

Probabilistic Modeling for Human Mesh Recovery

Nikos Kolotouros¹, Georgios Pavlakos², Dinesh Jayaraman¹ Kostas Daniilidis¹
¹ University of Pennsylvania ² UC Berkeley

Abstract

This paper focuses on the problem of 3D human reconstruction from 2D evidence. Although this is an inherently ambiguous problem, the majority of recent works avoid the uncertainty modeling and typically regress a single estimate for a given input. In contrast to that, in this work, we propose to embrace the reconstruction ambiguity and we recast the problem as learning a mapping from the input to a **distribution** of plausible 3D poses. Our approach is based on the normalizing flows model and offers a series of advantages. For conventional applications, where a single 3D estimate is required, our formulation allows for efficient mode computation. Using the mode leads to performance that is comparable with the state of the art among deterministic unimodal regression models. Simultaneously, since we have access to the likelihood of each sample, we demonstrate that our model is useful in a series of downstream tasks, where we leverage the probabilistic nature of the prediction as a tool for more accurate estimation. These tasks include reconstruction from multiple uncalibrated views, as well as human model fitting, where our model acts as a powerful image-based prior for mesh recovery. Our results validate the importance of probabilistic modeling, and indicate state-of-the-art performance across a variety of settings. Code and models are available at: <https://www.seas.upenn.edu/~nkolot/projects/prohmr>.

1. Introduction

Reconstructing 3D human pose from any form of 2D observations (image, 2D keypoints, silhouettes) is a fundamentally ambiguous problem. Of course, this is a very old insight, identified even from the very first approaches [25] dealing with the problem of single-view human pose reconstruction. However, the current norm for the state-of-the-art approaches is to return a single 3D estimate which is typically computed in a deterministic manner. In this work, we argue that there is great value at capturing a distribution of 3D poses conditioned on the preferred input.

Our reliance on systems that return a single deterministic

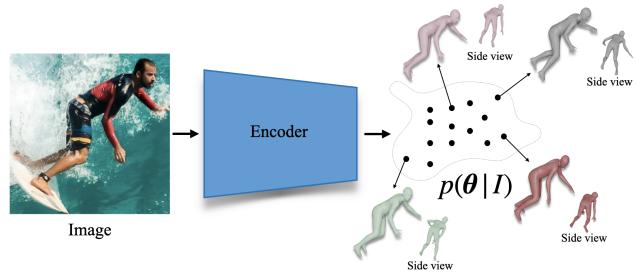


Figure 1: **Probabilistic modeling for 3D human mesh recovery.** We propose to recast the problem of 3D human reconstruction as learning a mapping from the input to a distribution of 3D poses. The output distribution has high probability mass on a diverse set of poses that are consistent with the 2D evidence.

3D pose output often happens out of convenience; it makes comparison on conventional benchmarks straightforward and fair, while a single output is enough for many downstream applications. Recent literature for 3D human pose reconstruction is currently dominated by such approaches and they are very popular for image [22] or keypoint [43] input, for skeleton-based [32] or mesh-based [23] reconstruction, as well as regression [17] or optimization-based [4] approaches. On the other end of the spectrum, there have always been approaches that advocate in favor of generating multiple predictions per input. Recent efforts have demonstrated interesting potential [3, 27], but often rely on ensemble-type prediction, modifying current systems into combining N output heads instead of one. This can lead to cumbersome architectural choices, inability to scale and/or limited expressivity for the output distribution.

Our approach aims to bridge this gap and demonstrate the value of predicting a distribution of 3D poses conditioned on the provided 2D input. To achieve this, we propose an elegant and efficient approach with many desirable properties missing from recent work, and we demonstrate its effectiveness. Instead of regressing a single estimate for the provided input, we use Normalizing Flows to regress a distribution of plausible poses. This allows us to train a network which returns a conditional distribution of 3D poses as a function of the input (*e.g.*, image or 2D keypoints), as depicted in Figure 1. Our probabilistic model

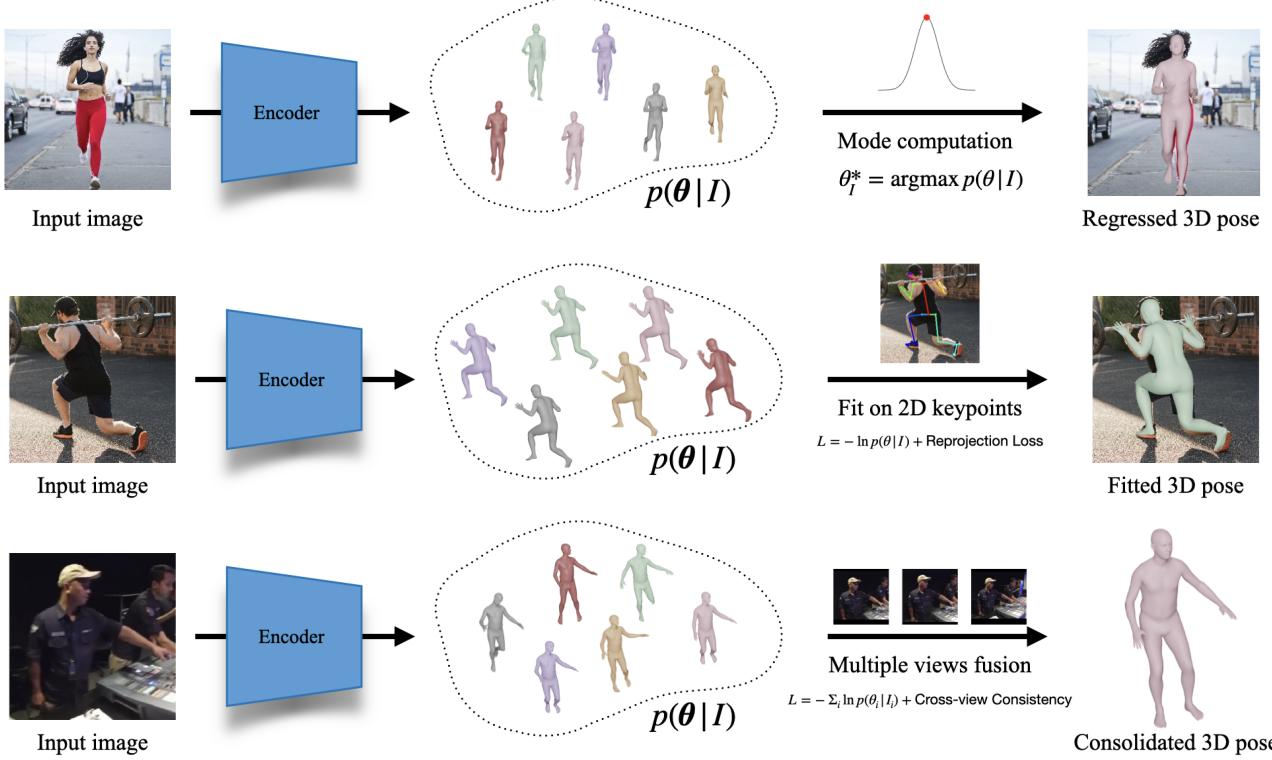


Figure 2: The value of probabilistic modeling for 3D human mesh estimation. We demonstrate that probabilistic modeling in the case of 3D human mesh estimation can be particularly useful because of its elegant and flexible form, which enables a series of downstream applications. First row: In the typical case of 3D mesh regression, we can naturally use the mode of the distribution and perform on par with approaches regressing a single 3D mesh. Second row: When keypoints (or other types of 2D evidence) are available we can treat our model as an image-based prior and fit a human body model to the keypoints by combining it with a 2D reprojection term. Third row: When multiple views are available, we can naturally consolidate all single-frame predictions by adding a cross-view consistency term. We underline that all these applications refer to test-time behavior and they use the same trained probabilistic model (no per-task training required).

allows for fast sampling of diverse outputs, we can efficiently compute the likelihood of each sample, and there is a fast and closed form solution to compute the mode of the distribution. The importance of the above is manifested in a variety of ways, which are summarized in Figure 2. First, we can easily compute the mode of the distribution, which returns the most likely 3D pose for the particular input. This is convenient, when a single estimate is required for some applications. Interestingly, this regressed value is on par with the state-of-the-art deterministic methods, so our model can be valuable even in the more conventional settings. More importantly though, by treating our trained probabilistic model as a conditional distribution, we can use it in many downstream applications to combine information from different sources. For example, when 2D keypoints are available, optimization approaches [4, 38], are used to fit parametric human body models to these 2D locations. In this case, our model can act as a powerful image-based prior that can guide the optimization towards accurate solutions that satisfy both 2D keypoint reprojection and image evidence. Similarly, when multiple views are available, we

can consolidate information from all conditional distributions, by optimizing for cross-view consistency and recover a 3D result that is consistent with the available observations. Last but not least, we highlight that all these applications are available at test-time with the same trained probabilistic model, without any need for task-specific retraining.

We conduct extensive experiments to demonstrate the importance of our learned probabilistic model. We focus primarily on image-based mesh recovery [17], proposing the **ProHMR** model, but we also investigate 2D keypoint input [32]. We achieve particularly strong performance across different tasks and evaluation settings. Our contributions can be summarized as follows:

- We propose a probabilistic model for human mesh recovery and demonstrate its value in various tasks.
- In the conventional evaluations with single estimate methods, our model is on par with the state of the art.
- We demonstrate that in the presence of additional information sources, *e.g.*, multiple views or 2D key-

points, our model offers an elegant and effective way to consolidate said sources.

- In the setting of human body model fitting, our model acts as a powerful image-based prior, achieving significant boost over previous baselines.

2. Related work

Although our formulation is quite general and can handle different inputs/outputs, here we focus mainly on human mesh recovery from a single image [17], while we briefly touch upon other settings, specifically 3D pose estimation from 2D keypoints [32]. Since the related work is vast, here we discuss the more relevant approaches. We direct the interested reader to a recent and extensive survey [51].

2.1. Human mesh recovery from a single image

Regression: Recent approaches for mesh recovery are following the regression paradigm, where the parameters of a parametric model [30, 38, 48, 36] are regressed from a deep network, given a single image as input. The canonical example here is HMR [17], with many of the design decisions being adopted also by follow-up works, *e.g.*, [2, 11, 23, 39, 6, 9, 15]. Here, our regression network also follows the principles of HMR, however, instead of regressing a single 3D pose estimate, it regresses a whole distribution of plausible 3D poses given the input image.

Optimization: These methods estimate iteratively the parameters of the body model, such that it is consistent with a set of 2D cues. The canonical example of SMPLify [4] optimizes SMPL parameters given 2D keypoints. Follow-up works investigate other inputs, *e.g.*, silhouettes [24], POFs [47], dense correspondences [11] or contact [34, 44]. However, most recent approaches [2, 22, 38] rely almost exclusively on 2D keypoints; losing the majority of pictorial cues, but gaining robustness. In this work, we demonstrate how our probabilistic model can leverage image-based information to guide the keypoint-based optimization.

Optimization-Regression hybrids: The idea of building a hybrid between the two paradigms has been explored extensively in recent work. HMR [17] and HUND [50] use a network to mimic the optimization steps and regress the updates to the model parameters. Song *et al.* [43] use the reprojection error of the model joints to guide their learning-based gradient descent approach. SPIN [22] initializes the optimization with a regression network and supervises the network with the output of the optimization. EFT [16] builds on that by updating the network weights during the fitting procedure. Our probabilistic model also investigates this type of collaboration by regressing a distribution of poses which can then be used as a prior term for the fitting.

2.2. Multiple hypotheses for 3D human pose

Multiple hypotheses methods have been used in the context of 3D human pose estimation to deal with the inherent ambiguities of the reconstruction such as occlusions, truncations or depth ambiguities. Jahangiri and Yuille [14] use a compositional model and anatomical constraints to generate multiple hypotheses consistent with 2D keypoint evidence. Li and Lee [27] use a Mixture Density Network instead and generate a fixed number of proposals based on the centroids of the Gaussian kernels, while Sharma *et al.* [42] tackle the same problem using a Conditional VAE. Recently, Biggs *et al.* [3] extend HMR [18] with N prediction heads. This leads to a discrete set of hypotheses, instead of a full probability of poses as we do. In a concurrent work, Sengupta *et al.* [41] use a Gaussian posterior to model the uncertainty in the parameter prediction. Differently from these methods, our approach is not limited to learning a generative model of plausible 3D poses, but rather shows how one can use such a model for useful downstream applications.

2.3. Normalizing Flows

Normalizing Flows are used to represent complex distributions as a series of invertible transformations of a simple base distribution. They were originally developed for modeling posterior distributions for variational inference [40, 20]. Popular examples include MADE [10], NICE [7], MAF [37], RealNVP [8] and Glow [19].

Normalizing Flows have been used in the context of 3D human pose estimation to learn a prior on the distribution of plausible poses [3, 48, 49]. These priors are usually trained using unpaired MoCap data [31]. Our work is fundamentally different from these methods in the sense that we are interested in learning a pose prior *conditioned* on 2D image evidence rather than a generic prior on the 3D pose space.

3. Method

In this Section, we present in detail our proposed approach. We start with an outline of Normalizing Flows [40] and the SMPL body model [30]. Then, we describe the model architecture and the training procedure. Finally, we show how our trained model can be used in downstream applications in a simple and straightforward manner.

3.1. Normalizing Flows

Let $Z \in \mathbb{R}^d$ be a random variable with distribution $p_Z(\mathbf{z})$ and $f : \mathbb{R}^d \rightarrow \mathbb{R}^d$ an invertible mapping. If we transform Z with f , then the resulting random variable $X = f(Z)$ has probability density function:

$$p_X(\mathbf{x}) = p_Z(\mathbf{z}) \left| \det \frac{\partial f}{\partial \mathbf{z}} \right|^{-1} \quad (1)$$

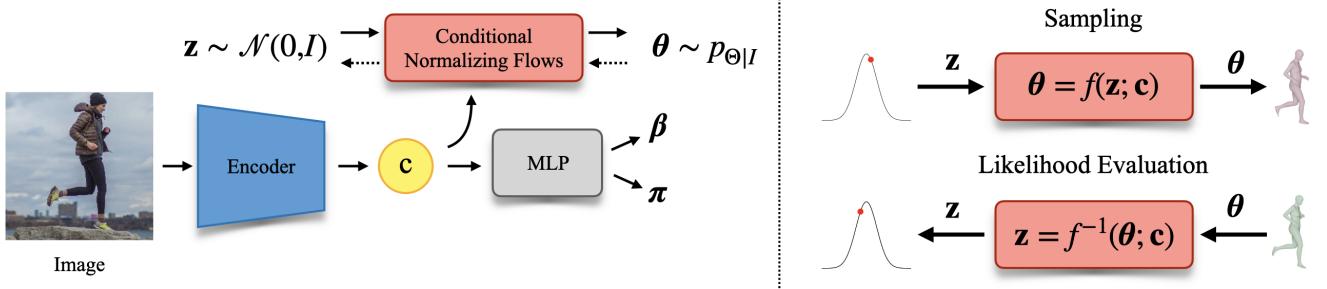


Figure 3: **Architecture of the proposed probabilistic model for human mesh recovery, ProHMR.** Left: Our image encoder regresses a hidden vector \mathbf{c} , which is used as the conditioning input to the flow model. In parallel, it is also decoded to shape parameters β and camera π . Right: Our flow model learns an invertible mapping which allows for two processing directions; depending on the desired function, we can perform both sampling and fast likelihood computation.

Normalizing Flow models are used to model arbitrarily complex distributions as a series of invertible transformations of a simple base distribution. Typically, the base distribution $p_Z(\mathbf{z})$ is chosen to be the standard multivariate Gaussian $\mathcal{N}(\mathbf{0}, I)$. If we write f as a composition of invertible transformations $\{f_k\}_{k=1}^K$ with $Z_0 = Z$, $Z_i = f_i(Z_{i-1})$ and $Z_K = X$, then the log-probability density of X can be computed as:

$$\ln p_X(\mathbf{x}) = \ln p_Z(\mathbf{z}) - \sum_{k=1}^K \ln \left| \det \frac{\partial f_i}{\partial \mathbf{z}_{i-1}} \right|. \quad (2)$$

Winkler *et al.* [46] extended Normalizing Flow models to model conditional distributions $p_{X|Y}(\mathbf{x}|\mathbf{y})$ by using transformations $\mathbf{x} = f(\mathbf{z}; \mathbf{y})$ that are bijective in \mathbf{x} and \mathbf{z} .

3.2. SMPL model

SMPL [30] is a parametric human body model. It defines a mapping $\mathcal{M}(\boldsymbol{\theta}, \boldsymbol{\beta})$ that takes as input a set of pose parameters $\boldsymbol{\theta}$ and shape parameters $\boldsymbol{\beta}$ and outputs a body mesh $M \in \mathbb{R}^{N \times 3}$, where $N = 6890$ is the number of mesh vertices. Additionally, given an output mesh, the body joints J can be expressed as a linear combination of the mesh vertices, $J = WM$, where W is a pretrained linear regressor.

3.3. Model design

Without loss of generality, we present our pipeline for the case where the input is an image of a person and the target output is the set of SMPL body model parameters. We call this model **ProHMR**, with the goal of *Probabilistic Human Mesh Recovery*. At the end of this section we also show how the same method can be applied in alternative scenarios with different input and output representations.

In our setting, we are given an input image I containing a person, and our goal is to learn a distribution of plausible poses for that person conditioned on I . Since we do not have access to accurate pairs of images-shape annotations, we choose to only model the uncertainty of the SMPL pose parameters $\boldsymbol{\theta}$. Our architecture follows closely the HMR

paradigm [17]. The output of our network is the conditional probability distribution $p_{\Theta|I}(\boldsymbol{\theta}|I)$ as well as point estimates for the shape and camera parameters $\boldsymbol{\beta}$ and π respectively.

The complete pipeline is depicted in Figure 3. Given an input image I , we encode it using a CNN g and obtain a context vector $\mathbf{c} = g(I)$. We model $p_{\Theta|I}(\boldsymbol{\theta}|\mathbf{c} = g(I))$ using Conditional Normalizing Flows. We learn a mapping $f : \mathbb{R}^d \times \mathbb{R}^c \rightarrow \mathbb{R}^d$ that is bijective in \mathbf{z} and $\boldsymbol{\theta}$, *i.e.*, $\boldsymbol{\theta} = f(\mathbf{z}; \mathbf{c})$ and $\mathbf{z} = f^{-1}(\boldsymbol{\theta}; \mathbf{c})$.

We employ Normalizing Flows instead of simpler alternatives such as Mixture Density Networks (MDN) [27] because of their expressiveness and ability to model more complex distributions, as we show later in the evaluation section. In our setting, Normalizing Flows have also clear advantages over VAEs, since VAEs do not offer an easy way to compute the likelihood of a given output sample, which is crucial when using our model in downstream tasks.

Our Normalizing Flow model is based on the Glow architecture [19]. Each building block f_i is comprised of 3 basic transformations:

$$f_i = f_{coupl} \circ f_{lin} \circ f_{norm}, \quad (3)$$

where $f_{norm}(\mathbf{z}) = \mathbf{a} \odot \mathbf{z} + \mathbf{b}$ (Instance Normalization), $f_{lin}(\mathbf{z}) = W\mathbf{z} + \mathbf{b}$ (Linear transformation) and $f_{coupl} = [\mathbf{z}_{1:k}, \mathbf{z}_{k+1:d} + \mathbf{t}(\mathbf{z}_{1:d}, \mathbf{c})]$ (Additive coupling). To make the inversion and the Jacobian computation faster, in the linear transformation we parametrize the LU decomposition of W . The final flow model is obtained by composing four of these building blocks.

The selected flow model allows us to perform both fast likelihood computation and fast sampling from the distribution. At the same time, a very important property is that the determinant of the Jacobian does not depend on \mathbf{z} , which in turn means that the mode of the output distribution is:

$$\boldsymbol{\theta}_I^* = \operatorname{argmax}_{\boldsymbol{\theta}} p_{\Theta|I}(\boldsymbol{\theta}|\mathbf{c}) = f(\mathbf{0}; \mathbf{c}). \quad (4)$$

This result allows us to use our model as a *predictive model* in a straightforward way; in the absence of any additional

side-information, we make predictions using the mode of the output distribution.

To regress the camera and the SMPL shape parameters, we use a small MLP h that takes as input the context vector \mathbf{c} and outputs a single point estimate, *i.e.*, $[\beta, \pi] = h(\mathbf{c})$. We also experimented with having β and π depend on θ , but there was no observable improvement.

3.4. Training objective

Let us assume that we have a collection of images paired with SMPL pose annotations. Typically, Normalizing Flow models are trained to minimize the negative log-likelihood of the ground truth examples θ_{gt} , *i.e.* the loss function is:

$$L_{nll} = -\ln p_{\Theta|I}(\theta_{gt}|\mathbf{c}). \quad (5)$$

However, for the task of 3D pose estimation, 3D annotations are generally not available except for a small number of indoor datasets captured in constrained studio environments [13, 33] and methods trained on those datasets fail to generalize in challenging in-the-wild scenes. Consequently, previous methods like [17] propose to use examples with only 2D keypoint annotations and minimize the keypoint reprojection loss jointly with an adversarial prior. To make such a mixed training possible within our framework, we propose to minimize the expectation of the above error with respect to the learned distribution, *i.e.*,

$$L_{exp} = \mathbb{E}_{\theta \sim p_{\Theta|I}}[L_{2D}(\theta, \beta, \pi) + L_{adv}(\theta, \beta)]. \quad (6)$$

To make this loss differentiable we use the *Law of the Unconscious Statistician* and rewrite the expectation as:

$$L_{exp} = \mathbb{E}_{\mathbf{z} \sim p_Z}[L_{2D}(f(\mathbf{z}; \mathbf{c}), \beta, \pi) + L_{adv}(f(\mathbf{z}; \mathbf{c}), \beta)]. \quad (7)$$

Conceptually, even though we do not have ground truth annotations, to maximize the conditional probability of these examples we can still constrain the form of the output distribution by forcing the output samples to have low reprojection error on average and lie on the manifold of valid poses. As in the case of VAEs [21], we approximate the expectation by drawing a single sample from the prior.

As mentioned previously, our goal is to use our model not only as a generative model but also as a predictive model. Thus, we propose to exploit the property that for each image I , the mode θ_I^* of the output distribution corresponds to the transformation of $\mathbf{z} = \mathbf{0}$. We do this by explicitly supervising θ_I^* with all the available annotations as in a standard regression framework and minimize:

$$L_{mode} = L_{3D}(\theta_I^*, \beta) + L_{2D}(\theta_I^*, \beta, \pi) + L_{adv}(\theta_I^*, \beta), \quad (8)$$

where L_{3D} is the loss on the available 3D annotations (3D joints and/or SMPL parameters) whenever they are available. As we show in the experimental section, this explicit

supervision of the mode of the output distribution helps boost the performance of our model in predictive tasks.

It is important to mention that L_{exp} is not redundant in the presence of L_{mode} ; the behavior of the mode is not indicative of the full distribution, whereas L_{exp} encourages the distribution to have certain desirable properties.

Finally, for modeling rotations we use the 6D representation proposed in [52]. One issue with this particular representation is that it is not unique. For example, for any 3D vectors x and y , $[x, y]$ and $[\alpha x + \gamma y, \beta x + \gamma y]$ are mapped to the same rotation matrix. Empirically we found that putting no constraints on the 6D representation results in large discrepancy between examples with full 3D SMPL parameter supervision and examples with only 2D keypoint annotations. Among other things, this caused mode collapse for the examples without 3D ground truth. Thus, we introduce another loss function L_{orth} that forces the 6D representations of the samples drawn from the distribution to be close to the orthonormal 6D representation.

Eventually, the final training objective becomes:

$$L = \lambda_{nll}L_{nll} + \lambda_{exp}L_{exp} + \lambda_{mode}L_{mode} + \lambda_{orth}L_{orth}. \quad (9)$$

3.5. Downstream applications

In this part we show how our learned conditional distribution can be used in a series of downstream applications. We highlight that all these applications refer to *test-time* processing with the same trained model without any special per-task training. Examples of such tasks are shown in Figure 2. These applications fall under the more general umbrella of Maximum a Posteriori estimation where we use all available evidence to make more informed predictions.

3D pose regression As already discussed in previous sections, we can use our model in conventional tasks such as 3D pose regression from a single image. In the absence of additional evidence, the most appropriate choice for making predictions is to pick the mode θ_I^* of the distribution.

Body model fitting SMPLify [4] is a popular method that fits the SMPL body model to a set of 2D keypoints using a traditional optimization approach. The objective is:

$$\lambda_J E_J + \lambda_\theta E_\theta + \lambda_\alpha E_\alpha + \lambda_\beta E_\beta, \quad (10)$$

where E_J penalizes the weighted 2D distance between the projected model joints and the detected joints, E_θ is a Mixture of Gaussian 3D pose prior, E_α is a pose prior penalizing unnatural rotations of elbows and knees and E_β is a quadratic penalty on the shape coefficients.

Fitting a parametric body model to 2D image landmarks is a very challenging and inherently ambiguous problem. The data term E_J is purely driven by the 2D keypoints and

disregards rich information contained in the input image. SPIN [22] partially addresses this issue by using an image-based regression network that provides a good initialization for the optimization, helping the fitting to converge to a better minimum. However, the image information is only used in the initialization phase, as SMPLify does not incorporate explicit image-specific priors that prevent the pose to deviate arbitrarily far from the set of plausible poses for the given image. The drifting problem is also an important limitation of [16], forcing the approach to rely on good initialization and carefully chosen stopping criteria.

Motivated by these limitations, we propose to replace the weaker generic 3D priors E_θ and E_α with an explicit pose prior $E_{\theta|I} = -\ln p_{\Theta|I}(\theta|\mathbf{c})$ that models the likelihood of a given pose conditioned on the image evidence. Thus, the final optimization objective becomes:

$$\lambda_J E_J - \ln p_{\Theta|I}(\theta|\mathbf{c}) + \lambda_\beta E_\beta. \quad (11)$$

As initialization for the fitting we use the mode θ_I^* of the conditional distribution. In the experimental section we show that by using this learned image-based prior we are able to consistently improve the fitting results, both qualitatively and quantitatively, as reflected in the 3D metrics.

Multiple views fusion Although our model has been trained for single-image reconstruction, we can still use the learned conditional distribution to obtain refined pose estimates in the presence of multiple views of a person. Let us assume that we have a set $\{I_n\}_1^N$ of uncalibrated views of the same subject. We partition the pose vector of each frame as $\theta_n = (\theta_n^g, \theta_n^b)$ where θ_n^g corresponds to the global rotation of the model and θ_n^b is the body pose. We propose to refine the pose by minimizing the following objective:

$$-\sum_{n=1}^N \ln p(\theta_n|\mathbf{c}_n) + \lambda \sum_{n=1}^N \|\theta_n^b - \bar{\theta}^b\|_2^2, \quad (12)$$

where $\bar{\theta}^b = \frac{1}{N} \sum_{n=1}^N \theta_n^b$. The second term of the objective is equivalent to minimizing the squared distance between all pairs of poses.

3.6. Additional details

ProHMR. Following previous works [17, 22] we use ResNet-50 [12] as the encoder. For the Normalizing Flows we use 4 building blocks f_i . For more details about the architecture, datasets and the training hyperparameters we refer the reader to the supplementary material.

2D pose lifting. Complementary to ProHMR, we use our approach to lift 2D poses to 3D skeletons, as in Martinez *et al.* [32]. We use the same Normalizing Flow architecture as in ProHMR. In this case the input is a set of 2D Hourglass detections [35] and the output is the 3D pose coordinates.

For the encoder g , instead of a CNN, we use the backbone from [32]. Since all examples have full 3D supervision, our training objective consists only of L_{nll} and L_{mode} .

Downstream tasks. For the fitting procedure employed in the downstream tasks, we found it beneficial to perform the optimization in the latent space instead of the pose space directly (similarly to SMPLify-X [38]). Thus, we leave \mathbf{z} as a free variable and decode it into the pose vector $\theta = f(\mathbf{z}; \mathbf{c})$. Also, since for our Normalizing Flow model the determinant of the Jacobian does not depend on \mathbf{z} , the likelihood term becomes $\ln p(\theta|\mathbf{c}) = -\|\mathbf{z}\|_2^2 + \text{const}$.

4. Experimental evaluation

In this Section we present the experimental evaluation of our approach. First we provide an outline of the datasets used for training and evaluation and then we will present detailed quantitative and qualitative evaluation results.

4.1. Datasets

We report results on Human3.6M [13], MPI-INF-3DHP [33], 3DPW [45] and Mannequin Challenge [28], where we use the annotations produced by Leroy *et al.* [26]. For training, we use datasets with 3D ground truth (Human3.6M [13] and MPI-INF-3DHP [33]), as well as datasets with 2D keypoint annotations (COCO [29] and MPII [1]) augmented with pseudo ground truth SMPL parameters from SPIN [22], whenever they are available.

4.2. Quantitative evaluation

In this part we evaluate different aspects of our proposed approach. We compare the predictive accuracy of our model with standard regression methods and show that it achieves comparable performance with the state of the art in human mesh recovery. We also benchmark the generative capabilities of our method in multiple hypotheses scenarios, where we outperform previous approaches. Finally, we demonstrate that our learned image-conditioned prior can boost the performance in downstream applications such as model fitting and multi-view refinement.

Human mesh recovery. First, we focus on the predictive performance of our model, comparing it against other state-of-the-art methods that regress SMPL body model parameters. For the evaluation of ProHMR, we use the mode θ_I^* of the learned distribution. For Biggs *et al.* [3] we report the metrics after quantizing to $n = 1$ sample. Based on the results of Table 1, using ProHMR as a regressor, leads to performance comparable to the state of the art. This shows that we can indeed recast the problem from point to density estimation without any significant loss in performance.

Multiple hypotheses. Next, we compare the representational power of ProHMR with different multiple hypotheses baselines, including Biggs *et al.* [3], as well as the MDN



Figure 4: **Samples from the learned distribution.** Pink colored mesh corresponds to the mode.

	3DPW	H36M	MPI-INF-3DHP
HMR [17]	81.3	56.8	89.8
SPIN [22]	59.1	41.1	67.5
Biggs <i>et al.</i> [3]	59.9	41.6	N/A
ProHMR	59.8	41.2	65.0

Table 1: **Evaluation on human mesh recovery.** Our model achieves accuracy comparable with the state of the art. Numbers reported are PA-MPJPE in mm.

	$n = 5$		$n = 10$		$n = 25$		min	
	[45]	[13]	[45]	[13]	[45]	[13]	[45]	[13]
[3] (MDN)	61.2	43.3	60.7	43.0	60.1	42.7	60.1	42.7
[3] (CVAE)	60.7	46.4	60.5	46.3	60.3	46.2	60.3	46.2
[3] (NF)	57.1	42.0	56.6	42.2	55.6	42.2	55.6	41.6
ProHMR	56.5	39.4	54.6	38.3	52.4	36.8	40.8	29.9

Table 2: **Multiple hypotheses evaluation.** Numbers are PA-MPJPE in mm. We report errors for small n and the *minimum* error over samples drawn from the distribution.

and Conditional VAE variants explored in the same paper. Following [3], we report results for small sample sizes n . Since we are interested in measuring the representational power of the learned distribution, we also compare the minimum 3D pose error of samples drawn from each distribution as proposed in [42]. We present the detailed results for Human3.6M and 3DPW in Table 2.

Model fitting. In this part we evaluate the accuracy of different methods that fit the SMPL body model to a set of 2D keypoints. The body model fitting baselines we compare include the standard SMPLify [4, 38], EFT [16], and our proposed fitting with the learned image-conditioned prior. For both SMPLify and EFT we use publicly available implementations and initialize the fitting process with SPIN, while for SMPLify we use two different versions for the pose prior, GMM [4] and VPoser [38]. For a fair evaluation

	3DPW	H36M (OP)	H36M (GT)
SPIN [22]	59.2	41.8	41.8
SPIN+SMPLify (GMM) [4]	66.5	54.6	43.3
SPIN+SMPLify (VPoser) [38]	70.9	53.5	39.9
SPIN+EFT [16]	56.6	41.6	38.7
ProHMR	59.8	41.2	41.2
ProHMR + fitting	55.1	39.3	34.8

Table 3: **Evaluation of different model fitting methods.** The fitting algorithms are initialized by the corresponding regression results. All numbers are PA-MPJPE in mm.

of the performance benefit, we compare methods that are trained on the same datasets and have similar regression performance. The results are presented in Table 3. While performing SMPLify on top of regression improves the model-image alignment, it increases the 3D pose errors, especially when using OpenPose detections [5]. We hypothesize that this happens because of the generic 3D pose prior terms of SMPLify. EFT on top of regression improves the 3D pose metrics, however our method manages to push the accuracy even further. In 3DPW our approach has a 4.7mm relative error improvement vs. 2.6mm for EFT, while if we use the ground truth 2D keypoints in Human3.6M we get a 6.3mm improvement vs 3.1mm for EFT.

Multi-view refinement. We evaluate the effect of our learned image-conditioned prior at refining the pose predictions in uncalibrated multi-view scenarios. For benchmarking, we use Human3.6M and the more challenging Mannequin Challenge dataset. We compare our fitting-based method against the individual per-view predictions and a baseline that performs rotation averaging in Table 4. For the rotation averaging we first average the per-view rotation matrices and then project them back to $SO(3)$ using SVD.

Ablation study. We also assess the significance of the term



Figure 5: **Model fitting results.** Pink: Regression. Green: ProHMR + fitting. Grey: Regression + SMPLify

	H36M		Mannequin	
	MPJPE	PA-MPJPE	MPJPE	PA-MPJPE
ProHMR	65.1	43.7	176.0	91.9
ProHMR + rot avg	64.8	35.2	174.4	85.1
ProHMR + fitting	62.2	34.5	171.3	83.9

Table 4: **Evaluation of multi-view refinement.** We compare single-image 3D reconstruction with a baseline refinement using rotation averaging and the proposed optimization-based refinement scheme.

	3DPW	H36M	MPI-INF-3DHP
ProHMR (w/o L_{mode})	67.4	54.8	76.5
ProHMR	59.8	41.2	65.0

Table 5: **Ablation for L_{mode} .** Numbers are PA-MPJPE.

L_{mode} that we use to explicitly supervise the mode of the learned distribution. We report results for training ProHMR with and without this loss in Table 5. We can see that including L_{mode} is crucial to achieve competitive performance in conventional regression tasks.

Additional evaluations. Finally, we show that the proposed modeling is general enough to handle different input and output representations. Here, we consider the setting of lifting a 2D pose input to a 3D skeleton output [32] and present results in Table 6. Our model performs on par with an equivalent regression approach [32], while it outperforms the MDN method of Li and Lee [27].

4.3. Qualitative results

In Figure 4 we show sample reconstructions of our method. Additionally, in Figure 5 we show comparisons of our model fitting approach with SMPLify. Our method produces more realistic reconstructions overall, particularly in

	MPJPE	PA-MPJPE
Martinez <i>et al.</i> [32]	62.9	47.7
Li and Lee [27] (mode)	64.5	47.8
Ours	62.9	47.6
Li and Lee [27] (min)	42.6	34.4
Ours (min)	42.4	32.9

Table 6: **Evaluation of 3D pose accuracy for skeleton-based 2D pose lifting** on Human3.6M. Top: Regression accuracy. Bottom: Minimum error of the distributions.

cases where there are missing or very low confidence key-point detections. In cases like that (*e.g.*, example of last row), our image-based prior, unlike SMPLify, does not let the pose deviate far from the image evidence.

5. Summary

This work presents a probabilistic model for 3D human mesh recovery from 2D evidence. Unlike most approaches that output a single point estimate for the 3D pose, we propose to learn a mapping from the input to a distribution of plausible poses. We model this distribution using Conditional Normalizing Flows. Our probabilistic model allows for sampling of diverse outputs, efficient computation of the likelihood of each sample, and a fast and closed-form solution for the mode. We demonstrate the effectiveness of our method with empirical results in several benchmarks. Future work could consider extending our approach to other classes of articulated or non-articulated objects and potentially model other ambiguities like the depth-size trade-off. **Acknowledgements:** Research was sponsored by the following grants: ARO W911NF-20-1-0080, NSF IIS 1703319, NSF TRIPODS 1934960, NSF CPS 2038873, ONR N00014-17-1-2093, the DARPA-SRC C-BRIC, and by Honda Research Institute. GP is supported by BAIR sponsors.

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. 6
- [2] Anurag Arnab, Carl Doersch, and Andrew Zisserman. Exploiting temporal context for 3D human pose estimation in the wild. In *CVPR*, 2019. 3
- [3] Benjamin Biggs, Sébastien Ehrhart, Hanbyul Joo, Benjamin Graham, Andrea Vedaldi, and David Novotny. 3D multibodies: Fitting sets of plausible 3D models to ambiguous image data. In *NeurIPS*, 2020. 1, 3, 6, 7
- [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. 1, 2, 3, 5, 7
- [5] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. *PAMI*, 43(1):172–186, 2019. 7
- [6] Vasileios Choutas, Georgios Pavlakos, Timo Bolkart, Dimitrios Tzionas, and Michael J Black. Monocular expressive body regression through body-driven attention. In *ECCV*, 2020. 3
- [7] Laurent Dinh, David Krueger, and Yoshua Bengio. NICE: non-linear independent components estimation. In *ICLR*, 2015. 3
- [8] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real NVP. In *ICLR*, 2017. 3
- [9] Georgios Georgakis, Ren Li, Srikrishna Karanam, Terrence Chen, Jana Košeká, and Ziyang Wu. Hierarchical kinematic human mesh recovery. In *ECCV*, 2020. 3
- [10] Mathieu Germain, Karol Gregor, Iain Murray, and Hugo Larochelle. MADE: masked autoencoder for distribution estimation. In *ICML*, 2015. 3
- [11] Riza Alp Guler and Iasonas Kokkinos. HoloPose: Holistic 3D human reconstruction in-the-wild. In *CVPR*, 2019. 3
- [12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 6
- [13] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. *PAMI*, 36(7):1325–1339, 2014. 5, 6, 7
- [14] Ehsan Jahangiri and Alan L Yuille. Generating multiple diverse hypotheses for human 3D pose consistent with 2D joint detections. In *ICCVW*, 2017. 3
- [15] Wen Jiang, Nikos Kolotouros, Georgios Pavlakos, Xiaowei Zhou, and Kostas Daniilidis. Coherent reconstruction of multiple humans from a single image. In *CVPR*, 2020. 3
- [16] 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. 3, 6, 7
- [17] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *CVPR*, 2018. 1, 2, 3, 4, 5, 6, 7
- [18] Angjoo Kanazawa, Shubham Tulsiani, Alexei A Efros, and Jitendra Malik. Learning category-specific mesh reconstruction from image collections. In *ECCV*, 2018. 3
- [19] Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In *NeurIPS*, 2018. 3, 4
- [20] Durk P Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling. Improved variational inference with inverse autoregressive flow. In *NIPS*, 2016. 3
- [21] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *ICLR*, 2014. 5
- [22] 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. 1, 3, 6, 7
- [23] Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. In *CVPR*, 2019. 1, 3
- [24] Christoph Lassner, Javier Romero, Martin Kiefel, Federica Bogo, Michael J Black, and Peter V Gehler. Unite the people: Closing the loop between 3D and 2D human representations. In *CVPR*, 2017. 3
- [25] Hsi-Jian Lee and Zen Chen. Determination of 3D human body postures from a single view. *CVIU*, 30(2):148–168, 1985. 1
- [26] Vincent Leroy, Philippe Weinzaepfel, Romain Brégier, Hadrien Combaluzier, and Grégory Rogez. SMPLy benchmarking 3D human pose estimation in the wild. In *3DV*, 2020. 6
- [27] Chen Li and Gim Hee Lee. Generating multiple hypotheses for 3D human pose estimation with mixture density network. In *CVPR*, 2019. 1, 3, 4, 8
- [28] Zhengqi Li, Tali Dekel, Forrester Cole, Richard Tucker, Noah Snavely, Ce Liu, and William T. Freeman. Learning the depths of moving people by watching frozen people. In *CVPR*, 2019. 6
- [29] 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. 6
- [30] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. SMPL: A skinned multi-person linear model. *ACM Transactions on Graphics (TOG)*, 34(6):248, 2015. 3, 4
- [31] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *ICCV*, 2019. 3
- [32] Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little. A simple yet effective baseline for 3D human pose estimation. In *ICCV*, 2017. 1, 2, 3, 6, 8
- [33] 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. 5, 6
- [34] Lea Muller, Ahmed AA Osman, Siyu Tang, Chun-Hao P Huang, and Michael J Black. On self-contact and human pose. In *CVPR*, 2021. 3

- [35] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hour-glass networks for human pose estimation. In *ECCV*, 2016. 6
- [36] Ahmed AA Osman, Timo Bolkart, and Michael J Black. STAR: Sparse trained articulated human body regressor. In *ECCV*, 2020. 3
- [37] George Papamakarios, Theo Pavlakou, and Iain Murray. Masked autoregressive flow for density estimation. In *NIPS*, 2017. 3
- [38] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3D hands, face, and body from a single image. In *CVPR*, 2019. 2, 3, 6, 7
- [39] Georgios Pavlakos, Nikos Kolotouros, and Kostas Daniilidis. TexturePose: Supervising human mesh estimation with texture consistency. In *ICCV*, 2019. 3
- [40] Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *JMLR*, 2015. 3
- [41] Akash Sengupta, Ignas Budvytis, and Roberto Cipolla. Probabilistic 3D human shape and pose estimation from multiple unconstrained images in the wild. In *CVPR*, 2021. 3
- [42] Saurabh Sharma, Pavan Teja Varigonda, Prashast Bindal, Abhishek Sharma, and Arjun Jain. Monocular 3D human pose estimation by generation and ordinal ranking. In *ICCV*, 2019. 3, 7
- [43] Jie Song, Xu Chen, and Otmar Hilliges. Human body model fitting by learned gradient descent. In *ECCV*, 2020. 1, 3
- [44] Omid Taheri, Nima Ghorbani, Michael J. Black, and Dimitrios Tzionas. GRAB: A dataset of whole-body human grasping of objects. In *ECCV*, 2020. 3
- [45] 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 *ECCV*, 2018. 6, 7
- [46] Christina Winkler, Daniel Worrall, Emiel Hoogeboom, and Max Welling. Learning likelihoods with conditional normalizing flows. *arXiv preprint arXiv:1912.00042*, 2019. 4
- [47] Donglai Xiang, Hanbyul Joo, and Yaser Sheikh. Monocular total capture: Posing face, body, and hands in the wild. In *CVPR*, 2019. 3
- [48] Hongyi Xu, Eduard Gabriel Bazavan, Andrei Zanfir, William T Freeman, Rahul Sukthankar, and Cristian Sminchisescu. GHUM & GHUML: Generative 3D human shape and articulated pose models. In *CVPR*, 2021. 3
- [49] Andrei Zanfir, Eduard Gabriel Bazavan, Hongyi Xu, William T Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Weakly supervised 3D human pose and shape reconstruction with normalizing flows. In *ECCV*, 2020. 3
- [50] Andrei Zanfir, Eduard Gabriel Bazavan, Mihai Zanfir, William T. Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Neural descent for visual 3D human pose and shape. In *CVPR*, 2021. 3
- [51] Ce Zheng, Wenhan Wu, Taojiannan Yang, Sijie Zhu, Chen Chen, Ruixu Liu, Ju Shen, Nasser Kehtarnavaz, and Mubarak Shah. Deep learning-based human pose estimation: A survey. *arXiv preprint arXiv:2012.13392*, 2020. 3
- [52] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *CVPR*, 2019. 5