optimization problem with respect to mesh parameters:

$$\boldsymbol{\Theta}_1^* = \arg\min_{\boldsymbol{\Theta}} L_{2D}(P(f(\boldsymbol{\Theta}), \boldsymbol{x}) + L_{\boldsymbol{\theta}}(\boldsymbol{\theta}) + L_{\text{shape}}, \tag{6}$$

where $L_\theta(\theta)$ stands for the pose prior (see supplementary material for details). This optimization order provides a good starting point, helping address the initialization issue noted in Section 3.2. Specifically, unlike existing work [4] that uses a mean SMPL pose as the starting point of optimization, using the result of Eq 5 provides a more data-driven (i.e., image-specific) initial pose vector.

Next, we use the $\boldsymbol{\Theta}_1^*$ from Eq 6 as an explicit regularization to optimize the CNN parameters, which we realize by modifying the problem of Eq 3 as:

$$\boldsymbol{\alpha}^* = \arg\min_{\boldsymbol{\alpha}} L_{2D}(P(f(\boldsymbol{\Phi}(I))), \boldsymbol{x}) + \|\boldsymbol{\Theta} - \boldsymbol{\Theta}_1^*\|_2^2 + L_{\text{shape}} \tag{7}$$

Given this $\mathbf{\Phi}^*$, we obtain a new $\mathbf{\Theta}$ as $\mathbf{\Theta}_2^* = [\boldsymbol{\theta}^*, \boldsymbol{\beta}^*, s^*, \boldsymbol{t}^*] = \mathbf{\Phi}^*(I)$. This can now be used to solve a new optimization problem of Eq 6 above, whose solution can be used to solve a new optimization problem of Eq 7, thereby leading to an iterative alternating optimization of $\mathbf{\Theta}$ and $\boldsymbol{\alpha}$. While the procedure above can give the desired final $\mathbf{\Phi}^*$ and $\mathbf{\Theta}^*$, we can further finetune the shape values by integrating our proposed L_{shape} into OMR. To do this, we simply consider the shape-only part $\boldsymbol{\beta}^*$ from $\mathbf{\Theta}^*$ and use it as a starting point to further minimize our shape loss: $\min_{\boldsymbol{\beta}} L_{\text{shape}}$, giving the final shape vector along with the other parameters.

OMR addresses limitations of prior work in a principled manner. First, step 0 ensures a good pose initialization for step 1. Since this only depends on $\mathbf{\Phi}$, OMR can be used as a drop-in to improve any pre-trained model's performance (e.g., we show results with SPIN [21], CMR [22], and HKMR [3]). Next, step 1 provides explicit regularization to address EFT's overfitting issue. Finally, OMR is flexible to be optimized with L_{shape}, resulting in a framework for both obese and general data, leading to reduced bias. Since OMR starts with a P-iteration solution for Eq 5 and subsequently alternates between an Q-iteration Eq 6 and an P-iteration Eq 7, we use the notation $(n+1)PnQ$ to refer to the number of OMR steps (in our experiments, $n = 4$ and each P/Q step has 20 iterations unless mentioned otherwise).

4 Experiments and Results

In this section, we discuss the results of a number of experiments we conducted to demonstrate the efficacy of both our shape constraints as well as OMR.

4.1 Datasets, Evaluation, and a New Metric

For datasets with only 2D keypoint annotations, we use LSP [13], LSP-extended [14], MPII [1] and MS COCO [24]. For datasets with both 2D and 3D keypoint annotations, we use MPI-INF-3DHP [28] and Human3.6M [10]. Furthermore, to demonstrate results on obese person data, we use SSP-3D [36] as well as an internally collected (by scraping the web and manually filtering) `LargeSize` dataset (see supplementary for some examples and all results on LargeSize). While SSP-3D has a varied set of annotated images, we only use data with extreme shape parameters for obese person evaluation, whereas our LargeSize dataset has 2D keypoints and body-part segmentation masks.

We report results with the standard mean per-joint-position-error (MPJPE) metric and its procrustes-aligned variant (PA-MPJPE) [16]. Since they only measure deviation between a set of sparse keypoints, they are insufficient to quantify shape errors. To address this issue, we propose a new metric, called per-vertex-error-T-pose (PVE-T). Given two shape vectors, PVE-T first computes one mesh corresponding to each vector (by setting the pose to mean pose, i.e., zero vector) and rescales the estimated one according to the height of the ground truth. Given the two meshes, it considers all pairs of corresponding vertices, measures their deviation using the Euclidean metric, and returns a mean over all these values. From Fig 3, PVE-T helps capture shape errors more representatively

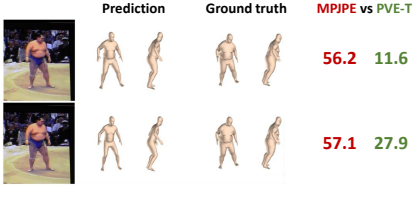

Figure 3: MPJPE vs. PVE-T.

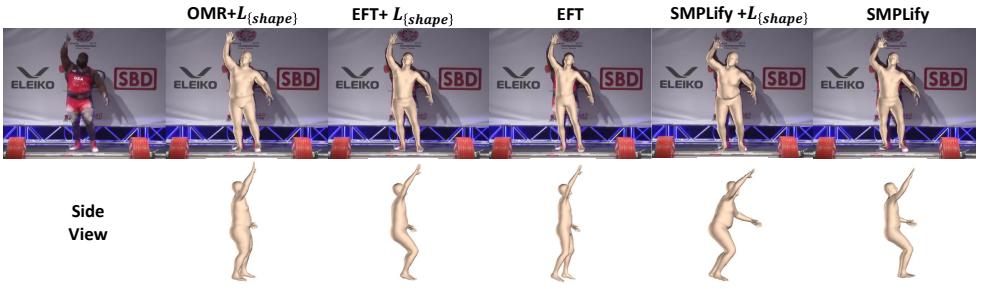

Figure 4: Improvements with our proposed L_{shape}.

(e.g., MPJPE values are almost similar but PVE-T values are more different, thereby being more representative of the deviations between the prediction and the ground truth). Note that such per-vertex comparisons are fairly standard in mesh recovery [20]. Further, unlike traditional 3D surface comparisons (e.g., with Hausdorff), here, given the $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$, the SMPL model produces the same number of vertices (6890) each time. This means the problem of sampling from a continuous 3D surface is not as pronounced as in other surface comparison problems. Since the 6890 SMPL vertices depend on both pose and shape, this "entanglement" is not helpful to isolate shape errors. To this end, with PVE-T, when computing the 6890 vertices, we use the same pose value (i.e., mean pose) across all cases, with shape being the only variation. This helps isolate shape's impact and capture shape deviations more precisely (e.g., as in Fig 3).

4.2 Evaluating Shape Constraints

We first present results of L_{shape} when used with SMPLify [4] and EFT [15]. While we use SPIN [21] as the base model, our method is applicable to and improves the performance of other methods as well (see supplementary for these results). Table 1 shows results on SSP-3D where we see the L_{shape} consistently reduces both PA-MPJPE and PVE-T errors across all methods (the right part of the table shows per-body-part PVE-T values to help understand local shape improvements with L_{shape}). Crucially, L_{shape} helps even a relatively weaker baseline (SMPLify) outperform SPIN on PA-MPJPE. Furthermore, the proposed OMR generalization outperforms both SMPLify and EFT with and without L_{shape} while also substantially reducing the error w.r.t. SPIN (46.26 mm PA-MPJPE vs. 53.57 for SPIN).

SSP-3D	PA-MPJPE (mm)	PVE-T (mm)
SPIN	53.57	35.68
SMPLify w/o L_{shape}	56.78	31.86
SMPLify+L_{shape}	53.23	28.99
EFT w/o L_{shape}	51.96	34.03
EFT+L_{shape}	50.68	32.56
OMR w/o L_{shape}	49.67	32.71
OMR+L_{shape}	**46.26**	**21.52**

SSP-3D (PVE-T)	Torso	Legs	Arms	Head
SPIN	53.56	26.15	38.00	32.16
SMPLify w/o L_{shape}	50.09	19.99	36.68	24.79
SMPLify+L_{shape}	46.31	19.06	34.32	23.68
EFT w/o L_{shape}	54.87	22.54	38.23	29.11
EFT+L_{shape}	51.99	21.61	35.08	29.90
OMR w/o L_{shape}	52.32	21.68	35.35	29.88
OMR+L_{shape}	**26.41**	**19.69**	**20.70**	**21.53**

Table 1: Improving SMPLify and EFT with L_{shape}. OMR outperforms both.

Finally, from Figure 4, one can note how L_{shape} helps improve the shape fits of EFT

and SMPLify. Crucially, while L_{shape} helps improve SMPLify/EFT's shape, the pose gets degraded (see side views). This is addressed by OMR where we see much better results since the P-iteration step explores reasonable poses and provides better initialization for the Q-iteration step, which in turn recovers better shape to guide the P-iteration step.

4.3 Generalized Model Fitting Evaluation

We next evaluate OMR's ability to be used in conjunction with existing methods. To this end, we start with pre-trained base models (we show SPIN [21] in Table 2, and CMR [22] and HKMR [9] in supplementary) and run SMPLify/EFT/OMR on Human3.6M to infer Θ and compute the 3D keypoints (see Section 3.1). Note that while 3D keypoints ground truth are available, we do not use them in any capacity during the optimization process (i.e., they are only used for reporting evaluation metrics). We repeat this for all images in the evaluation set and report average error values. From Table 2, increasing the number of iterations (from 20 to 100) in both SMPLify and EFT leads to overfitting (note increasing errors), whereas OMR is able to address this issue with a decrease in MPJPE/PA-MPJPE.

Human3.6M	MPJPE	PA-MPJPE
SPIN [21]	64.95	43.78
SMPLify - 20	71.99	45.17
SMPLify - 100	82.90	50.23
EFT - 20	61.24	40.82
EFT - 100	63.26	37.95
OMR (1P1Q)	63.18	40.80
OMR (5P4Q)	**61.07**	**37.70**

Table 2: OMR vs. SMPLify/EFT.

Note that OMR's 5P4Q strategy gives the lowest errors that are each substantially better than the corresponding base model's performance (e.g., 64.95 mm for SPIN vs. 61.07 mm for OMR), suggesting OMR's flexibility to be used as a drop-in across multiple different techniques (see supplementary for more results). Finally, the strong performance of OMR across both Tables 1 and 2 suggest its generalizability for both (specific) obese mesh fitting and (generic) non-obese mesh fitting.

Human3.6M	Protocol #1	Protocol #2
	PA-MPJPE	PA-MPJPE
SPIN [21]	44.1	41.1
SPIN - SMPLify	45.2	42.6
SPIN - EFT	44.3	41.5
SPIN - OMR	**43.7**	**41.0**
HKMR [9]	45.9	43.2
HKMR - SMPLify	47.3	44.4
HKMR - EFT	46.3	43.4
HKMR - OMR	**45.6**	**42.9**

Table 3: Improving baseline models.

4.4 Generating Annotations and Model Retraining

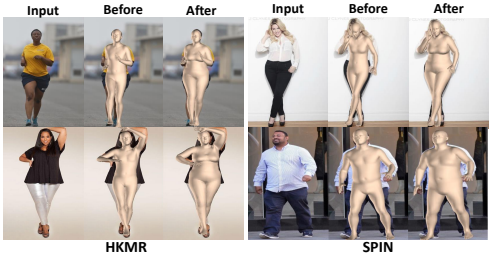

Figure 5: Before and after retraining.

As noted previously, current state-of-the-art methods fail to fit accurate meshes to obese data, leading to biased estimation (see Section 3.2 and Figure 1). Furthermore, as noted in Section 1, while accurate mesh fitting needs 3D parameter supervision, it is expensive to generate these. To address both these issues, we use our proposed OMR algorithm and L_{shape} to automatically generate Θ parameters for LargeSize, LSP, LSP-extended, MPII and MSCOCO. We then retrain state-of-the-art methods with these automatically generated parameters. To demonstrate how this leads to reduced bias on obese data and improved standard benchmark performance, we use HMR [16], CMR [22], SPIN [21] and HKMR

[9]. The corresponding default training configurations are used for a fair evaluation. Figure 5 shows substantially improved shape fits with HKMR and SPIN when compared to their corresponding baseline versions, qualitatively demonstrating the impact of our proposed method. To quantify these gains, in Table 3, we show results on Human3.6M protocols 1 (data from all cameras) and 2 (only data from frontal camera) (more results, including on SSP-3D, in supplementary), where one can note substantial performance improvements after retraining across all the baseline methods. Finally, OMR-generated parameters give improved performance compared to the corresponding SMPLify and EFT versions (see SPIN and HKMR in Table 3), further validating OMR's design.

We finally compare OMR-retrained models with the state of the art. Table 4 shows our results on four test sets, where one can note clear performance improvements. For instance, SPIN-OMR obtains the lowest error on Human 3.6M whereas HKMR-OMR obtains the highest part segmentation accuracy on LSP.

Human3.6M	Protocol #2
	PA-MPJPE
HMR [18]	56.8
CMR [29]	50.1
SPIN [20]	41.1
HKMR [8]	43.2
Pose2Mesh [8]	47.0
HKMR - OMR	42.9
SPIN - OMR	**41.0**

MPI-INF-3DHP	MPJPE
Mehta *et al.* [25]	117.6
VNect [24]	124.7
HMR [18]	124.2
HKMR [8]	108.9
SPIN [20]	105.2
HKMR - OMR	**100.1**
SPIN - OMR	100.9

3DPW	PA-MPJPE
HMR [18]	81.3
CMR [29]	70.2
HKMR [8]	76.7
SPIN [20]	59.2
Pose2Mesh [8]	58.9
I2LMeshNet [23]	57.7
HKMR - OMR	**56.1**
SPIN - OMR	56.5

LSP	Part acc.
Oracle [8]	88.82
SMPLify [8]	87.71
HMR [18]	87.12
SPIN [20]	89.41
HKMR [8]	89.59
HKMR - OMR	**89.86**
SPIN - OMR	89.76

Table 4: Comparison with competing state-of-the art methods.

5 Summary

In this work, we considered the problem of human mesh recovery with a particular emphasis on mesh estimation for images of obese people. We noted that the current state-of-the-art methods produce biased estimates for obese images, discussed our reasoning behind this issue, and proposed ways to overcome this problem. Specifically, we first proposed new 2D shape constraints that can be flexibly used in conjunction with existing mesh fitting algorithms. We showed how this results in an immediate improvement in baseline performance. We then proposed a generalized mesh fitting algorithm, called OMR, that optimizes a reprojection error cost function in a space of both mesh parameters and CNN model parameters, showing how this results in a holistic approach that addresses the limitations of existing mesh optimization algorithms. We then showed the proposed 2D shape constraints can be easily integrated into OMR while also having the flexibility to be used with any contemporary regression-based mesh recovery algorithm. We demonstrated the efficacy of our algorithms by means of extensive experiments on both obese person data and standard benchmark data, establishing new baseline results for obese mesh recovery and state-of-the-art performance of benchmark human mesh recovery. While the proposed method is able to generate reasonably reliable annotations, this depends on the 2D data (e.g., keypoints) being relatively accurate. In cases when this does not happen (e.g., occlusions), one possible direction for future research can be to exploit adjacent or additional data modalities (e.g., depth or multi-view RGB) to "fill-in" the missing/noisy data.

References

[1] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *CVPR*, 2014.

[2] Anurag Arnab, Carl Doersch, and Andrew Zisserman. Exploiting temporal context for 3d human pose estimation in the wild. In *CVPR*, 2019.

[3] Sandika Biswas, Sanjana Sinha, Kavya Gupta, and Brojeshwar Bhowmick. Lifting 2d human pose to 3d: A weakly supervised approach. In *IJCNN*, 2019.

[4] 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*, 2016.

[5] Ching-Hang Chen and Deva Ramanan. 3d human pose estimation= 2d pose estimation+ matching. In *CVPR*, 2017.

[6] William Ching, John Robinson, and Mark McEntee. Patient-based radiographic exposure factor selection: a systematic review. *Journal of medical radiation sciences*, 61 (3):176–190, 2014.

[7] Hongsuk Choi, Gyeongsik Moon, and Kyoung Mu Lee. Pose2mesh: Graph convolutional network for 3d human pose and mesh recovery from a 2d human pose. ECCV, 2020.

[8] Yicheng Fang, Huangqi Zhang, Jicheng Xie, Minjie Lin, Lingjun Ying, Peipei Pang, and Wenbin Ji. Sensitivity of chest ct for covid-19: comparison to rt-pcr. *Radiology*, 296(2):E115–E117, 2020.

[9] Georgios Georgakis, Ren Li, Srikrishna Karanam, Terrrence Chen, Jana Kosecka, and Ziyan Wu. Hierarchical kinematic human mesh recovery. In *ECCV*, 2020.

[10] 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*, 36(7):1325–1339, 2013.

[11] Umar Iqbal, Andreas Doering, Hashim Yasin, Björn Krüger, Andreas Weber, and Juergen Gall. A dual-source approach for 3d human pose estimation from single images. *Computer Vision and Image Understanding*, 172:37–49, 2018.

[12] Wen Jiang, Nikos Kolotouros, Georgios Pavlakos, Xiaowei Zhou, and Kostas Daniilidis. Coherent reconstruction of multiple humans from a single image. In *CVPR*, 2020.

[13] Sam Johnson and Mark Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *BMVC*, 2010.

[14] Sam Johnson and Mark Everingham. Learning effective human pose estimation from inaccurate annotation. In *CVPR*, 2011.

[15] Hanbyul Joo, Natalia Neverova, and Andrea Vedaldi. Exemplar fine-tuning for 3d human pose fitting towards in-the-wild 3d human pose estimation. *arXiv preprint arXiv:2004.03686*, 2020.

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

[17] Yangyuxuan Kang, Anbang Yao, Shandong Wang, Ming Lu, Yurong Chen, and Enhua Wu. Explicit residual descent for 3d human pose estimation from 2d joint locations. In *BMVC*, 2020.

[18] Srikrishna Karanam, Ren Li, Fan Yang, Wei Hu, Terrence Chen, and Ziyan Wu. Towards contactless patient positioning. *IEEE Transactions on Medical Imaging*, 2020.

[19] David A Kass, Priya Duggal, and Oscar Cingolani. Obesity could shift severe covid-19 disease to younger ages. *Lancet (London, England)*, 2020.

[20] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. Vibe: Video inference for human body pose and shape estimation. In *CVPR*, 2020.

[21] 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 *ICCV*, 2019.

[22] Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. In *CVPR*, 2019.

[23] Jogendra Nath Kundu, Mugalodi Rakesh, Varun Jampani, Rahul Mysore Venkatesh, and R Venkatesh Babu. Appearance consensus driven self-supervised human mesh recovery. *arXiv preprint arXiv:2008.01341*, 2020.

[24] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, 2014.

[25] Matthew Loper, Naureen Mahmood, and Michael J Black. Mosh: Motion and shape capture from sparse markers. *ACM Transactions on Graphics*, 33(6):1–13, 2014.

[26] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM Transactions on Graphics*, 34(6):1–16, 2015.

[27] Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little. A simple yet effective baseline for 3d human pose estimation. In *ICCV*, 2017.

[28] 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.

[29] 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):1–14, 2017.

[30] Gyeongsik Moon and Kyoung Mu Lee. I2l-meshnet: Image-to-lixel prediction network for accurate 3d human pose and mesh estimation from a single rgb image. *arXiv preprint arXiv:2008.03713*, 2020.

[31] Francesc Moreno-Noguer. 3d human pose estimation from a single image via distance matrix regression. In *CVPR*, 2017.

[32] Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. Learning to estimate 3d human pose and shape from a single color image. In *CVPR*, 2018.

[33] Georgios Pavlakos, Nikos Kolotouros, and Kostas Daniilidis. TexturePose: Supervising human mesh estimation with texture consistency. In *ICCV*, 2019.

[34] Gerard Pons-Moll, David J Fleet, and Bodo Rosenhahn. Posebits for monocular human pose estimation. In *CVPR*, 2014.

[35] Nadine Rueegg, Christoph Lassner, Michael J Black, and Konrad Schindler. Chained representation cycling: Learning to estimate 3d human pose and shape by cycling between representations. *arXiv preprint arXiv:2001.01613*, 2020.

[36] Akash Sengupta, Ignas Budvytis, and Roberto Cipolla. Synthetic training for accurate 3d human pose and shape estimation in the wild. In *British Machine Vision Conference (BMVC)*, September 2020.

[37] Akash Sengupta, Ignas Budvytis, and Roberto Cipolla. Synthetic training for accurate 3d human pose and shape estimation in the wild. *arXiv preprint arXiv:2009.10013*, 2020.

[38] Donglai Xiang, Hanbyul Joo, and Yaser Sheikh. Monocular total capture: Posing face, body, and hands in the wild. In *CVPR*, 2019.

[39] Fan Yang, Ren Li, Georgios Georgakis, Srikrishna Karanam, Terrence Chen, Haibin Ling, and Ziyan Wu. Robust multi-modal 3d patient body modeling. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2020.

[40] Andrei Zanfir, Eduard Gabriel Bazavan, Hongyi Xu, Bill Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Weakly supervised 3d human pose and shape reconstruction with normalizing flows. *arXiv preprint arXiv:2003.10350*, 2020.

[41] Jason Y Zhang, Sam Pepose, Hanbyul Joo, Deva Ramanan, Jitendra Malik, and Angjoo Kanazawa. Perceiving 3d human-object spatial arrangements from a single image in the wild. In *ECCV*, 2020.