Learning Analytical Posterior Probability for Human Mesh Recovery

Qi Fang[1] Kang Chen[1] Yinghui Fan[1] Qing Shuai[2] Jiefeng Li[3] Weidong Zhang[1]
[1]NetEase Games AI Lab [2]Zhejiang University [3]Shanghai Jiao Tong University

Abstract

Despite various probabilistic methods for modeling the uncertainty and ambiguity in human mesh recovery, their overall precision is limited because existing formulations for joint rotations are either not constrained to $\mathcal{SO}(3)$ or difficult to learn for neural networks. To address such an issue, we derive a novel analytical formulation for learning posterior probability distributions of human joint rotations conditioned on bone directions in a Bayesian manner, and based on this, we propose a new posterior-guided framework for human mesh recovery. We demonstrate that our framework is not only superior to existing SOTA baselines on multiple benchmarks but also flexible enough to seamlessly incorporate with additional sensors due to its Bayesian nature. The code is available at https://github.com/NetEase-GameAI/ProPose.

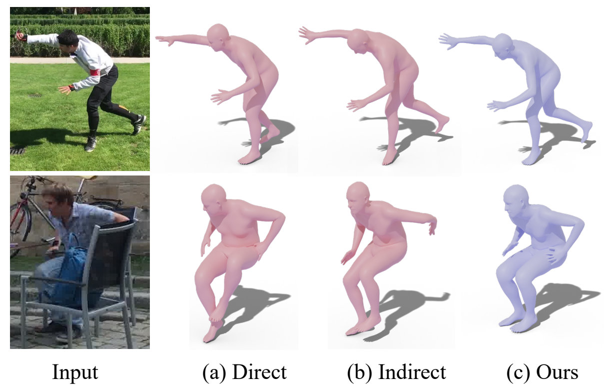

Figure 1. Comparisons of (a) the direct method [29], (b) the indirect method [36], and (c) our method.

1. Introduction

Human mesh recovery is a task of recovering body meshes and 3D joint rotations of human actors from images, which has ubiquitous applications in animation production, sports analysis, etc. To achieve this goal, various approaches have been proposed in the computer vision community. Existing methods can be divided into two categories, *i.e.*, direct and indirect, respectively. Direct methods use neural networks to regress the rotations (*e.g.*, axis angle [22], rotation matrix [34], 6D vector [29,37,74]) of each humanoid joint in an end-to-end way, while indirect methods recover joint rotations based on some intermediately predicted proxies (*e.g.*, 3D human keypoints [19, 36, 42], 2D heatmaps [52] or part segmentation [27]). However, both methods have obvious weaknesses. Generally, the estimated poses from direct solutions are not so well-aligned with the images (Fig. 1(a)), because joint rotations are more difficult to regress compared with keypoints [19,36]. On the contrary, though indirect solutions tend to have better estimation precision, their performance heavily relies on the precision of the intermediate proxies and thus are vulnerable to noise and error in the predicted keypoints or part segmentation (Fig. 1(b)).

To simultaneously achieve high precision and high robustness, some probabilistic methods are developed, which, instead of seeking a unique solution, try to explicitly model the uncertainty of human poses by learning some kind of probability distribution. Prevalent ways of modeling the distribution include multivariate Gaussian distributions [48, 56], normalizing flows [31], and neural networks [51, 53]. In practice, these learned probability distributions can notably improve the estimation results in some extreme cases (*e.g.*, under large occlusion), however, only minor differences can be found in terms of the overall performances on large datasets. One reason is that these probability models cannot truly reflect the rotational uncertainty since they are not strictly constrained to $\mathcal{SO}(3)$. Recently, [55] proposes to adopt the matrix Fisher distribution over $\mathcal{SO}(3)$ [8, 25] to model the rotational uncertainty caused by depth ambiguity. However, even with this mathematically-correct formulation, the actual performance does not improve much either, because the parameters of the matrix Fisher distribution are not easy for deep neural networks to learn directly.

To address this problem, we propose a new learning-friendly and mathematically-correct formulation for learning probability distributions for human mesh recovery. Our formulation is derived based on the facts that, (i) the joint rotations follow the matrix Fisher distribution over $\mathcal{SO}(3)$,

(ii) the unit directions of bones follow the von Mises-Fisher distribution [44], (iii) the bone direction can be viewed as the observation of joint rotation (*i.e.*, the latent variable). It can be proven that the probability distributions of joint rotations conditioned on bone directions still follow the matrix Fisher distribution, which allows us to regress the posterior probability distribution of the 3D joint rotations in a Bayesian manner, and more importantly, in an analytical form. Moreover, we mathematically prove that the posterior probability of human joint rotations is more concentrated than the prior probability. Our experimental results demonstrate that such a characteristic makes the posterior probability an easier form to learn (for neural networks) than its prior counterpart.

Apart from the theoretical contributions, we also propose a new human mesh recovery framework that can utilize the learned analytical posterior probability. We demonstrate that this framework successfully achieves high precision and high robustness at the same time, and outperforms existing SOTA baselines. Furthermore, our framework enables seamless integration with additional sensors that can yield directional/rotational observations (*e.g.*, multi-view cameras, optical markers, IMUs) due to its Bayesian nature. Different from naive multi-sensor fusion algorithms (*e.g.*, Kalman filter [21]) that typically perform fusion at the inference stage, our framework allows fusion in the training stage to learn the noise characteristics of sensors, and thus has the potential to produce better precision. We demonstrate that our fusion mechanism can achieve similar effects to fusing the latent features from multiple sensor input branches, but is much more flexible since it does not require modification of the main backbone.

The key contributions of this paper are thereby:

- We derive a novel analytical formulation for learning probability distributions for human joint rotations, and theoretically prove that such formulation allows the regression of posterior probability distribution in a Bayesian manner.
- We propose a new framework for human mesh recovery by leveraging the learned analytical posterior probability and show that this framework outperforms existing SOTA baselines.
- We introduce a novel and flexible multi-sensor fusion mechanism that allows fusing different observations in the training stage.

2. Related work

In this section, we discuss related studies on human mesh recovery, which can be achieved by optimization-based and learning-based methods. Leveraging the parametric human model [41, 54], optimization-based approaches [2, 9, 15, 51] fit the parameters via iteration while learning-based approaches regress the parameters with neural networks. Our work follows the learning paradigm, therefore we here mainly review recent advances in learning-based methods.

Direct methods: Given images as input, this kind of approach directly regresses the model parameters with neural networks. Different representations of rotation [22, 34, 74], supervision schemes [20, 29, 37] as well as temporal context [5, 13, 23, 26] are explored to improve performance. However, the gap between the image space and the abstract parameter space of statistical models makes it difficult to generate well-aligned estimations.

Indirect methods: Instead of regressing rotation representations from RGB images directly, plenty of works introduce proper intermediate or proxy representations, such as segmentation [27, 49, 68], IUV maps [64, 69, 70], keypoints [6, 14, 36, 52] or surface landmarks [30, 32, 42], to guide the learning of neural networks efficiently. HybrIK [36] decomposes the 3D rotation into solvable swings from 3D keypoints and extra predicted twists. PARE [27] learns to predict attention masks which are fused with image feature maps to provide body part information. These solutions may achieve higher precision, but generating only deterministic results and ignoring the uncertainty of estimation make them sensitive to noisy or erroneous proxy predictions.

Probabilistic methods: To deal with the uncertainty from occlusions or depth ambiguities, several works manage to produce multiple hypotheses [1] or a probability distribution [53]. I2L-MeshNet [48] predicts lixel-based 1D heatmaps for each human mesh vertex for uncertainty modeling. Sengupta *et al.* [56] assume simple multivariate Gaussian distributions over the parameters of the human model. ProHMR [31] learns a distribution of plausible 3D poses represented by normalizing flows, which is more powerful and expressive than Gaussian distributions. Recently Sengupta *et al.* [55] further represent the essential distribution of human joint rotation over $\mathcal{SO}(3)$ by adopting the matrix Fisher distribution [25], which can provide quantified uncertainty estimation. Despite a better explanation for ambiguities, the parameters of the above distribution are not easy to learn, limiting their overall performance on complicated scenes.

Multi-sensor fusion: Recently an increasing number of approaches attempt to integrate extra observations from other sensors, such as IMUs [10, 65, 66] and muli-view cameras [7, 71, 73], to obtain more reliable estimations. One simple strategy is combining all observations properly with

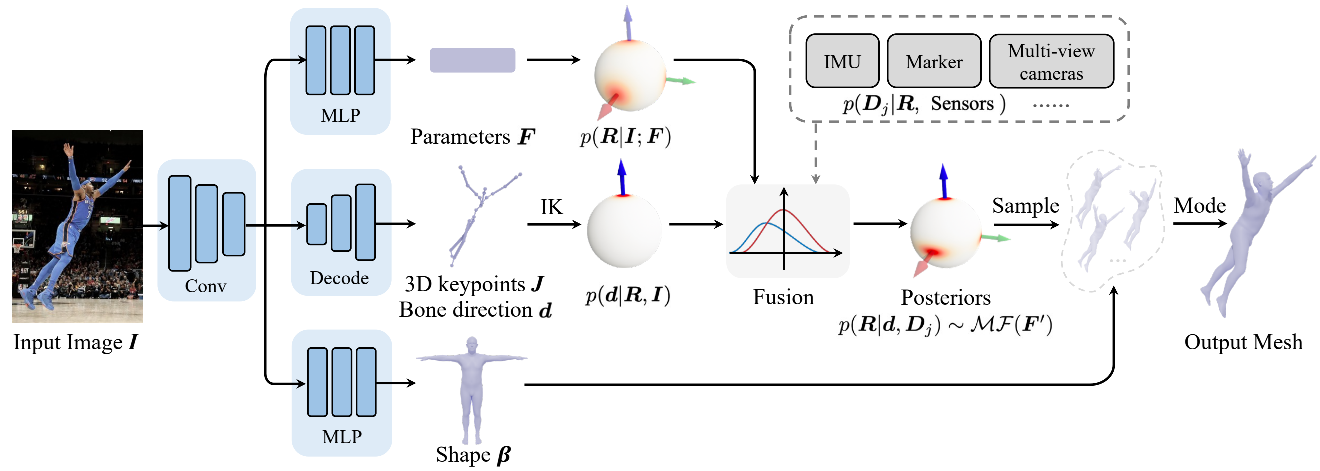

Figure 2. **Overview of our framework.** Given an input image, the multi-branch network predicts the prior matrix Fisher parameters $\boldsymbol{F}$, the 3D keypoints $\boldsymbol{J}$, and the SMPL shape parameters $\boldsymbol{\beta}$, respectively. The bone direction $\boldsymbol{d}$ calculated from $\boldsymbol{J}$ serves as the likelihood conditioned on 3D rotation. The posterior probability can be obtained based on Bayesian rules (Fusion), which still follows the matrix Fisher distribution, but with different parameters and larger confidences. Observations from additional sensors can also be fused into the posterior probability in the same manner. The corresponding human mesh can then be recovered using the estimated rotation and shape.

Kalman filter [21], which can be treated as a baseline. Some works [43,62,63] fit the human model to evidence including images and IMUs through joint optimization. Apart from these test-time fusion schemes, several approaches [10, 61] incorporate the fusing process into training by concatenating the features from images and IMUs directly. GeoFuse [72] reinforces the image features guided by IMUs to infer the occluded joints. Our framework is also flexible to perform multi-sensor fusion and generates competitive results without specific modification to the backbone.

3. Methods

In this section, we first mathematically introduce the probability distributions for orientations regarding rotations and directions (Sec. 3.1). Then, we model the human joint rotation and bone direction with the corresponding distribution, and derive the analytical formulation of the posterior probability of joint rotation conditioned on the bone direction with crucial conclusions and discussion (Sec. 3.2). Finally, we describe the proposed framework (Sec. 3.3) and learning details (Sec. 3.4).

3.1. Orientation probability distribution

Before delving into human modeling, we investigate the orientation representation for general rigid entities. Suppose $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ is the parametric representation of the entity orientation, each column of which depicts the direction of a basis. $\boldsymbol{X}$ is on the Stiefel manifold $\mathcal{V}(n,p)$ if $\boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{I}_p$, and when $n = p$, it further belongs to the orthogonal group $\mathcal{O}(n)$. Additionally, the components of $\mathcal{O}(n)$ with determinant $+1$ are referred to as the special orthogonal group $\mathcal{SO}(n)$, which is used to represent the rotation of n degrees of freedom. Meanwhile, if $p = 1$, the normalized $\boldsymbol{X}$, *i.e.*, n-dimensional unit vector on the manifold $(n-1)$-sphere $\mathcal{S}^{n-1}$, can represent the single direction as well. From a probabilistic perspective, when $\boldsymbol{X}$ is a random matrix, there are two common cases.

Rotation ($n = p$)**:** For rotation matrix $\boldsymbol{R} \in \mathcal{SO}(n)$, the matrix Fisher distribution $\mathcal{MF}(\cdot)$ has been proposed to characterize its probabilistic properties on $\mathcal{SO}(n)$ [8, 25]. The probability density function is as follows:

$$p(\boldsymbol{R};\boldsymbol{F}) = \frac{1}{c(\boldsymbol{F})} exp(tr(\boldsymbol{F}^T\boldsymbol{R})) \sim \mathcal{MF}(\boldsymbol{F}), \quad (1)$$

where $\boldsymbol{F} \in \mathbb{R}^{n\times n}$ is the distribution parameter, and $c(\boldsymbol{F})$ is a normalizing constant. Algebraically, $\boldsymbol{F}$ can be decomposed into a concentration matrix $\boldsymbol{K}$ and a mean rotation matrix $\boldsymbol{M}$ via SVD decomposition:

$$\boldsymbol{F} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^T = \underbrace{(\boldsymbol{U}\Delta\boldsymbol{V}^T)}_{\boldsymbol{M}\in\mathbb{R}^{n\times p}}\underbrace{(\boldsymbol{V}\Delta\boldsymbol{S}\boldsymbol{V}^T)}_{\boldsymbol{K}\in\mathbb{R}^{p\times p}}, \quad (2)$$

where $\Delta = diag(1, 1, |\boldsymbol{U}\boldsymbol{V}|)$ is a diagonal orthogonal matrix to ensure the determinant of $\boldsymbol{M}$ is $+1$. $\boldsymbol{K}$ is symmetric positive definite as long as $\boldsymbol{F}$ is full rank. A rotation estimation $\hat{\boldsymbol{R}}$ can be calculated from the mode of distribution:

$$\hat{\boldsymbol{R}} = \boldsymbol{M} = \boldsymbol{U} diag(1, 1, |\boldsymbol{U}\boldsymbol{V}|)\boldsymbol{V}^T. \quad (3)$$

Direction ($p = 1$)**:** The probability density function for a unit vector $\boldsymbol{d} \in \mathcal{S}^{n-1}$ is similar to Eq. (1) if we set $p = 1$, which corresponds to the classical von Mises-Fisher distribution $\mathcal{VMF}(\cdot)$ [44]:

$$p(\boldsymbol{d};\kappa,\boldsymbol{m}) = \frac{1}{c(\kappa)}exp(\kappa\boldsymbol{m}^T\boldsymbol{d}) \sim \mathcal{VMF}(\boldsymbol{m},\kappa), \quad (4)$$

where $c(\kappa)$ is a normalizing constant. $\boldsymbol{m}$ denotes the mean direction and κ denotes the concentration parameter, which have a close meaning to $\boldsymbol{M}$ and $\boldsymbol{K}$ in Eq. (2), respectively, and thus $\hat{\boldsymbol{d}} = \boldsymbol{m}$ becomes a direction estimation.

Theoretically, if κ is 0, $\mathcal{VMF}(\boldsymbol{m},0)$ is equivalent to the uniform distribution on the sphere, while if κ is large, it is close to the wrapped normal distribution $\mathcal{WN}(\boldsymbol{m},\kappa^{-1})$ that adds up the densities of vectors representing the same direction on the sphere due to the periodicity. Thus, κ can be viewed as the inverse of the variance and denotes the concentration of the distributions.

3.2. Human modeling

The human joint rotation can be represented as rotation matrix $\boldsymbol{R} \in \mathcal{SO}(3)$. Inspired by recent advances in object pose estimation [3, 33, 46, 67], we incorporate the probabilistic modeling for human poses. Specifically, we adopt the matrix Fisher distribution over $\mathcal{SO}(3)$ as the prior distribution for joint rotation. Moreover, as the bone direction can be easily calculated from the joint rotation, we regard the joint rotation $\boldsymbol{R}$ as the latent variable and the bone direction $\boldsymbol{d}$ as the corresponding observation, which follows the von Mises-Fisher distribution:

$$p(\boldsymbol{d}|\boldsymbol{R}) = \frac{1}{c(\kappa)}exp(\kappa\boldsymbol{l}^T\boldsymbol{R}^T\boldsymbol{d}) \sim \mathcal{VMF}(\boldsymbol{R}\boldsymbol{l},\kappa), \quad (5)$$

where $\boldsymbol{l}$ is the unit direction of the bone in the reference pose (*e.g.*, T-pose), ideally satisfying $\boldsymbol{R}\boldsymbol{l} = \boldsymbol{d}$.

Leveraging Bayesian inference, given the prior distribution (Eq. (1)) and the likelihood function (Eq. (5)), the posterior probability of joint rotation conditioned on the bone direction can be derived as follows:

$$\begin{aligned} p(\boldsymbol{R}|\boldsymbol{d}) &= \frac{p(\boldsymbol{R})\cdot p(\boldsymbol{d}|\boldsymbol{R})}{p(\boldsymbol{d})} \propto p(\boldsymbol{R})\cdot p(\boldsymbol{d}|\boldsymbol{R}) \quad (6)\\ &= \frac{1}{c}exp(tr[(\boldsymbol{F}+\kappa\boldsymbol{d}\boldsymbol{l}^T)^T\boldsymbol{R}]) \sim \mathcal{MF}(\boldsymbol{F}+\kappa\boldsymbol{d}\boldsymbol{l}^T). \end{aligned}$$

It can be concluded from Eq. (6) that the posterior probability $p(\boldsymbol{R}|\boldsymbol{d})$ also follows the matrix Fisher distribution with an updated parameter $\boldsymbol{F}' = \boldsymbol{F} + \kappa\boldsymbol{d}\boldsymbol{l}^T$.

Property: From another perspective, the posterior parameter $\boldsymbol{F}'$ can be viewed as the multiplication of the same mean term $\boldsymbol{M}$ and a new concentration term $\boldsymbol{K}'$:

$$\boldsymbol{F}' = \boldsymbol{F} + \kappa\boldsymbol{d}\boldsymbol{l}^T = \boldsymbol{M}(\underbrace{\boldsymbol{K} + \kappa\boldsymbol{M}^T\boldsymbol{d}\boldsymbol{l}^T}_{\boldsymbol{K}'}). \quad (7)$$

It can be proved that $\boldsymbol{M}^T\boldsymbol{d}\boldsymbol{l}^T = \boldsymbol{l}\boldsymbol{l}^T$ is a real symmetric matrix with rank 1, and $\boldsymbol{K}$ from Eq. (2) is also real symmetric, thus the posterior concentration term $\boldsymbol{K}'$ is a real symmetric matrix. According to the interlacing theorem for Hermitian matrices from matrix analysis [17], the eigenvalues for a Hermitian matrix with a rank-1 Hermitian perturbation satisfy the following inequality:

$$\lambda_1 \leq \lambda_1' \leq \lambda_2 \leq \cdots \leq \lambda_{p-1}' \leq \lambda_p \leq \lambda_p', \quad (8)$$

where λ_i and λ_i' denote the eigenvalues of $\boldsymbol{K}$ and $\boldsymbol{K}'$, respectively. Note that the eigenvalues of the concentration term equal the singular values of the distribution parameter, which reflect the confidence of the distribution. From Eq. (8) we can get the conclusion that the posterior estimation is more concentrated than the prior estimation as long as the likelihood term is non-zero, and is validated to be a more easily learnable formulation in the experiment and the supplementary material.

General form: Similarly, if other sensors that yield directional $\boldsymbol{d}_i$ or rotational $\boldsymbol{D}_j$ observations are available, the analytical posterior probability is thereby as follows:

$$p(\boldsymbol{R}|\{\boldsymbol{d}_i,\boldsymbol{D}_j\}) \sim \mathcal{MF}(\boldsymbol{F}+\sum_{i\in\mathcal{Z}_1}\kappa_i g(\boldsymbol{d}_i)+\sum_{j\in\mathcal{Z}_3}\boldsymbol{D}_j\boldsymbol{K}_j^T), \quad (9)$$

where κ_i and $\boldsymbol{K}_j$ are the concentration terms for weighting. $g(\cdot)$ is a mapping of IK that converts the directional observation to a rotation estimation, which is not limited to a specific IK algorithm as long as it supports gradient backpropagation (*e.g.*, the simple solution $\boldsymbol{d}\boldsymbol{l}^T$ in Eq. (6)). $\mathcal{Z}_1$ denotes the set of sensors providing directional observations such as accelerometers, while $\mathcal{Z}_3$ denotes the set of rotational sensors like gyroscopes. We simplify the original derivation by assuming the sensors are unbiased. Please refer to the supplementary material for the derivation.

Discussion: There are several advantages of our approach. First, adopting the matrix representation is more reasonable than other rotation representations. As presented in [34], a continuous 9D unconstrained representation followed by SVD can achieve comparable or even better performance than the widely used 6D vector [74]. Second, the Gaussian distribution is unsuitable for cases with large uncertainty where the assumption of local linearity cannot hold [11, 12], while the matrix Fisher distribution does not have this problem. Third, the posteriors are easier to learn than the priors in that learning the posteriors can converge to the mode preferred by the likelihood function quickly, while learning the priors may face multiple local minima in the initial stage and thus cannot converge well.

To recognize the proposed scheme intuitively, we show the schematic diagrams of probabilistic modeling in Fig. 3. For a method without probabilistic modeling (*e.g.*, using IK

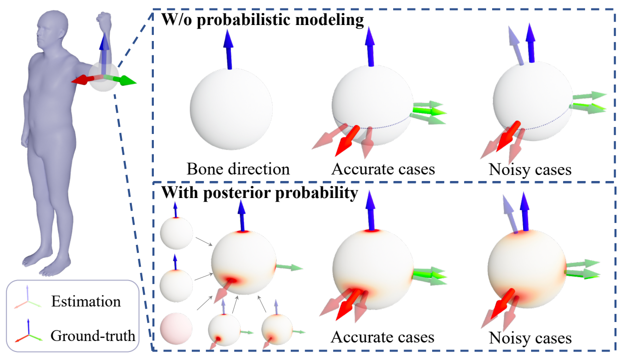

Figure 3. **Schematic diagrams of probabilistic modeling.** The opaque coordinate system is the ground-truth 3D rotation. The transparent rotation represents a deterministic estimation for a method without probabilistic modeling (row 1), while it denotes a sample from the estimated posterior distribution (row 2). The red region on the sphere represents the probability of a certain rotation, and it could cover the ground-truth even for noisy cases.

to solve rotations from keypoints), its underlying model is a single direction, thus it may be erroneous when the estimated bone direction deviates from the ground-truth, as shown in the noisy cases. In contrast, the posterior model can be fused with various models, and for noisy keypoints, it has the potential to recover the exact rotations since the negative impact of its partial reliance on the estimated keypoints can be eliminated by the prior or other observations.

3.3. Learning Framework

Our proposed framework that leverages the derived posterior probability for human mesh recovery is demonstrated in Fig. 2. We adopt the parametric model SMPL [41] as our human representation, which can also be replaced by other human models [51,54]. Given an input image, a CNN backbone is used to extract image features, followed by three output branches, including prior distribution parameter $\boldsymbol{F}$, 3D keypoints $\boldsymbol{J}$, and shape parameter $\boldsymbol{\beta}$. The adopted keypoints branch consists of normalized 2D keypoints and relative depth to the root joint decoded from the feature, as well as the human scale predicted by a small MLP branch, so as to recover absolute 3D keypoints. Note that other strategies for 3D keypoints estimation are also applicable. The bone direction $\boldsymbol{d}$ is calculated from $\boldsymbol{J}$. Then we utilize Eq. (9) to fuse $\boldsymbol{F}$, $\boldsymbol{d}$, or other optional observations analytically in the sense of probability. With the new distribution parameter $\boldsymbol{F}'$, we can get the rotation estimation according to Eq. (3). As for κ, it is related to the physical properties (measure covariance) of specific sensors that could be set in advance. If there is no prior knowledge of the sensor, it's feasible to tune it manually or learn it from the data. We simply use the scaled scores of estimated keypoints. For multiple sensors, the concentration term $\boldsymbol{K}$ is simplified as a diagonal matrix.

Objective functions: For the following objective functions, symbols with superscript '*' indicate the ground-truth, and symbols with hat indicate the estimation. The L_1 loss is used to supervise the 3D keypoints, while the L_2 loss is applied to other variables:

$$\mathcal{L}_J = ||\hat{\boldsymbol{J}} - \boldsymbol{J}^*||_1, \tag{10}$$

$$\mathcal{L}_\beta = ||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*||_2^2, \tag{11}$$

The mode of the posterior probability distribution $\hat{\boldsymbol{R}}$ can be solved according to Eq. (6) and (3), and is supervised if the annotations of SMPL pose parameters $\boldsymbol{\theta}$ are provided as follows:

$$\mathcal{L}_\theta = ||\hat{\boldsymbol{R}} - \mathrm{expm}(\boldsymbol{\theta}^*)||_2^2, \tag{12}$$

where expm denotes the exponential map implemented with the Rodrigues' formula.

Apart from the mode, the whole distribution also needs to be supervised. Since the normalizing constant in the distribution is hard to be calculated stably due to the numerical integration, we propose to supervise the distribution by sampling. Specifically, we adopt the rejection sampling technique to sample Bingham distribution of unit quaternions on $\mathcal{S}^3$ based on its equivalence to matrix Fisher distribution. The proposal distribution in rejection sampling is angular central Gaussian (ACG) distribution [24, 55]. Thus the sampling loss is as follows:

$$\mathcal{L}_s = \sum_{i=1}^{N_s} \rho(||\hat{\boldsymbol{R}}_i - \mathrm{expm}(\boldsymbol{\theta}^*)||_2^2), \tag{13}$$

where N_s is the number of samples. ρ is an simple activation function for relaxation, which tolerates small deviations.

The total objective function is as follows:

$$\mathcal{L} = w_1\mathcal{L}_J + w_2\mathcal{L}_\beta + w_3\mathcal{L}_\theta + w_4\mathcal{L}_s, \tag{14}$$

where w_1, w_2, w_3, and w_4 are weight scalars.

3.4. Implementation details

We adopt ResNet-34 [16] and HRNet-W48 [58] as backbones. The ResNet backbone is followed by three deconvolutional layers to generate 3D heatmaps with the size of $64 \times 64 \times 64$ for keypoints and three MLPs for shape parameters $\boldsymbol{\beta}$ (10), distribution parameter $\boldsymbol{F}$ (216) and human scale (1). The feature from HRNet backbone is upsampled and directly followed by similar output branches with dimensions motioned above. The input image has a resolution of 256×256. The network is trained for 50 epochs with Adam and an initial learning rate of 1×10^{-3}, decayed with a factor of 10. w_1 and w_2 are set to 1. w_3 and w_4 are set to 0.1 and increased to 1 in the later stage of training.

4. Experiments

In this section, we demonstrate the effectiveness of our framework on human mesh recovery and multi-sensor fusion, evaluate our key designs via an ablation study, and discuss the limitation and future work of our method.

4.1. Datasets and metrics

To maintain the fairness of comparison, we adopt the same datasets and metrics as previous methods.
Human3.6M [18] provides 3D keypoints annotations, and the corresponding SMPL annotations are from MoSh [40]. We use (S1, S5, S6, S7, S8) for training and (S9, S11) for evaluation, following standard practice [22, 36].
3DPW [62] provides SMPL annotations. Following [36, 37], we add its training set only for experiments on it.
MS COCO [39] contains in-the-wild images and 2D keypoints annotations. We use its training set to improve the generalization ability of our method.
MPI-INF-3DHP [45] is a multi-view dataset that provides 3D keypoints annotations. We only use it for training.
AGORA [50] is a synthetic dataset with challenging scenes and SMPL annotations of adults and kids. Only when evaluating our algorithm on it will we add its training set.
TotalCapture [61] contains multi-view videos, IMUs and 3D keypoints annotations for the evaluation of sensor fusion algorithms. We follow [61, 72] to divide it.
Metrics include MPJPE, PA-MPJPE, and PVE all in mm. MPJPE measures the 3D keypoints error, while PA-MPJPE is similar to MPJPE except that a rigid alignment is performed at first. PVE measures the human mesh vertex error.

4.2. Human mesh recovery

Table 1 shows the evaluation results on public benchmarks. With either ResNet or HRNet as the backbone, our approach outperforms SOTA methods. Besides, we surpass the prior counterpart [55] by a large margin, indicating that our posterior estimation is easier to learn than the single prior. Table 2 shows the results on the AGORA test set. Our framework is more accurate than others, especially for kids. Note that for 2D datasets, despite that the pseudo-GT annotator of CLIFF [37] and EFT dataset [20] can be incorporated to further improve our performance, we only use the original keypoints supervision for fair comparisons.

Posterior effects: We compare different designs to thoroughly validate the posterior effects and the feature choice in our framework, as shown in Table 3, which include: (a) regressing the parameters $\boldsymbol{F}$ only without the keypoint branch; (b) solving rotations from keypoints via IK without probabilistic modeling; (c) deactivating the learned prior parameters in testing (*i.e.*, setting $\boldsymbol{F}$ to zero); (d) using the feature close to the end of the backbone to regress prior $\boldsymbol{F}$

Methods	3DPW			Human3.6M	
	PA ↓	MPJPE ↓	PVE ↓	PA ↓	MPJPE ↓
HMR (*R-50*) [22]	81.3	130.0	-	56.8	88.0
GraphCMR (*R-50*) [30]	70.2	-	-	50.1	-
SPIN (*R-50*) [29]	59.2	96.9	116.4	41.1	62.5
Sengupta. (*H-48*) [55]*	59.2	84.7*	-	-	-
HMR-EFT (*R-50*) [20]	52.4	-	-	43.9	-
I2L-MeshNet (*R-50*) [48]	58.6	93.2	-	41.7	55.7
SPEC (*R-50*) [28]	53.2	96.5	118.5	-	-
BEV (*H-32*) [60]	46.9	78.5	92.3	-	-
PARE (*H-32*) [27]	46.5	74.5	88.6	-	-
Graphormer (*H-64*) [38]	45.6	74.7	87.7	34.5	51.2
PyMAF (*H-48*) [70]	45.3	74.2	87.0	37.2	54.2
HybrIK (*R-34*) [36]	45.0	74.1	86.5	33.6	55.4
FastMETRO (*R-50*) [4]	48.3	77.9	90.6	37.3	53.9
FastMETRO (*H-64*) [4]	44.6	73.5	84.1	33.7	52.2
CLIFF (*R-50*) [37]	45.7	72.0	85.3	35.1	50.5
CLIFF (*H-48*) [37]	43.0	69.0	81.2	32.7	47.1
Ours (*R-34*)	**44.1**	**71.8**	**84.9**	**31.6**	**48.7**
Ours (*H-48*)	**40.6**	**68.3**	**79.4**	**29.1**	**45.7**

Table 1. **Results on standard benchmarks.** 'PA' is PA-MPJPE. '*R*' and '*H*' mean ResNet and HRNet. * with scale correction.

Methods	AGORA			
	MPJPE ↓	PVE ↓	Kid-MPJPE ↓	Kid-PVE ↓
HMR [22]	180.5	173.6	219.4	209.3
HMR-EFT [20]	165.4	159.0	202.7	193.5
SPIN [29]	153.4	148.9	191.7	186.7
PARE [27]	146.2	140.9	193.9	186.4
SPEC [28]	112.3	106.5	171.0	163.2
ROMP [59]	108.1	103.4	159.8	156.6
BEV [60]	105.3	100.7	129.1	125.9
Hand4Whole [47]	89.8	84.8	153.3	146.4
CLIFF [37]	81.0	76.0	94.1	89.6
HybrIK [36]	77.0	73.9	90.2	86.6
Ours	**74.4**	**70.9**	**84.5**	**80.5**
PLIKS [57]	71.5	67.3	88.3	84.2
NIKI [35]	67.3	63.9	83.9	80.2

Table 2. **Results on the AGORA test set.** The metrics with the prefix 'Kid-' are calculated only for kids, otherwise for all ages. Two concurrent works are shown in gray for completeness.

Designs	Human3.6M	
	PA-MPJPE ↓	MPJPE ↓
(a) W/o 3D keypoints	45.8	76.5
(b) W/o prior $\boldsymbol{F}$	43.2	63.6
(c) W/o prior $\boldsymbol{F}$ (in testing)	42.7	58.7
(d) Late feature for $\boldsymbol{F}$ and $\boldsymbol{\beta}$	29.9	48.6
(e) Early feature for $\boldsymbol{F}$ and $\boldsymbol{\beta}$	29.3	46.5
(f) Ours (full model)	**29.1**	**45.7**

Table 3. **Ablation study of designs on the Human3.6M dataset.**

and shape $\boldsymbol{\beta}$; (e) using the feature from the initial stage; (f) our full model with all branches and intermediate feature.

The performance of design (a) is not good since its estimation cannot align precisely with the image, and it is observed that this design exhibits a slower speed of convergence, reflecting the importance of the likelihood function

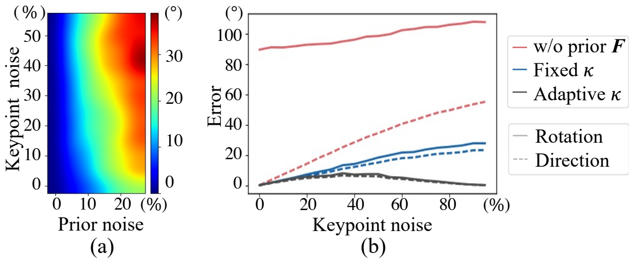

Figure 4. **Noise simulation.** (a) Rotation error v.s. the mixed noises of both prior and keypoint. (b) Rotation/direction error v.s. keypoint noise. The noise level is the ratio of the noise amplitude to the maximum of the variable. The simulation step is 5%. 'W/o prior $\boldsymbol{F}$' is an indirect strategy without probability. 'Fixed κ' uses predefined κ. 'Adaptive κ' means κ varies with the noise variance.

Noise amplitude	5mm	10mm	15mm	30mm	50mm
W/o prior $\boldsymbol{F}$	61.3	62.7	65.0	71.5	80.4
Posterior-based	48.5 (12.8↓)	49.6 (13.1↓)	51.4 (13.6↓)	54.5 (17.0↓)	59.0 (21.4↓)

Table 4. **Noise test on the Human3.6M dataset.** The 3D keypoints suffer from different levels of noises. MPJPE is reported.

from keypoints. The comparison between design (b) and (e) shows that the prior $\boldsymbol{F}$ is crucial in fusion. Furthermore, design (c) is better than design (b), indicating the prior $\boldsymbol{F}$ has guided the keypoints learning to some extent. The difference is more significant in MPJPE since the global rotation is supervised for design (c). As for the feature choice, the intermediate feature adopted by our framework shows slightly better performance than the late and early features.

Noise robustness: We evaluate the robustness of our framework when suffering from noise. Fig. 4 shows the simulation results with two metrics. The rotation error is the angle to be rotated from the estimated rotation $\hat{\boldsymbol{R}}$ to the ground-truth. The direction error is the angle between the estimated bone vector and the ground-truth. Fig. 4 (b) reveals that 'w/o prior $\boldsymbol{F}$' has a small direction error but a large rotation error for the unsolved twist, while the posterior strategy has a much smaller error even with a high keypoint noise level and shows a slower error growth rate. Note that the performance of a simple fixed κ is also acceptable. Fig. 4 (a) shows that when the prior noise level is less than 15%, the keypoint noise has little effect on the posterior result with the adaptive κ, reflecting the tolerance to noises of the posterior scheme. Table 4 shows the noise test on the Human3.6M dataset. From the difference listed in parentheses, when the noise level is higher, the error of the posterior method increases more slowly compared with the baseline without prior $\boldsymbol{F}$.

Samples illustration: Fig. 5 illustrates the samples from the posterior distribution. The right hand has a relatively

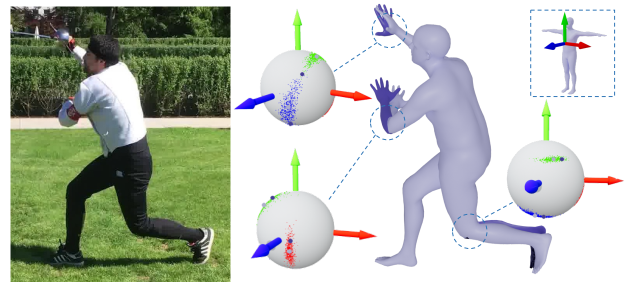

Figure 5. **Samples from the distribution.** The mode and samples of the relative rotations are shown. The light color indicates the mode, while the dark color indicates an extra sample. The canonical coordinate is at the top right as a reference.

Methods	Sensors	TotalCapture	
		PA-MPJPE ↓	MPJPE ↓
IMUPVH [10]	mv + IMUs	-	42.6
GeoFuse [72]	mv + IMUs	20.6	24.6
Ours	mv + IMUs	**19.4**	**23.5**
VIP [62]	sv + IMUs	26.0	-
Kalman filter	sv + IMUs	23.1	34.7
Ours	sv	29.0	42.1
Ours (w/o ref, *R-50*)	sv + IMUs	25.8	41.7
Ours (w/o ref, *H-48*)	sv + IMUs	22.3	38.8
Ours (*R-50*)	sv + IMUs	25.0	32.3
Ours (*H-48*)	sv + IMUs	**21.2**	**28.5**

Table 5. **Results on the TotalCapture dataset.** 'mv' and 'sv' denote multi-view and single-view. 'w/o ref' means lacking a reference skeleton (only a statistical one from training set is adopted), which would slightly weaken the performance of our framework.

large uncertainty on rotating around the X-axis due to the uncertain twist angle, as shown by the widespread blue and green samples. The left elbow has a vertical uncertainty on the direction of the forearm since the left hand cannot be easily determined. The left ankle has large confidence, except in the depth direction, therefore the green samples spread horizontally.

Fig. 6 shows the qualitative comparison with the SOTA methods. The indirect methods that use part segmentation [27] or 3D keypoints [36] perform well in most cases, but may suffer from wrong segmentation for distant people or generate unnatural poses. While the direct method CLIFF [37] may not align the image well for complicated scenes.

4.3. Multi-sensor fusion

We perform experiments on the TotalCapture dataset [61] using the feed-forward network directly without any iterative optimization. As the original dataset does not provide SMPL annotations, we adopt the human skeleton defined by 19 joints and 11 attached IMUs, as shown in Fig. 7. We choose the optimization-based method VIP [62] and the feature-fused method GeoFuse [72] as baselines.

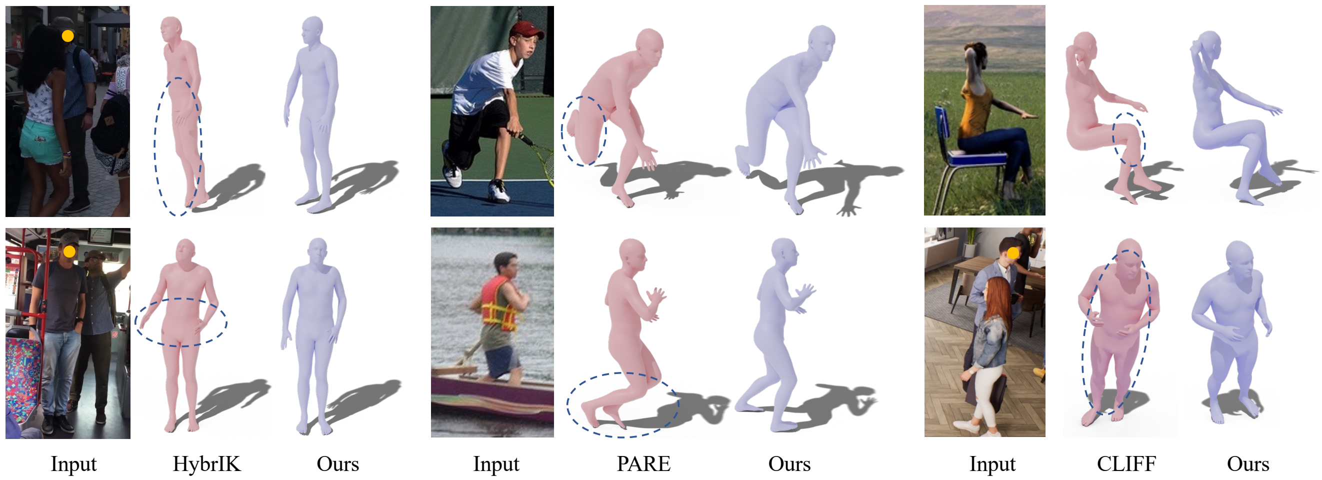

Figure 6. **Qualitative comparison.** The input images are from 3DPW [62], MS COCO [39], and AGORA [50], respectively. We compare our approach with SOTA methods including HybrIK [36], PARE [27], and CLIFF [37]. For images with multiple people, the person with a solid yellow circle on the face is estimated.

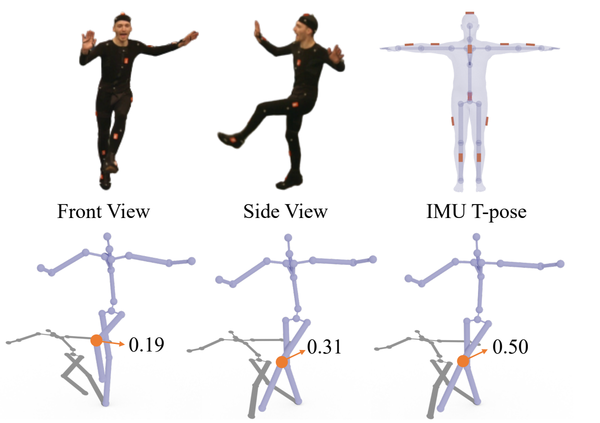

Figure 7. **An example of a multi-sensor case.** The estimations from the front-view camera, side-view camera, and IMUs are plotted. The normalized confidence score of the left knee is listed.

Table 5 shows the experimental results. Our method outperforms baseline methods, indicating the effectiveness of our posterior scheme. In the 'Kalman filter' setup, we apply a weighted sum to the 3D rotations separately obtained from images and IMUs. As the observation variance required for calculating the Kalman gain is unknown, we pick the weight pairs that can yield the best estimation through grid searching, *i.e.*, from $(0.7, 0.3)$ to $(0.3, 0.7)$. Note that unlike Kalman filter which fuses the observations in the testing stage, our method performs fusing in the training stage, and thus has the potential to obtain higher precision. Our framework is also more flexible than feature-level fusing methods since we do not require modifications of the backbone to incorporate new sensors.

Fig. 7 shows an example of a scene with multiple cameras and IMUs. The confidences from the two cameras are calculated via the normalized differential entropy of the estimated distribution parameters, while the confidence from IMUs is set to a relatively larger value since IMUs can provide accurate measurements. Note that other metrics that represent the uncertainty from the distribution can also be adopted. It can be observed that the front view produces erroneous knee bending due to the depth ambiguity, therefore its confidence is lower than the side view. As a result, the fused result will be less affected by the noisy estimation.

4.4. Limitation and future work

Our work has several limitations. First, we only consider the uncertainty of poses, not including that of shapes, which can also be modeled as probability distributions. Second, we model the human joint independently, which is only affected by the parent node. Therefore, how to derive the analytical form of joint rotations conditioned on other hierarchical joints to incorporate anatomical constraints explicitly is still unsolved. Besides, with the single-view uncertainty, the temporal extension also deserves further investigation.

5. Conclusion

In this paper, we derive a novel analytical posterior probability for human joint rotations in a Bayesian manner and prove the property that the posteriors are more concentrated than the priors. Based on the derivation, we propose a new framework for human mesh recovery by leveraging the learned posteriors, which has high precision and robustness, outperforming existing SOTA baselines. Furthermore, our framework can be seamlessly incorporated with additional sensors in the training due to its Bayesian nature. Our research also provides a sound foundation for incorporating more advanced prior conditions or physical constraints.

References

[1] 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*, volume 33, pages 20496–20507, 2020. 2

[2] 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*, pages 561–578, 2016. 2

[3] Hansheng Chen, Pichao Wang, Fan Wang, Wei Tian, Lu Xiong, and Hao Li. Epro-pnp: Generalized end-to-end probabilistic perspective-n-points for monocular object pose estimation. In *CVPR*, pages 2781–2790, 2022. 4

[4] Junhyeong Cho, Kim Youwang, and Tae-Hyun Oh. Cross-attention of disentangled modalities for 3d human mesh recovery with transformers. In *ECCV*, 2022. 6

[5] Hongsuk Choi, Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. Beyond static features for temporally consistent 3d human pose and shape from a video. In *CVPR*, pages 1964–1973, 2021. 2

[6] Hongsuk Choi, Gyeongsik Moon, and Kyoung Mu Lee. Pose2mesh: Graph convolutional network for 3d human pose and mesh recovery from a 2d human pose. In *ECCV*, pages 769–787, 2020. 2

[7] Junting Dong, Qi Fang, Wen Jiang, Yurou Yang, Hujun Bao, and Xiaowei Zhou. Fast and robust multi-person 3d pose estimation and tracking from multiple views. *IEEE TPAMI*, 44(10):6981–6992, 2021. 2

[8] Thomas D Downs. Orientation statistics. *Biometrika*, 59(3):665–676, 1972. 1, 3

[9] Qi Fang, Qing Shuai, Junting Dong, Hujun Bao, and Xiaowei Zhou. Reconstructing 3d human pose by watching humans in the mirror. In *CVPR*, pages 12814–12823, 2021. 2

[10] Andrew Gilbert, Matthew Trumble, Charles Malleson, Adrian Hilton, and John Collomosse. Fusing visual and inertial sensors with semantics for 3d human pose estimation. *IJCV*, 127(4):381–397, 2019. 2, 3, 7

[11] Igor Gilitschenski, Roshni Sahoo, Wilko Schwarting, Alexander Amini, Sertac Karaman, and Daniela Rus. Deep orientation uncertainty learning based on a bingham loss. In *ICLR*, 2019. 4

[12] Jared Marshall Glover. *The quaternion Bingham distribution, 3D object detection, and dynamic manipulation*. PhD thesis, MIT, 2014. 4

[13] Shanyan Guan, Jingwei Xu, Michelle Z He, Yunbo Wang, Bingbing Ni, and Xiaokang Yang. Out-of-domain human mesh reconstruction via dynamic bilevel online adaptation. *IEEE TPAMI*, 2022. 2

[14] Riza Alp Guler and Iasonas Kokkinos. Holopose: Holistic 3d human reconstruction in-the-wild. In *CVPR*, pages 10884–10894, 2019. 2

[15] Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J Black. Resolving 3d human pose ambiguities with 3d scene constraints. In *ICCV*, pages 2282–2292, 2019. 2

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 5

[17] Roger A Horn and Charles R Johnson. *Matrix analysis*. Cambridge university press, 2012. 4

[18] 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 TPAMI*, 36(7):1325–1339, 2013. 6

[19] Umar Iqbal, Kevin Xie, Yunrong Guo, Jan Kautz, and Pavlo Molchanov. Kama: 3d keypoint aware body mesh articulation. In *3DV*, pages 689–699, 2021. 1

[20] 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*, pages 42–52, 2021. 2, 6

[21] Rudolph Emil Kalman. A new approach to linear filtering and prediction problems. 1960. 2, 3

[22] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *CVPR*, pages 7122–7131, 2018. 1, 2, 6

[23] Angjoo Kanazawa, Jason Y Zhang, Panna Felsen, and Jitendra Malik. Learning 3d human dynamics from video. In *CVPR*, pages 5614–5623, 2019. 2

[24] John T Kent, Asaad M Ganeiber, and Kanti V Mardia. A new method to simulate the bingham and related distributions in directional data analysis with applications. *arXiv*, 2013. 5

[25] CG Khatri and Kanti V Mardia. The von mises–fisher matrix distribution in orientation statistics. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):95–106, 1977. 1, 2, 3

[26] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. Vibe: Video inference for human body pose and shape estimation. In *CVPR*, pages 5253–5263, 2020. 2

[27] Muhammed Kocabas, Chun-Hao P Huang, Otmar Hilliges, and Michael J Black. Pare: Part attention regressor for 3d human body estimation. In *ICCV*, pages 11127–11137, 2021. 1, 2, 6, 7, 8

[28] Muhammed Kocabas, Chun-Hao P Huang, Joachim Tesch, Lea Müller, Otmar Hilliges, and Michael J Black. Spec: Seeing people in the wild with an estimated camera. In *ICCV*, pages 11035–11045, 2021. 6

[29] 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*, pages 2252–2261, 2019. 1, 2, 6

[30] Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. In *CVPR*, pages 4501–4510, 2019. 2, 6

[31] Nikos Kolotouros, Georgios Pavlakos, Dinesh Jayaraman, and Kostas Daniilidis. Probabilistic modeling for human mesh recovery. In *ICCV*, pages 11605–11614, 2021. 1, 2

[32] 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*, pages 6050–6059, 2017. 2

[33] Taeyoung Lee. Bayesian attitude estimation with the matrix fisher distribution on so (3). *IEEE Transactions on Automatic Control*, 63(10):3377–3392, 2018. 4

[34] Jake Levinson, Carlos Esteves, Kefan Chen, Noah Snavely, Angjoo Kanazawa, Afshin Rostamizadeh, and Ameesh Makadia. An analysis of svd for deep rotation estimation. In *NeurIPS*, volume 33, pages 22554–22565, 2020. 1, 2, 4

[35] Jiefeng Li, Siyuan Bian, Qi Liu, Jiasheng Tang, Fan Wang, and Cewu Lu. Niki: Neural inverse kinematics with invertible neural networks for 3d human pose and shape estimation. In *CVPR*, 2023. 6

[36] Jiefeng Li, Chao Xu, Zhicun Chen, Siyuan Bian, Lixin Yang, and Cewu Lu. Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation. In *CVPR*, pages 3383–3393, 2021. 1, 2, 6, 7, 8

[37] 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*, pages 590–606, 2022. 1, 2, 6, 7, 8

[38] Kevin Lin, Lijuan Wang, and Zicheng Liu. Mesh graphormer. In *ICCV*, pages 12939–12948, 2021. 6

[39] 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*, pages 740–755, 2014. 6, 8

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

[41] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM TOG*, 34(6):1–16, 2015. 2, 5

[42] Meysam Madadi, Hugo Bertiche, and Sergio Escalera. Smplr: Deep smpl reverse for 3d human pose and shape recovery. *arXiv*, 2018. 1, 2

[43] Charles Malleson, Andrew Gilbert, Matthew Trumble, John Collomosse, Adrian Hilton, and Marco Volino. Real-time full-body motion capture from video and imus. In *3DV*, pages 449–457, 2017. 3

[44] Kanti V Mardia, Peter E Jupp, and KV Mardia. *Directional statistics*, volume 2. Wiley Online Library, 2000. 2, 4

[45] 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*, pages 506–516, 2017. 6

[46] David Mohlin, Josephine Sullivan, and Gérald Bianchi. Probabilistic orientation estimation with matrix fisher distributions. In *NeurIPS*, volume 33, pages 4884–4893, 2020. 4

[47] Gyeongsik Moon, Hongsuk Choi, and Kyoung Mu Lee. Accurate 3d hand pose estimation for whole-body 3d human mesh estimation. In *CVPRW*, pages 2308–2317, 2022. 6

[48] Gyeongsik Moon and Kyoung Mu Lee. I2l-meshnet: Image-to-lixel prediction network for accurate 3d human pose and mesh estimation from a single rgb image. In *ECCV*, pages 752–768, 2020. 1, 2, 6

[49] Mohamed Omran, Christoph Lassner, Gerard Pons-Moll, Peter Gehler, and Bernt Schiele. Neural body fitting: Unifying deep learning and model based human pose and shape estimation. In *3DV*, pages 484–494, 2018. 2

[50] 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*, pages 13468–13478, 2021. 6, 8

[51] 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*, pages 10975–10985, 2019. 1, 2, 5

[52] Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. Learning to estimate 3d human pose and shape from a single color image. In *CVPR*, pages 459–468, 2018. 1, 2

[53] Davis Rempe, Tolga Birdal, Aaron Hertzmann, Jimei Yang, Srinath Sridhar, and Leonidas J Guibas. Humor: 3d human motion model for robust pose estimation. In *ICCV*, pages 11488–11499, 2021. 1, 2

[54] Javier Romero, Dimitris Tzionas, and Michael J Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM TOG*, 36(6), 2017. 2, 5

[55] 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*, pages 11219–11229, 2021. 1, 2, 5, 6

[56] Akash Sengupta, Ignas Budvytis, and Roberto Cipolla. Probabilistic 3d human shape and pose estimation from multiple unconstrained images in the wild. In *CVPR*, pages 16094–16104, 2021. 1, 2

[57] Karthik Shetty, Annette Birkhold, Srikrishna Jaganathan, Norbert Strobel, Markus Kowarschik, Andreas Maier, and Bernhard Egger. Pliks: A pseudo-linear inverse kinematic solver for 3d human body estimation. *arXiv*, 2022. 6

[58] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, pages 5693–5703, 2019. 5

[59] Yu Sun, Qian Bao, Wu Liu, Yili Fu, Michael J Black, and Tao Mei. Monocular, one-stage, regression of multiple 3d people. In *ICCV*, pages 11179–11188, 2021. 6

[60] 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*, pages 13243–13252, 2022. 6

[61] Matthew Trumble, Andrew Gilbert, Charles Malleson, Adrian Hilton, and John Collomosse. Total capture: 3d human pose estimation fusing video and inertial sensors. In *BMVC*, pages 1–13, 2017. 3, 6, 7

[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*, pages 601–617, 2018. 3, 6, 7, 8

[63] Timo Von Marcard, Gerard Pons-Moll, and Bodo Rosenhahn. Human pose estimation from video and imus. *IEEE TPAMI*, 38(8):1533–1547, 2016. 3

[64] Yuanlu Xu, Song-Chun Zhu, and Tony Tung. Denserac: Joint 3d pose and shape estimation by dense render-and-compare. In *ICCV*, pages 7760–7770, 2019. 2

[65] Xinyu Yi, Yuxiao Zhou, Marc Habermann, Soshi Shimada, Vladislav Golyanik, Christian Theobalt, and Feng Xu. Physical inertial poser (pip): Physics-aware real-time human motion tracking from sparse inertial sensors. In *CVPR*, pages 13167–13178, 2022. 2

[66] Xinyu Yi, Yuxiao Zhou, and Feng Xu. Transpose: real-time 3d human translation and pose estimation with six inertial sensors. *ACM TOG*, 40(4):1–13, 2021. 2

[67] Yingda Yin, Yingcheng Cai, He Wang, and Baoquan Chen. Fishermatch: Semi-supervised rotation regression via entropy-based filtering. In *CVPR*, pages 11164–11173, 2022. 4

[68] 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*, pages 14484–14493, 2021. 2

[69] Wang Zeng, Wanli Ouyang, Ping Luo, Wentao Liu, and Xiaogang Wang. 3d human mesh regression with dense correspondence. In *CVPR*, pages 7054–7063, 2020. 2

[70] 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*, pages 11446–11456, 2021. 2, 6

[71] Yuxiang Zhang, Liang An, Tao Yu, Xiu Li, Kun Li, and Yebin Liu. 4d association graph for realtime multi-person motion capture using multiple video cameras. In *CVPR*, pages 1324–1333, 2020. 2

[72] Zhe Zhang, Chunyu Wang, Wenhu Qin, and Wenjun Zeng. Fusing wearable imus with multi-view images for human pose estimation: A geometric approach. In *CVPR*, pages 2200–2209, 2020. 3, 6, 7

[73] Zhe Zhang, Chunyu Wang, Weichao Qiu, Wenhu Qin, and Wenjun Zeng. Adafuse: Adaptive multiview fusion for accurate human pose estimation in the wild. *IJCV*, 129(3):703–718, 2021. 2

[74] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *CVPR*, pages 5745–5753, 2019. 1, 2, 4