CondiMen: Conditional Multi-Person Mesh Recovery

Romain Brégier[1] Fabien Baradel[1] Thomas Lucas[1] Salma Galaaoui[1,2]
Matthieu Armando[1] Philippe Weinzaepfel[1] Grégory Rogez[1]
[1] NAVER LABS Europe
[2]LIGM, École des Ponts, Univ Gustave Eiffel, CNRS, Marne-la-Vallée, France

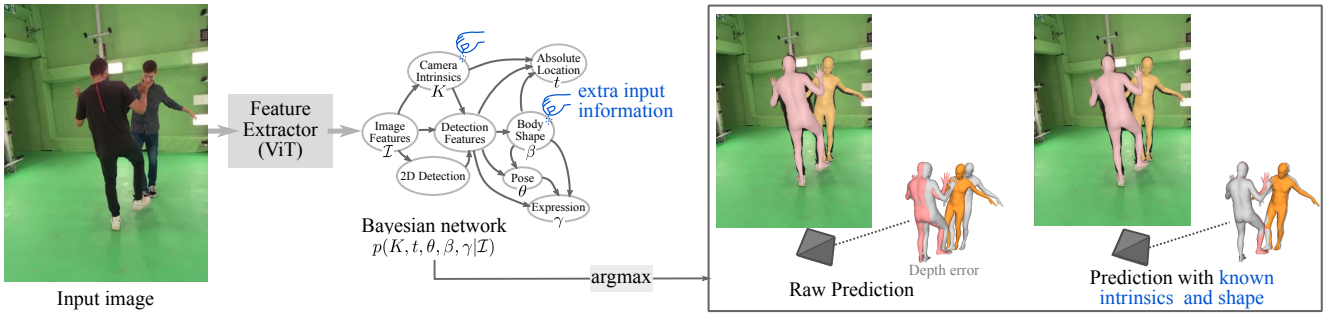


Figure 1. **CondiMen: A Recipe for HMR.** Recovering 3D human meshes from an image is challenging as predictions that look plausible in 2D can be inaccurate in 3D (ground-truth meshes in gray). To improve predictions, we propose a method leveraging additional information that may be available – such as camera calibration, body shape, distance to the camera, or multi-view observations. We decompose mesh recovery into a human detection and attribute estimation problem, modeling the joint probability of these attributes (pose, body shape, *etc*.) using a Bayesian network. It enables efficient inference, compatible with 20FPS real-time applications.

Abstract

Multi-person human mesh recovery (HMR) consists in detecting all individuals in a given input image, and predicting the body shape, pose, and 3D location for each detected person. The dominant approaches to this task rely on neural networks trained to output a single prediction for each detected individual. In contrast, we propose CondiMen, a method that outputs a joint parametric distribution over likely poses, body shapes, intrinsics and distances to the camera, using a Bayesian network. This approach offers several advantages. First, a probability distribution can handle some inherent ambiguities of this task – such as the uncertainty between a person's size and their distance to the camera, or more generally the loss of information that occurs when projecting 3D data onto a 2D image. Second, the output distribution can be combined with additional information to produce better predictions, by using e.g. *known camera or body shape parameters, or by exploiting multi-view observations. Third, one can efficiently extract the most likely predictions from this output distribution, making the proposed approach suitable for real-time applications. Empirically we find that our model i) achieves performance on par with or better than the state-of-the-art, ii) captures uncertainties and correlations inherent in pose estimation and iii) can exploit additional information at test time, such as multi-view consistency or body shape priors. CondiMen spices up the modeling of ambiguity, using just the right ingredients on hand.*

1. Introduction

Recovering people characteristics in 3D from images is essential for a variety of applications, ranging from human behavior analysis to robotic systems in crowded environments. In this work, we present a method that detects individuals in images and then predicts 3D meshes, encoding their pose, body shape, and location.

Human mesh recovery is an ill-posed problem, as different meshes could be plausible for a single input image due to factors like clothing, occlusions, and the projective nature of 2D imaging. In particular, the apparent 2D size of a person in an image depends on their actual size in 3D, the distance to the camera, and the camera's focal length (see supp. mat. for examples). Despite this uncertainty, state-of-the-art methods [1, 28, 58, 59] typically predict deterministic outputs. These methods are optimized to minimize an empirical loss over the training data, and in the presence

of ambiguity, they tend to predict average attributes – those that occur most frequently in the training data – resulting in a loss of accuracy when faced with aleatoric uncertainty.

A probabilistic framework offers an elegant way to handle such uncertainty. Methods that predict probability distributions – allowing for the sampling of hypotheses [31, 32, 40] – or that provide confidence estimates [14, 31, 55] have been proposed in previous work. However, most of them focus solely on relative pose estimation and fail to consider the body shape or 3D location, which are critical in many applications. Moreover, there are scenarios where additional information such as camera calibration, body shapes previously scanned, or multi-view image captures are available and could provide useful cues to improve predictions. Incorporating such external inputs into existing methods is unfortunately challenging and often requires iterative optimization techniques [31], limiting their practical application.

In this paper, we propose a solution to this issue by treating mesh recovery as a task of jointly predicting various attributes (pose, body shape, location, *etc*.), and by regressing a joint probability distribution over these attributes. Specifically, we introduce a parametric Bayesian model [3] illustrated in Fig. 1 that, given an input image, outputs a joint probability distribution over camera intrinsics, human detections, poses, body shapes and 3D locations. This probabilistic formulation accounts for ambiguities of the multi-person mesh recovery tasks and allows for the efficient use of external information during inference. Additionally, multi-view images can seamlessly be used within our framework to improve 3D mesh recovery – something which is not straightforward with monocular deterministic methods. This ability to make multi-view predictions at test time, despite being trained on monocular data only is advantageous because high-quality, diverse monocular data is far more abundant and easier to collect than multi-view data, which often requires complex setups.

For our experiments, we train a model from synthetic-only data, using BEDLAM [4] as well as images we generated to further increase data variability. We demonstrate that the Bayesian nature of our model offers flexibility and the ability to exploit available external knowledge, at test time. For instance, we show that leveraging known ground-truth quantities (*e.g*. camera intrinsics or body shape) within the model can improve predictions in a zero-shot manner. It can also be used to exploit consistency across multiple input images, for instance by enforcing a constant body shape for a given person. We validate the performance of our approach against existing state-of-the-art methods on monocular and multi-view datasets [22, 23, 38, 62, 70], showing competitive results and the ability to handle uncertainty effectively.

2. Related work

Human Mesh Recovery from a Single Image. The development of parametric 3D models [37, 44, 67] has advanced the field of human pose estimation from 3D skeleton regression [9, 48, 65] to the prediction of full human body meshes [15, 19, 27, 28, 30, 35, 50, 73]. More recently, research has focused on estimating expressive human meshes [1, 4, 20, 42, 57] and placing these meshes within real-world coordinate systems [54, 63]. The pioneering work by Kanazawa et al. [28] introduces a method to predict SMPL [37] parameters, along with weak-perspective re-projection parameters, from a single cropped image of a person. Subsequent methods have improved on this approach, either by refining the network architecture [19, 35, 73] or by leveraging new training data and protocols [4, 42]. Recent advances have also enabled whole-body pose estimation, *i.e*., including facial expression and hand poses [8, 11, 17, 20, 36, 39, 43, 49, 74, 78]. A few methods now tackle the detection and regression of multiple human meshes within a single network [1, 47, 57–59], though these approaches typically produce deterministic outputs. In contrast, our proposed Bayesian network for multi-person whole-body mesh recovery outputs a probability distribution, allowing it to handle ambiguities, integrate external information, and fuse multi-view predictions for more robust performance in complex scenarios.

Multi-view Human Mesh Recovery A first category of methods to tackle multi-view mesh recovery assumes known camera calibration to extend single view reconstruction to multi-view settings [5, 12, 76, 79]. In particular, SMPLify-X [5] performs 3D reconstruction in a unified coordinate system by iteratively minimizing 2D keypoint reprojection errors. Some learning-based methods also follow this calibrated approach [10, 24, 60, 68]. For example, Iskakov et al. [24] propose a learnable triangulation solution. A second category of methods addresses the uncalibrated setup, where predictions from multiple views are used to estimate camera parameters and merged either through handcrafted averaging [34, 45, 71] or learning-based techniques [46, 79]. Recent adaptive frameworks can handle both calibrated and uncalibrated settings but are still limited to single-person [26, 66]. In contrast, our approach supports multi-person mesh recovery across multiple views.

Probabilistic Human 3D Pose. Early methods addressed pose uncertainty through optimization-based formulations [55, 56] or by predicting multiple 3D poses from 2D cues [25, 32, 75]. Li and Lee [32] use a Mixture Density Network to infer a distribution of 3D joints from 2D joints, while Zhang et al. [75] model separate Gaussian distributions for 2D keypoints and depth. Recent works leverage Normalizing Flow (NF) to model pose or shape probability distributions [31, 33, 53, 64], with Kolotouros et al. [31]

conditioning NF on image features to predict SMPL parameters. Another line of research leverages denoising diffusion models [21, 40, 72] that can account for uncertainties in depth, body shape, and camera intrinsics. Extracting the most likely prediction with such approach is often computationally intensive however. Other strategies include using multiple prediction heads with a *best-of-N* loss [2], estimating Gaussian distributions over SMPL parameters [52], or quantizing human mesh representations [18]. Sengupta et al. [51] use a hierarchical matrix-Fisher distribution for each SMPL rotation parameters following the kinematic tree. In contrast, we introduce a Bayesian network head that can handle uncertainty while allowing for efficient inference, use of external information, and multi-view predictions in a simple and elegant way.

3. Method

We present our method for detecting people and estimating their 3D whole-body attributes from a single input image. This method called *CondiMen* (short for *conditional multi-person mesh recovery*) is illustrated in Fig. 1. We extract image features using a Vision Transformer (ViT) backbone [13], which we use as conditioning variables in a joint probability distribution that models people appearing in the image with their different attributes (3D location, body shape, *etc*.). We model this joint distribution as a trained Bayesian network. At inference, we efficiently detect humans and predict their attributes, by sequentially extracting modes of the conditional distributions. This probabilistic framework allows us to leverage different information available at inference, such as known camera intrinsics or specific body shapes to enhance prediction accuracy.

3.1. Problem formulation

Human parametrization. To encode human meshes, we rely on a parametric body model, namely SMPL-X [11]. SMPL-X provides a whole-body parametrization decoupled into an absolute 3D location t, a list of bone orientations θ modeling the pose, a vector β modeling the body shape and a vector γ modeling the facial expression.

Bayesian network. We model the multi-person mesh recovery problem as a probabilistic optimization problem. Given some input image features $\mathcal{I}$, we aim at predicting the value of different random variables: the intrinsic parameters K of the camera, as well as attributes of people visible in the image $(t, \theta, \beta, \gamma)$[1]. We consider the joint probability distribution of these variables conditioned on image features, and aim to extract the most likely prediction:

$$\hat{K}, \hat{t}, \hat{\theta}, \hat{\beta}, \hat{\gamma} = \text{argmax } p(K, t, \theta, \beta, \gamma | \mathcal{I}), \quad (1)$$

where $p(K, t, \theta, \beta, \gamma|\mathcal{I})$ denotes the associated probability density. While conceptually appealing, modeling such joint distribution from limited data is challenging due to the high dimension of the representation space ($\dim(K)$=3, $\dim(t)$=3, $\dim(\beta)$=11, $\dim(d)$=1, $\dim(\theta)$=53×3, $\dim(\gamma)$=10 in our setting).

A classical solution to this *curse of dimensionality* is to adopt the naive Bayes assumption, which posits that different variables are conditionally independent given the image features. This results in a factorized form: $p(K, t, \theta, \beta, \gamma|\mathcal{I}) \approx p(K|\mathcal{I}) \cdot p(t|\mathcal{I}) \cdot p(\theta|\mathcal{I}) \cdot p(\beta|\mathcal{I}) \cdot p(\gamma|\mathcal{I})$, where $p(x|y)$ denotes probability density at x conditioned on the value y. In practice, each conditional density $p(\cdot|\mathcal{I})$ is generally assumed to belong to a parametric family, with parameters that are functions of the input $\mathcal{I}$, *e.g*. regressed using a neural network. Multi-HMR [1] and most other deterministic methods [47, 58, 59] can be thought of as special cases within this framework. They use regression objectives equivalent to a naive Bayes formulation, assuming probability distributions with constant dispersion terms. However, a key limitation of the naive Bayes assumption is that it ignores inter-variable dependencies. These dependencies are often crucial for scene understanding: for instance a small person A appearing the same size in 2D as a taller person B is likely to be closer to the camera than B, other things being equal.

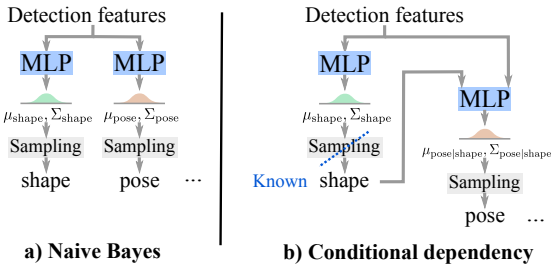
To address these challenges, we adopt a relaxed hypothesis by modeling the joint distribution as a Bayesian network, decoupling variables into a directed acyclic graph of conditional distributions, illustrated in Fig. 1. The ability of deep neural networks to model complex, high-dimensional data in an auto-regressive manner has been demonstrated across various domains including text [7, 61] and images [16]. Our Bayesian model is inspired by auto-regressive approaches, encoding relationships between variables in a cascaded fashion.

3.2. Conditional distributions

We implement our Bayesian network using a cascade of Multi-Layer Perceptrons (MLP). Each MLP outputs parameters of a probability distribution associated with a random variable of our mesh recovery problem, such as the pose of a detected person. This distribution is conditioned on the value of its parent variables in the Bayesian network's graphical model, which allows us to, for example, encode the likelihood of a body pose given a specific body shape and image detection features (see Fig. 1 and Fig. 2). In this section, we present the variables and distributions considered in our experiments.

Camera intrinsics. We assume a pinhole camera model with focal length f>0 and principal point $p \in \mathbb{R}^2$. We predict the parameters of Gaussian distributions for $\ln(f)$ and p , conditioned on image features, specifically on the [CLS] token output by the ViT image encoder. Considering $\ln(f)$ instead of f ensures that $f > 0$, and is mathematically

[1]We describe the single-person case here to simplify notations.



a) Naive Bayes **b) Conditional dependency**

Figure 2. **Modeling conditional dependency**. We predict probability distributions for different human attributes (e.g., pose, shape, distance) and efficiently sample the most likely predictions. Rather than treating each attribute independently (a), we capture their interdependency by modeling conditional distributions (b), leading to more coherent results. Our framework can also incorporate additional input information when available (such as a known shape attribute, shown in blue), to further improve prediction accuracy.

equivalent to modeling f with a log-normal distribution.

2D detection. Image features produced by the ViT encoder consist of patch tokens $P_{u,v}$ defined along a 2D regular grid $G = \{(u,v)\}_{u=1\dots w, v=1\dots h}$. We encode people detections as binary variables $s_{u,v}$ along this grid, which indicate whether a reference keypoint of a person projects into a grid cell $(u,v) \in G$, following the CenterNet object detection framework [77]. For each variable, we predict a score $p(s_{u,v}|\mathcal{I})$ encoding the detection likelihood, which is regressed from the corresponding patch features $P_{u,v}$. In practice, we use the person's head as reference keypoint (see Fig. 6a), assuming that at most one person is detected in each grid cell. At inference, detection is performed via score thresholding and local non-maxima suppression.

Detection features. For each detected person, we consider a latent variable consisting of image patch features $P_{u,v}$ of the detected grid cell (u,v) augmented with camera ray embeddings, following an approach similar to Multi-HMR [1]. These *detection features* thus depend on camera intrinsics, and will serve as main conditioning variable for predicting the different human attributes.

Human attributes. SMPL-X parameterizes body shape and expressions as latent vectors of a PCA space of dimension D (D=11 for shape, D=10 for expression, in our setup). We therefore model the conditional distributions for shape and expression as multivariate diagonal Gaussians. The absolute 3D location of the person is decomposed into 2D coordinates c of the reference keypoint in the image, and distance d to the image plane. To encode this distance, we use the variable $\ln(d/f)$ referred to as *encoded depth*, and model the conditional distributions for c and $\ln(d/f)$ as normal distributions. The $\ln(d/f)$ encoding ensures that d remains positive and allows for stronger conditioning on camera intrinsics. Lastly, pose is parameterized as a tuple $\theta \in SO(3)^J$ of $J = 53$ bone orientations. Since $SO(3)$ has a more complex topology than the PCA space of shapes and expressions, we model the conditional pose distribution as a product of independent matrix Fisher distributions $\Pi_{j=1}^{J}\mathcal{F}(\boldsymbol{F}_j)$ of density defined for a rotation matrix $\boldsymbol{R} \in SO(3)$ and some distribution parameters $\boldsymbol{F} \in \mathcal{M}_{3\times3}(\mathbb{R})$ by:

$$p_{\mathcal{F}(\boldsymbol{F})}(\boldsymbol{R}) = c(\boldsymbol{F})\exp(\mathrm{Tr}(\boldsymbol{F}^\top \boldsymbol{R})), \tag{2}$$

with $c(\boldsymbol{F})$ a normalization constant.

Distribution parametrization. To model a D-dimensional Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ while avoiding degenerate cases, we regress the mode $\boldsymbol{\mu} \in \mathbb{R}^D$ of the distribution along with dispersion parameters $\boldsymbol{\sigma} \in \mathbb{R}^D$, from which we construct the diagonal covariance matrix $\boldsymbol{\Sigma} = \mathrm{diag}(1+\exp(\boldsymbol{\sigma}))^2$. Similarly, we decompose the parameter $\boldsymbol{F} \in \mathcal{M}_{3\times3}(\mathbb{R})$ of a matrix Fisher distribution $\mathcal{F}(\boldsymbol{F})$ into a mode $\boldsymbol{R} \in SO(3)$ and dispersion parameters ($\boldsymbol{O} \in SO(3)$, $\boldsymbol{\Lambda} \in \mathbb{R}^3$), more suitable to be regressed by a MLP. These components are combined as follows: $\boldsymbol{F} = \boldsymbol{R}\boldsymbol{O}\,\mathrm{diag}(\lambda\,\mathrm{sigmoid}(\boldsymbol{\Lambda}))\boldsymbol{O}^\top$, where sigmoid denotes the element-wise sigmoid function and λ is a strictly positive scaling constant (we use $\lambda = 2$ in our experiments). Rotations are regressed as 3×3 matrices, which are orthonormalized using a differentiable special Procrustes operator implemented in RoMa [6].

Normalization constant. The matrix Fisher probability density function of Eq. (2) is defined up to a constant $c(\boldsymbol{F})$. To evaluate this constant during training, we use numerical integration by sampling 36,864 rotations on a uniform SO(3) grid proposed by Yershova et al. [69].

3.3. Inference

At inference, we extract predictions from our Bayesian network in an efficient feed-forward manner. Given a predicted distribution for a random variable (*e.g.* body shape), we can sample a value for this variable and use it to predict conditioned distributions (*e.g.* pose distribution, illustrated Fig. 2b). We iterate this process until all variables are sampled to generate hypotheses.

Mode extraction In practical applications, one is typically interested in extracting the most likely predictions given observations, *i.e.* solutions of Eq. (1). Finding such solution is not straightforward however, notably due to the non-linearities introduced by the MLPs regressing conditional distributions parameters. We therefore use a greedy but much more efficient alternative. Similar to the sampling procedure described above, we proceed in a feed-forward iterative manner through the Bayesian graph. We assign to each variable a value corresponding to the mode of the associated conditional distribution, and we iterate until all variables are evaluated. This algorithm can be implemented very efficiently with the normal and Fisher distributions considered in our experiments, as mode values are readily available in the regressed distribution parameters.

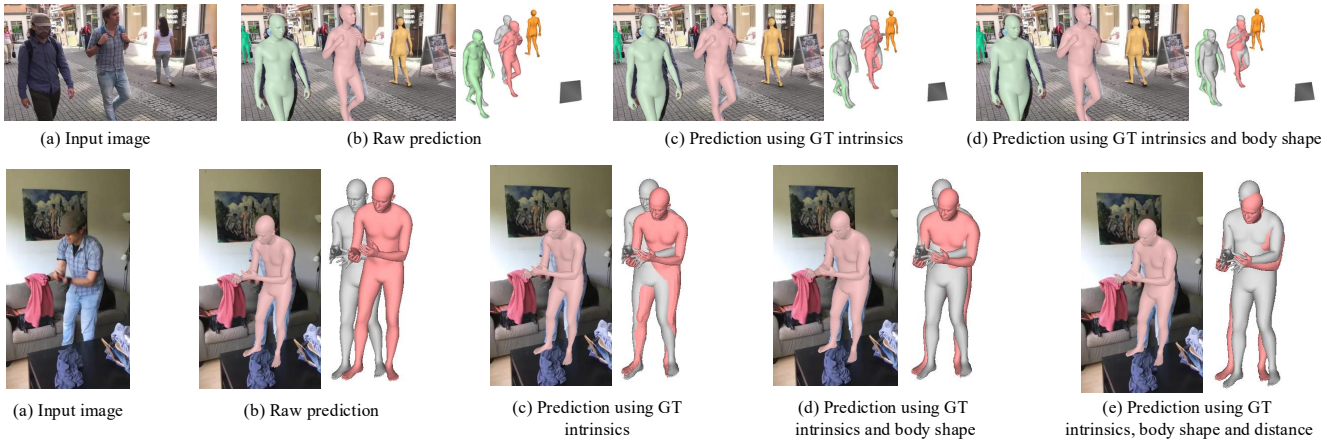


Figure 3. **Qualitative results**. Leveraging additional inputs, such as camera intrinsics and body shape, reduces errors and improves mesh accuracy. Ground-truth meshes are shown in grey for comparison.

Note that more advanced distributions such as normalizing flows [31, 53] could be used for greater expressivity, but these do not enable as efficient mode extraction as the simpler alternatives considered in this work.

Using known variables. A major benefit of modeling conditional distributions over deterministic regression is the ability to exploit additional information available. In many applications, prior knowledge such as camera intrinsic parameters (from calibration or image metadata), a person's body shape (when imaging a known individual), or their distance from the camera (using depth sensors, for instance) can be leveraged. During inference, we simply inject the known values of corresponding variables into our Bayesian network instead of performing mode extraction, as shown Fig. 2b, to improve prediction consistency.

Multi-view prior. Better results can be obtained by exploiting k simultaneous observations of the same person from different viewpoints, when available. In our experiments, we assume that the camera poses are unknown. We decompose the pose parameters into a global rigid orientation parameter θ_0 and intrinsic, viewpoint-independent, pose parameters $\tilde{\theta}_j = \theta_0^{-1}\theta_j$, with $j = 1 \dots J-1$. We then look for the multi-view prediction maximizing the product of posterior probabilities conditioned by image features:

$$\prod_{i=1}^{k} p(K^i, t^i, (\theta_0^i, \theta_1 \dots \theta_{J-1}), \beta, \gamma | \mathcal{I}^i), \tag{3}$$

where variables specific to a view $i = 1 \dots k$ are denoted with superscript i. For greater efficiency, we also solve this problem greedily. We start by an initial rigid alignment of predictions (obtained separately for each view) to estimate global orientations $(\theta_0^i)_{i=1\dots k}$, and then proceed to find the optimal intrinsic orientations $\tilde{\theta}_j$. These are determined by minimizing the product of Fisher probability densities $\prod_{i=1}^{k} p_{\mathcal{F}(\theta_0^i \boldsymbol{F}_j^i)}(\tilde{\theta}_j)$ for each bone orientation $j = 1 \dots J-1$ and view i , see Eq. (2). This optimization admits a closed-form solution, which consists for each bone orientation $\tilde{\theta}_j$ in the special Procrustes orthonormalization of $\sum_{i=1}^{k} \theta_0^i \boldsymbol{F}_{\boldsymbol{j}}^{\boldsymbol{i}}$.

Matching. With multiple input views and multiple people per view, predictions from each view need to be matched to ensure they correspond to the same person, before they can be combined. For simplicity, we rely on Hungarian matching to obtain these matches, with cost matrices computed from pairs of single-view predictions after rigid alignment.

3.4. Training

We train CondiMen to regress conditional distributions using an empirical cross-entropy objective function denoted $\mathcal{L}_{\texttt{prob}}$. This objective function aims to maximize the predicted log-probability density of the ground truth variables $(K^{*i}, s^{*i}, (t_j^{*i}, \theta_j^{*i}, \beta_j^{*i}, \gamma_j^{*i})_j)$ given images $i = 1 \dots n$:

$$\mathcal{L}_{\texttt{prob}} = -\frac{1}{n}\sum_{i=1}^{n} \log p(K^{*i}, s^{*i}, (t_j^{*i}, \theta_j^{*i}, \beta_j^{*i}, \gamma_j^{*i})_j | \mathcal{I}^i), \tag{4}$$

with visible people indexed by j. This joint probability density corresponds to the product of conditional probability densities predicted in our Bayesian network. We minimize the objective function by mini-batch gradient descent.

Mode guiding. To achieve better predictions, we found it beneficial to additionally guide the mode extraction procedure of our method. At training time, we consider some input camera intrinsics K and generate human attribute hypotheses using the procedure described in Sec. 3.3. This results in human mesh predictions $\hat{\mathrm{V}}$, composed of $|V|$ vertices centered at 3D location $\hat{t}$. Denoting π_K as the 2D projection operator onto the image plane, we aim to minimize a reprojection error on the vertices relative to the ground truth (K^*, V^*, t^*): $\mathcal{L}_{\texttt{reproj}} = \frac{1}{|V|}\sum_n \left|\pi_K(\hat{\mathrm{V}}_n + \hat{t}_n) - \pi_{K^*}(\mathrm{V}_n^* + t_n^*)\right|$. For 50% of the mini-batches, we randomly sample camera intrinsics with an horizontal field-of-view

uniformly chosen between 5 and 170°. For the remaining mini-batches, we use ground-truth camera intrinsics and introduce an additional objective function to minimize a human-centered vertices loss: $\mathcal{L}_{\texttt{mesh}} = \frac{1}{|V|}\sum_n \left|\hat{\mathrm{V}}_n - \mathrm{V}_n^*\right|$. With the addition of these two deterministic losses, our total objective function can be expressed as:

$$\mathcal{L} = \mathcal{L}_{\texttt{prob}} + \mathcal{L}_{\texttt{mesh}} + \mathcal{L}_{\texttt{reproj}}. \tag{5}$$

3.5. Implementation details

Architecture. For our experiments, we use an architecture inspired by the single-shot framework of Multi-HMR [1], for its simplicity and state-of-the-art performance. The input RGB image is encoded via a transformer-based architecture to produce 1024-dimensional image patch features $\mathcal{I}$, alongside detection features of similar dimensions. We train the entire network in an end-to-end manner to minimize the objective function (5), starting from DINOv2 [41] weights initialization for the image encoder. In our default settings, we use a ViT-Large encoder with an image resolution of 518×518 and a patch size of 14×14. The MLPs outputting conditional distributions parameters take the conditioning variables as input and combine them through a sum in an hidden space (typically of dimension 256) after linear projection followed by a rectilinear activation. We use the Adam optimizer [29] with a learning rate of $5 \cdot 10^{-6}$ and train for 500k steps with a batch size of 16 images. We consider all image patches with a detection score above 0.5 as detections in our experiments, after applying a non-maxima suppression of 3×3 patch window.

Training Data. We train models using only synthetic data in our experiments: synthetic data has the advantage of mitigating personal privacy issues, it provides potentially perfect ground-truth annotations, and was shown to transfer well to real word applications in practice [1, 4, 42]. Namely, we rely on the BEDLAM [4] dataset, which consists of 286k images depicting 951k persons. To further increase the body shape diversity we additionally render a synthetic dataset of 8k scenes. It contains images depicting 7 persons on average, with 40 multi-view renderings per scene, leading to more than 300k images and 2M human instances rendered. Additional details are provided in the supplementary material.

Computing Resources. Training our ViT-L model took about 51 hours using one NVIDIA H100 GPU ($36h$ and 35.5 hours resp. for ViT-B and ViT-S variants). Inference time depends on the number of people visible in each image. For reference, we report average inference times on the 3DPW dataset in Table 2a.

4. Experiments

To evaluate the effectiveness of our approach, we assess how CondiMen can leverage additional input information, and compare its performance against state-of-the-art methods. We conduct experiments on single-view datasets (3DPW [62] and MuPoTS [38]) as well as multi-view datasets (HI4D [22], Human3.6M [23] and RICH [22]). We report Mean Per Vertex Error in *mm*, with or without Procrustes alignment (*PVE* and *PA-PVE*, resp.), along with mean absolute position error, (*PE*) in *mm*. Following previous works [1, 58, 59], we also provide Mean Per Joint Error in *mm* (*PJE*) on 3DPW and Human3.6M, and Percentage of Correct Keypoints (*PCK*) within $15cm$ on MuPoTS. For experiments with additional input information, we convert SMPL annotations of 3DPW to the SMPL-X format using the code of Choutas et al. [11]. Note that we do not perform these experiments on MuPoTS due to the absence of mesh annotations.

Graph connectivity. Our Bayesian network architecture introduces conditional dependencies in the estimation of human attributes. For instance, pose prediction depends on body shape, as shown in the graphical model in Fig. 1. To evaluate the effectiveness of this strategy, we compare our approach with a *Naive-Bayes* baseline model, in which human attributes are predicted from detection features independently, without conditional dependencies, as illustrated in Fig. 2. Such baseline is similar to Multi-HMR [1], except that it regresses parametric distributions instead of deterministic values, ensuring a more meaningful comparison with *CondiMen*. Results are summarized in Fig. 4, and additional results can be found in the supp. mat. Our approach consistently outperforms the *Naive-Bayes* baseline, especially when leveraging additional input information for certain variables. This result underscores the importance of modeling dependencies between human attributes.

Additional input information. We report in Fig. 4 performance gains achieved when incorporating additional input information, using the approach described in Sec. 3.3. As expected, providing distance to the camera in addition to the camera intrinsics (*intrinsics-distance*) significantly enhances mean absolute vertex position accuracy, reducing the average PE error across datasets to 91mm for *CondiMen* and 103mm for *Naive Bayes*. Exploiting body shape information (*intrinsics-shape*) similarly boosts relative pose estimation, yielding up to a 25% reduction in PVE error. It also brings significant improvement in term of absolute position error (-57% on HI4D, -57% on RICH for *CondiMen*), which suggests that the model is able to capture the relationship between visual appearance, body shape and distance to the camera, and to exploit this dependency to produce better predictions (see examples Fig. 1 and 3). In contrast, the *Naive Bayes* baseline which does not model attribute interdependencies shows only minor improvements in absolute position error (-1.6% HI4D and -0.5% on RICH). On 3DPW, using additional shape information has marginal impact, as the model's initial shape estimates are already rather

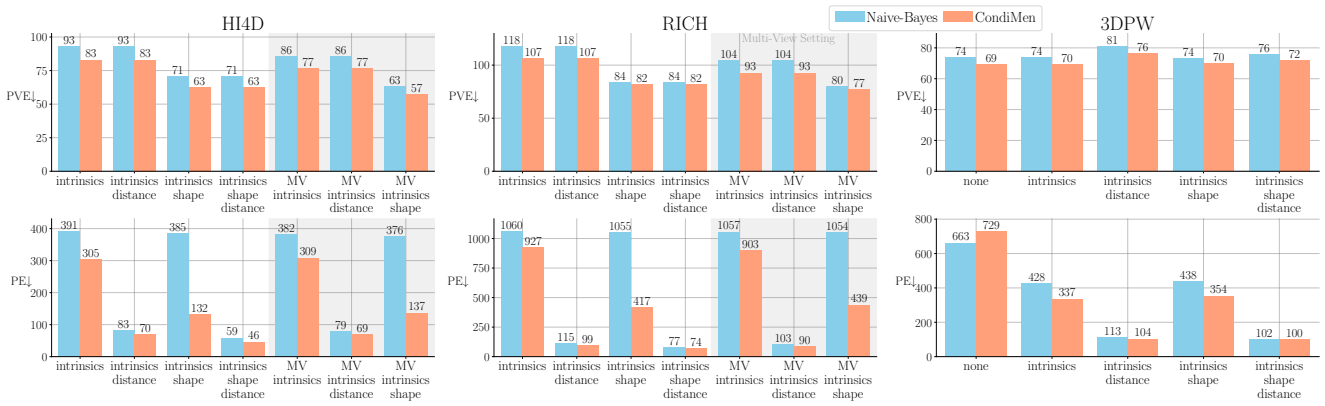


Figure 4. **Impact of additional information**. We report metrics for mesh reconstruction accuracy (Per Vertex Error, PVE) and 3D positioning (Positioning Error, PE) across different datasets. Exploiting additional information such as camera intrinsics (*intrinsics*), body shape parameters (*shape*), distance to the camera (*distance*), or multi-view observations (MV) can significantly reduce prediction errors. By modeling relationships between variables, *CondiMen* can consistently achieve lower errors than a *Naive Bayes* baseline.

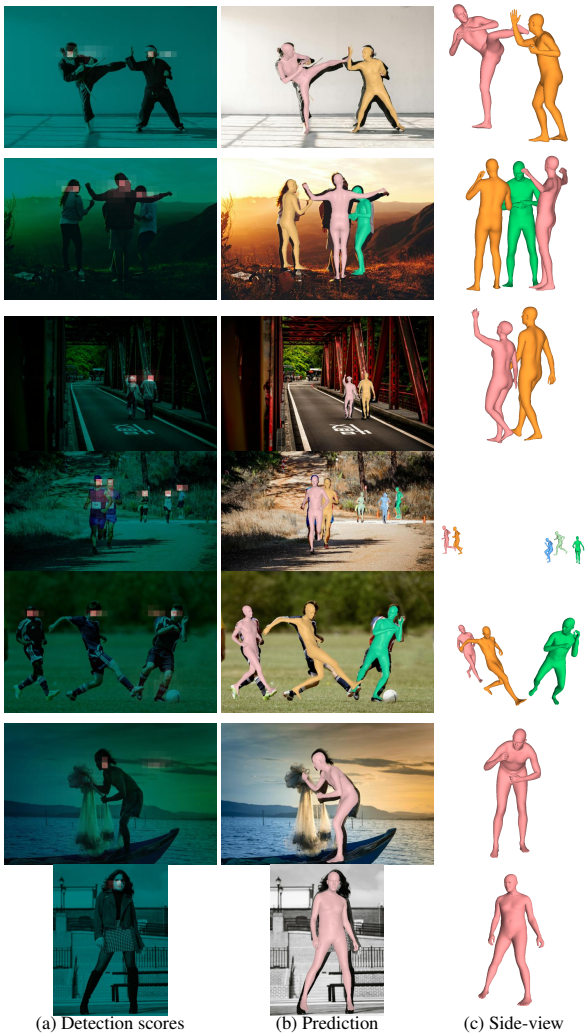
(a) Input (b) Ground truth (c) Single-view (d) Multi-view

Figure 5. **Improved prediction using multi-view prior.**

close to ground truth. However, using camera intrinsics significantly reduces localization error (by 53% between *none* and *intrinsics*). Finally, incorporating external information regarding camera intrinsics, body shape and distance to the camera provides further improvements for both methods.

Multi-view. We also report in Fig. 4 results obtained using a multi-view consistency prior, using the approach described in Sec. 3.3. We observe that the use of multiple views results in a smaller PVE error compared to the monocular case (see also Fig 5), and that providing external input leads to further performance improvements, with observations similar to the ones made in the monocular case.

State-of-the-art comparison. Finally, we compare the performance of our proposed method with recent approaches in both monocular and multi-view settings. Table 1 presents the performance results as reported by the authors, except for Multi-HMR [1], which we retrained on our dataset using a ViT-Large backbone at 518×518 resolution for a fair comparison. Following established practices [79], we report results for Human3.6M, HI4D and RICH after fine-tuning on each corresponding train set (for $15k$ steps), and we use ground-truth camera intrinsics for evaluation [1]. We also report results of a strong multi-view baseline – denoted *Multi-HMR+avg* which averages (after a rigid registration) the shape and expression vectors, as well as the absolute bone orientations predicted by Multi-HMR across multiple views. CondiMen achieves highly competitive results in both monocular and multi-view settings, and



(a) Detection scores (b) Prediction (c) Side-view

Figure 6. **Qualitative results** on free-to-use internet images. Our method can produce plausible predictions even when no additional inputs (such as camera intrinsics or body shape) are available.

Method	MV	Human3.6M ↓ PJE	PA-PJE	PVE	PA-PVE	HI4D ↓ PVE	PA-PVE
ProHMR [31]		65.1	43.7	–	–		
ROMP [58]		–	–	–	–	215.3	–
BEV [59]		–	–	–	–	153.9	–
HMR2.0 [19]		50.0	**32.4**	–	–	141.2	–
Yu et al. [71]		–	41.6	–	46.4	–	–
SMPLer-X[8]		–	38.9	–	42.8	–	–
MUC [79]		–	44.3	–	45.8	–	–
Multi-HMR [1]		50.8	40.0	62.2	43.3	49.0	36.2
Ours		**49.0**	38.9	**59.1**	**42.2**	**48.8**	**35.7**
ProHMR [31]	✓	62.2	34.5	–	–	–	–
Yu et al. [71]	✓	–	33.0	–	34.4	–	–
SMPLer-X[8] + avg	✓	–	33.4	–	37.1	–	–
MUC [79]	✓	–	31.9	–	33.4	–	–
Calib-free PaFF [26]	✓	44.8	28.2	–	–	–	–
OUVR [66]	✓	-	**27.1**	-	**28.9**	–	–
Multi-HMR [1] + avg	✓	42.8	30.0	51.3	32.5	51.1	28.7
Ours	✓	**41.1**	28.9	**49.0**	31.2	**44.7**	**27.8**

(a) **Multi-view setting.**

Method	MV	Whole-body↓ PVE	PA-PVE	Hands↓ PA-PVE	Face↓ PA-PVE
MUC [79]		–	47.7	8.2	4.1
Multi-HMR [1]		69.4	38.2	7.4	3.7
Ours		**65.4**	**36.5**	**7.3**	**3.6**
MUC [79]	✓	–	33.5	**6.7**	**3.2**
Multi-HMR [1]+avg	✓	61.7	29.5	6.8	3.5
Ours	✓	**57.8**	**27.9**	**6.7**	3.4

(b) **Body-part specific results on RICH.**

Method	3DPW ↓ PJE	PA-PJE	MuPoTS PCK↑ Matched	All
ProHMR [31]	–	59.8	–	–
ROMP [58]	76.7	47.3	69.9	72.2
BEV [59]	78.5	46.9	75.2	70.2
Multi-HMR [1]	70.6	47.5	77.6	**82.7**
HMR2.0 [19]	81.3	54.3	–	–
POCO [14]	69.7	**42.8**	–	–
Ours	**69.5**	46.4	**84.7**	74.0

(c) **Monocular setting.**

Table 1. **Comparison with state-of-the-art methods** in monocular and multi-view settings.

we provide qualitative examples in Fig. 3 and 6.

4.1. Additional ablations

Backbone. Tab. 2a shows the results when experimenting with various image encoder backbones. As expected, larger backbones yield better results: a ViT-Large model performs better than a ViT-Base which itself performs better than ViT-Small, likely due to the large scale of the training sets. However, these improvements in prediction quality come with additional computation costs both at training and inference time. Still, inference takes less than 50ms per image on average with a ViT-Large backbone, making it suitable for some real-time applications.

Training procedure. We also conduct ablation studies on various aspects of our training procedure, with results presented in Table 2b. First, we remove the mode-guiding objectives described in Sec. 3.4 and observe a drop in performance across all datasets. We also train a variant of our model without the additional synthetic data we generated. While this variant performs better on Human3.6M and HI4D, its performance is worse on RICH, 3DPW, and MuPoTS. We posit this is due to the presence of less standard camera parameters or body shape attributes in these benchmarks, for which the diversity of additional training data proves beneficial.

Uncertainty modeling Empirically, we observe a correlation between the predicted likelihood $p(K, t, \theta, \beta, \gamma|\mathcal{I})$ and the prediction errors (see supp. mat.). This suggests that the proposed model is able to estimate prediction uncertainty to some extent, an information that could be valuable for downstream applications.

Backbone	Human3.6M↓	HI4D↓	RICH↓	3DPW↓	MuPoTS↑	Inference↓ (ms)
ViT-Small	117.9	96.2	114.2	83.2	67.8	**29**
ViT-Base	99.7	93.3	108.2	76.6	69.7	31
ViT-Large	**88.8**	**77.0**	**92.8**	**69.5**	**74.0**	50

(a) **Backbone size.**

Training	Human3.6M↓	HI4D↓	RICH↓	3DPW↓	MuPoTS↑
Ours	88.8	77.0	**92.8**	**69.5**	**74.0**
w/o mode guiding	92.4	78.5	106.1	75.3	71.8
w/o additional synth. data	**75.9**	**73.8**	96.4	74.4	71.5

(b) **Training.**

Table 2. **Ablation study.** We report PVE metric in a multi-view setting on Human3.6M, HI4D and RICH. We report PJE for 3DPW and PCK-All for MuPoTS in a monocular setting. Inference durations are measured with an NVidia V100 GPU.

5. Discussion

We propose a novel approach for multi-person human mesh recovery, based on a Bayesian network. This method enables the seamless incorporation of additional input information – such as camera intrinsics, body shape, distance from the camera, or even multi-view acquisitions – to improve the predictions. We believe this has a significant practical value, as such data is readily available in many real-world applications. Our approach achieves on par or better performances than existing state-of-the-art methods on standard benchmarks, while still maintaining real-time capabilities. The motivation behind our work was to exploit interdependencies between various attributes in the mesh recovery problem (*e.g.* detection, camera parameters, and pose estimation). To this end, we adopted a Bayesian network framework, which is conceptually simple and supports highly efficient inference. Our findings confirms the effectiveness of this approach and suggest that exploring more sophisticated probabilistic modeling techniques could be a promising direction for future research.

References

[1] Fabien Baradel, Matthieu Armando, Salma Galaaoui, Romain Brégier, Philippe Weinzaepfel, Grégory Rogez, and Thomas Lucas. Multi-hmr: Multi-person whole-body human mesh recovery in a single shot. In *ECCV*, 2024. 1, 2, 3, 4, 6, 7, 8

[2] Benjamin Biggs, David Novotny, Sebastien Ehrhardt, Hanbyul Joo, Ben Graham, and Andrea Vedaldi. 3d multi-bodies: Fitting sets of plausible 3d human models to ambiguous image data. In *NeurIPS*, 2020. 3

[3] Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*. Springer, 2006. 2

[4] Michael J Black, Priyanka Patel, Joachim Tesch, and Jinlong Yang. Bedlam: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In *CVPR*, 2023. 2, 6

[5] 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. 2

[6] Romain Brégier. Deep regression on manifolds: a 3D rotation case study. In *3DV*, 2021. 4

[7] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *NeurIPS*, 2020. 3

[8] Zhongang Cai, Wanqi Yin, Ailing Zeng, Chen Wei, Qingping Sun, Yanjun Wang, Hui En Pang, Haiyi Mei, Mingyuan Zhang, Lei Zhang, et al. Smpler-x: Scaling up expressive human pose and shape estimation. In *NeurIPS*, 2023. 2, 8

[9] Wenzheng Chen, Huan Wang, Yangyan Li, Hao Su, Zhenhua Wang, Changhe Tu, Dani Lischinski, Daniel Cohen-Or, and Baoquan Chen. Synthesizing training images for boosting human 3d pose estimation. In *3DV*, 2016. 2

[10] Rohan Choudhury, Kris M Kitani, and László A Jeni. Tempo: Efficient multi-view pose estimation, tracking, and forecasting. In *ICCV*, 2023. 2

[11] Vasileios Choutas, Georgios Pavlakos, Timo Bolkart, Dimitrios Tzionas, and Michael J Black. Monocular expressive body regression through body-driven attention. In *ECCV*, 2020. 2, 3, 6

[12] Vandad Davoodnia, Saeed Ghorbani, Marc-André Carbonneau, Alexandre Messier, and Ali Etemad. Upose3d: Uncertainty-aware 3d human pose estimation with cross-view and temporal cues. In *ECCV*, 2024. 2

[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 3

[14] Sai Kumar Dwivedi, Cordelia Schmid, Hongwei Yi, Michael J Black, and Dimitrios Tzionas. Poco: 3d pose and shape estimation with confidence. In *3DV*, 2024. 2, 8

[15] Sai Kumar Dwivedi, Yu Sun, Priyanka Patel, Yao Feng, and Michael J Black. Tokenhmr: Advancing human mesh recovery with a tokenized pose representation. In *CVPR*, 2024. 2

[16] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *CVPR*, 2021. 3

[17] Yao Feng, Vasileios Choutas, Timo Bolkart, Dimitrios Tzionas, and Michael J Black. Collaborative regression of expressive bodies using moderation. In *3DV*, 2021. 2

[18] Guénolé Fiche, Simon Leglaive, Xavier Alameda-Pineda, Antonio Agudo, and Francesc Moreno-Noguer. Vq-hps: Human pose and shape estimation in a vector-quantized latent space. *ECCV*, 2024. 3

[19] Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. Humans in 4d: Reconstructing and tracking humans with transformers. In *ICCV*, 2023. 2, 8

[20] Charlie Hewitt, Fatemeh Saleh, Sadegh Aliakbarian, Lohit Petikam, Shideh Rezaeifar, Louis Florentin, Zafiirah Hosenie, Thomas J Cashman, Julien Valentin, Darren Cosker, et al. Look ma, no markers: holistic performance capture without the hassle. In *SIGGRAPH Asia*, 2024. 2

[21] Buzhen Huang, Chen Li, Chongyang Xu, Liang Pan, Yangang Wang, and Gim Hee Lee. Closely interactive human reconstruction with proxemics and physics-guided adaption. In *CVPR*, 2024. 3

[22] Chun-Hao P. Huang, Hongwei Yi, Markus Höschle, Matvey Safroshkin, Tsvetelina Alexiadis, Senya Polikovsky, Daniel Scharstein, and Michael J. Black. Capturing and inferring dense full-body human-scene contact. In *CVPR*, 2022. 2, 6

[23] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Trans. PAMI*, 2014. 2, 6

[24] Karim Iskakov, Egor Burkov, Victor Lempitsky, and Yury Malkov. Learnable triangulation of human pose. In *ICCV*, 2019. 2

[25] Ehsan Jahangiri and Alan L Yuille. Generating multiple diverse hypotheses for human 3d pose consistent with 2d joint detections. In *ICCVW*, 2017. 2

[26] Kai Jia, Hongwen Zhang, Liang An, and Yebin Liu. Delving deep into pixel alignment feature for accurate multi-view human mesh recovery. In *AAAI*, 2023. 2, 8

[27] Hanbyul Joo, Natalia Neverova, and Andrea Vedaldi. Exemplar fine-tuning for 3d human model fitting towards in-the-wild 3d human pose estimation. In *3DV*, 2021. 2

[28] 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

[29] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 6

[30] 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. 2

[31] Nikos Kolotouros, Georgios Pavlakos, Dinesh Jayaraman, and Kostas Daniilidis. Probabilistic modeling for human mesh recovery. In *ICCV*, 2021. 2, 5, 8

[32] Chen Li and Gim Hee Lee. Generating multiple hypotheses for 3d human pose estimation with mixture density network. In *CVPR*, 2019. 2

[33] Jiefeng Li, Siyuan Bian, Ailing Zeng, Can Wang, Bo Pang, Wentao Liu, and Cewu Lu. Human pose regression with residual log-likelihood estimation. In *ICCV*, 2021. 2

[34] Zhongguo Li, Magnus Oskarsson, and Anders Heyden. 3d human pose and shape estimation through collaborative learning and multi-view model-fitting. In *WACV*, 2021. 2

[35] Zhihao Li, Jianzhuang Liu, Zhensong Zhang, Songcen Xu, and Youliang Yan. Cliff: Carrying location information in full frames into human pose and shape estimation. In *ECCV*, 2022. 2

[36] Jing Lin, Ailing Zeng, Haoqian Wang, Lei Zhang, and Yu Li. One-stage 3d whole-body mesh recovery with component aware transformer. In *CVPR*, 2023. 2

[37] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics*, 2015. 2

[38] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Srinath Sridhar, Gerard Pons-Moll, and Christian Theobalt. Single-shot multi-person 3d pose estimation from monocular rgb. In *3DV*, 2018. 2, 6

[39] Gyeongsik Moon, Hongsuk Choi, and Kyoung Mu Lee. Accurate 3d hand pose estimation for whole-body 3d human mesh estimation. In *CVPRW*, 2022. 2

[40] Lea Müller, Vickie Ye, Georgios Pavlakos, Michael Black, and Angjoo Kanazawa. Generative Proxemics: A Prior for 3D Social Interaction from Images. In *CVPR*, 2024. 2, 3

[41] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *TMLR*, 2024. 6

[42] Priyanka Patel, Chun-Hao P Huang, Joachim Tesch, David T Hoffmann, Shashank Tripathi, and Michael J Black. Agora: Avatars in geography optimized for regression analysis. In *CVPR*, 2021. 2, 6

[43] 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

[44] 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

[45] Georgios Pavlakos, Jitendra Malik, and Angjoo Kanazawa. Human mesh recovery from multiple shots. In *CVPR*, 2022. 2

[46] Haibo Qiu, Chunyu Wang, Jingdong Wang, Naiyan Wang, and Wenjun Zeng. Cross view fusion for 3d human pose estimation. In *ICCV*, 2019. 2

[47] Zhongwei Qiu, Qiansheng Yang, Jian Wang, Haocheng Feng, Junyu Han, Errui Ding, Chang Xu, Dongmei Fu, and Jingdong Wang. Psvt: End-to-end multi-person 3d pose and shape estimation with progressive video transformers. In *CVPR*, 2023. 2, 3

[48] Gregory Rogez, Philippe Weinzaepfel, and Cordelia Schmid. Lcr-net: Localization-classification-regression for human pose. In *CVPR*, 2017. 2

[49] Yu Rong, Takaaki Shiratori, and Hanbyul Joo. Frankmocap: A monocular 3d whole-body pose estimation system via regression and integration. In *ICCV*, 2021. 2

[50] István Sárándi and Gerard Pons-Moll. Neural localizer fields for continuous 3d human pose and shape estimation. *NeurIPS*, 2024. 2

[51] Akash Sengupta, Ignas Budvytis, and Roberto Cipolla. Hierarchical kinematic probability distributions for 3d human shape and pose estimation from images in the wild. In *ICCV*, 2021. 3

[52] 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

[53] Akash Sengupta, Ignas Budvytis, and Roberto Cipolla. Humaniflow: Ancestor-conditioned normalising flows on so (3) manifolds for human pose and shape distribution estimation. In *CVPR*, 2023. 2, 5

[54] Soyong Shin, Juyong Kim, Eni Halilaj, and Michael J Black. Wham: Reconstructing world-grounded humans with accurate 3d motion. In *CVPR*, 2024. 2

[55] Hedvig Sidenbladh, Michael J Black, and David J Fleet. Stochastic tracking of 3d human figures using 2d image motion. In *ECCV*, 2000. 2

[56] Cristian Sminchisescu and Bill Triggs. Covariance scaled sampling for monocular 3d body tracking. In *CVPR*, 2001. 2

[57] Qingping Sun, Yanjun Wang, Ailing Zeng, Wanqi Yin, Chen Wei, Wenjia Wang, Haiyi Mei, Chi-Sing Leung, Ziwei Liu, Lei Yang, et al. Aios: All-in-one-stage expressive human pose and shape estimation. In *CVPR*, 2024. 2

[58] Yu Sun, Qian Bao, Wu Liu, Yili Fu, Michael J Black, and Tao Mei. Monocular, one-stage, regression of multiple 3d people. In *ICCV*, 2021. 1, 3, 6, 8

[59] Yu Sun, Wu Liu, Qian Bao, Yili Fu, Tao Mei, and Michael J Black. Putting people in their place: Monocular regression of 3d people in depth. In *CVPR*, 2022. 1, 2, 3, 6, 8

[60] Hanyue Tu, Chunyu Wang, and Wenjun Zeng. Voxelpose: Towards multi-camera 3d human pose estimation in wild environment. In *ECCV*, 2020. 2

[61] Aaron Van den Oord, Nal Kalchbrenner, Lasse Espeholt, Oriol Vinyals, Alex Graves, et al. Conditional image generation with pixelcnn decoders. In *NeurIPS*, 2016. 3

[62] 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. 2, 6

[63] Yufu Wang, Ziyun Wang, Lingjie Liu, and Kostas Daniilidis. Tram: Global trajectory and motion of 3d humans from in-the-wild videos. In *ECCV*, 2024. 2

[64] Tom Wehrbein, Marco Rudolph, Bodo Rosenhahn, and Bastian Wandt. Probabilistic monocular 3d human pose estimation with normalizing flows. In *ICCV*, 2021. 2

[65] Philippe Weinzaepfel, Romain Brégier, Hadrien Combaluzier, Vincent Leroy, and Grégory Rogez. Dope: Distillation of part experts for whole-body 3d pose estimation in the wild. In *ECCV*, 2020. 2

[66] Zhenyu Xie, Huanyu He, Gui Zou, Jie Wu, Guoliang Liu, Jun Zhao, Yingxue Wang, Hui Lin, and Weiyao Lin. Visibility-guided human body reconstruction from uncalibrated multi-view cameras. In *ICMR*, 2024. 2, 8

[67] 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*, 2020. 2

[68] Hang Ye, Wentao Zhu, Chunyu Wang, Rujie Wu, and Yizhou Wang. Faster voxelpose: Real-time 3d human pose estimation by orthographic projection. In *ECCV*, 2022. 2

[69] Anna Yershova, Swati Jain, Steven M Lavalle, and Julie C Mitchell. Generating uniform incremental grids on so (3) using the hopf fibration. *IJRR*, 2010. 4

[70] Yifei Yin, Chen Guo, Manuel Kaufmann, Juan Zarate, Jie Song, and Otmar Hilliges. Hi4d: 4d instance segmentation of close human interaction. In *CVPR*, 2023. 2

[71] Zhixuan Yu, Linguang Zhang, Yuanlu Xu, Chengcheng Tang, Luan Tran, Cem Keskin, and Hyun Soo Park. Multi-view human body reconstruction from uncalibrated cameras. In *NeurIPS*, 2022. 2, 8

[72] Ye Yuan, Jiaming Song, Umar Iqbal, Arash Vahdat, and Jan Kautz. Physdiff: Physics-guided human motion diffusion model. In *ICCV*, 2023. 3

[73] Hongwen Zhang, Yating Tian, Xinchi Zhou, Wanli Ouyang, Yebin Liu, Limin Wang, and Zhenan Sun. Pymaf: 3d human pose and shape regression with pyramidal mesh alignment feedback loop. In *ICCV*, 2021. 2

[74] Hongwen Zhang, Yating Tian, Yuxiang Zhang, Mengcheng Li, Liang An, Zhenan Sun, and Yebin Liu. Pymaf-x: Towards well-aligned full-body model regression from monocular images. *IEEE Trans. PAMI*, 2023. 2

[75] Jinlu Zhang, Yujin Chen, and Zhigang Tu. Uncertainty-aware 3d human pose estimation from monocular video. In *ACMMM*, 2022. 2

[76] Siwei Zhang, Qianli Ma, Yan Zhang, Sadegh Aliakbarian, Darren Cosker, and Siyu Tang. Probabilistic human mesh recovery in 3d scenes from egocentric views. In *ICCV*, 2023. 2

[77] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. In *arXiv preprint arXiv:1904.07850*, 2019. 4

[78] Yuxiao Zhou, Marc Habermann, Ikhsanul Habibie, Ayush Tewari, Christian Theobalt, and Feng Xu. Monocular real-time full body capture with inter-part correlations. In *CVPR*, 2021. 2

[79] Yitao Zhu, Sheng Wang, Mengjie Xu, Zixu Zhuang, Zhixin Wang, Kaidong Wang, Han Zhang, and Qian Wang. Muc: Mixture of uncalibrated cameras for robust 3d human body reconstruction. *arXiv preprint arXiv:2403.05055*, 2024. 2, 7, 8