Supplementary Material: Learning Analytical Posterior Probability for Human Mesh Recovery

In this supplementary material, we provide more explanations about distribution parameters (Sec. 1), and detailed derivations of crucial theorems (Sec. 2). We also describe the sampling procedure (Sec. 3). To demonstrate the probability models intuitively, we visualize them contrastively (Sec. 4). For some unmentioned details, we simply list them for completeness (Sec. 5).

1. Parameter explanation

1.1. matrix Fisher distribution

In this part, we try to illustrate the parameter $\boldsymbol{F}$ of matrix Fisher distribution from a simple geometric perspective. As stated in the main manuscript, the probability density function of matrix Fisher distribution $\mathcal{MF}(\cdot)$ over $\mathcal{SO}(n)$ [1, 5] is as follows:

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

Suppose there are two sets of unit vectors $\{\boldsymbol{l}_i\}$ and $\{\boldsymbol{d}_i\}$, and the vectors in $\{\boldsymbol{l}_i\}$ are linearly independent. To calculate a rotation $\boldsymbol{R}$ that can transform $\{\boldsymbol{l}_i\}$ to $\{\boldsymbol{d}_i\}$, the estimation can be derived by minimizing the L_2 distance:

$$\begin{aligned}
\hat{\boldsymbol{R}} &= \arg\min_{\boldsymbol{R}} \sum_i ||\boldsymbol{d}_i - \boldsymbol{R}\boldsymbol{l}_i||_2^2 \\
&= \arg\min_{\boldsymbol{R}} \sum_i (\boldsymbol{d}_i^T\boldsymbol{d}_i + \boldsymbol{l}_i^T\boldsymbol{l}_i - 2\boldsymbol{d}_i^T\boldsymbol{R}\boldsymbol{l}_i) \\
&= \arg\max_{\boldsymbol{R}} \sum_i \boldsymbol{d}_i^T\boldsymbol{R}\boldsymbol{l}_i = \arg\max_{\boldsymbol{R}} \sum_i tr(\boldsymbol{l}_i\boldsymbol{d}_i^T\boldsymbol{R}) \\
&= \arg\max_{\boldsymbol{R}}\ tr[\sum_i \boldsymbol{l}_i\boldsymbol{d}_i^T\boldsymbol{R}].
\end{aligned}$$

If we define $\boldsymbol{F} \triangleq s\sum_i \boldsymbol{d}_i\boldsymbol{l}_i^T$ with the scale factor s, then the above equation has a similar formulation as the standard matrix Fisher distribution. In this sense, we have:

$$\boldsymbol{F}' = \boldsymbol{F} + \kappa\boldsymbol{d}\boldsymbol{l}^T = s\sum_i(\boldsymbol{d}_i\boldsymbol{l}_i^T) + \kappa\boldsymbol{d}\boldsymbol{l}^T,$$

which means the original parameter $\boldsymbol{F}$ can be comprehend as the sum of the outer product of some paired vectors ($\boldsymbol{l}_i$, $\boldsymbol{d}_i$), while the updated parameter $\boldsymbol{F}'$ is equivalent to fusing more paired vectors on the basis of the original data.

1.2. von Mises-Fisher distribution

In this part, we illustrate the relation between normal distribution and von Mises-Fisher distribution. As stated in the main manuscript, the probability density function of von Mises-Fisher distribution $\mathcal{VMF}(\cdot)$ [6] is as follows:

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

Remark. *The von Mises-Fisher distribution becomes the uniform distribution on the sphere for $\kappa = 0$, and it approximates the wrapped normal distribution with the same mean $\boldsymbol{m}$ and variance κ^{-1} for a large κ:*

$$\begin{aligned}
&\mathcal{VMF}(\boldsymbol{m}, 0) = \mathcal{U}(\mathcal{S}^{n-1}), \\
&\mathcal{VMF}(\boldsymbol{m},\kappa) \approx \mathcal{WN}(\boldsymbol{m},\kappa^{-1}), \quad \kappa \gg 0.
\end{aligned}$$

Proof. It's easy to prove the statement for $\kappa = 0$. Therefore only the case of a large κ is discussed here. Note that both $\boldsymbol{m}$ and $\boldsymbol{d}$ can be assumed as unit vectors, leaving their scale factors to the normalizing constant. For convenience, we represent $\boldsymbol{m}$ and $\boldsymbol{d}$ as trigonometric functions, *i.e.*, $(cos\ \theta_{\boldsymbol{m}}, sin\ \theta_{\boldsymbol{m}})$ and $(cos\ \theta_{\boldsymbol{d}}, sin\ \theta_{\boldsymbol{d}})$, respectively. Therefore, the probability density function is as follows:

$$p(\theta_{\boldsymbol{d}};\kappa,\theta_{\boldsymbol{m}}) = \frac{1}{c(\kappa)} exp(\kappa\cdot cos(\theta_{\boldsymbol{d}} - \theta_{\boldsymbol{m}})).$$

Let $\xi = \kappa^{1/2}(\theta_{\boldsymbol{d}} - \theta_{\boldsymbol{m}})$, then we can derive its probability formulation as follows:

$$\begin{aligned}
p(\xi) &\propto exp(-\kappa)\cdot exp(\kappa\cdot cos(\kappa^{-1/2}\xi)) \\
&\approx exp(-\kappa + \kappa(1 - \frac{1}{2}\kappa^{-1}\xi^2)) \\
&= exp(-\frac{1}{2}\xi^2) \sim \mathcal{N}(0,1),
\end{aligned}$$

where $exp(-\kappa)$ is a constant term independent of $\theta_{\boldsymbol{d}}$ for transformation, and the first two terms of Taylor series of $cos(\cdot)$ is used. Based on the fact that $p(\xi)$ follows the standard normal distribution, we can derive that $\boldsymbol{d} \sim \mathcal{VMF}(\boldsymbol{m},\kappa) \approx \mathcal{WN}(\boldsymbol{m},\kappa^{-1})$ for large κ (the condition for Taylor expansion). Note that the wrapped normal distribution is used here due to periodicity. □

2. Theorem derivation

In this section, we derive some theorems about the posterior probability and the corresponding property thoroughly.

2.1. Posterior distribution

Remark. *For $\boldsymbol{l}, \boldsymbol{d} \in \mathbb{R}^n$, $\boldsymbol{R} \in \mathbb{R}^{n\times n}$, the following equation holds:*

$$tr(\boldsymbol{l}\boldsymbol{d}^T\boldsymbol{R}) = \boldsymbol{l}^T\boldsymbol{R}^T\boldsymbol{d}.$$

Proof.

$$\begin{aligned} left &= tr(\boldsymbol{l}\boldsymbol{d}^T\boldsymbol{R}) \\ &= \sum_i (\boldsymbol{l}\boldsymbol{d}^T\boldsymbol{R})_{ii} = \sum_i \boldsymbol{l}_i(\boldsymbol{d}^T\boldsymbol{R})_i = \boldsymbol{l}^T(\boldsymbol{d}^T\boldsymbol{R})^T \\ &= \boldsymbol{l}^T\boldsymbol{R}^T\boldsymbol{d} = right. \end{aligned}$$

□

Theorem 2.1. *The analytical probability of rotation $\boldsymbol{R} \in \mathcal{SO}(n)$ conditioned on bone direction $\boldsymbol{d} \in \mathcal{S}^{n-1}$ satisfies:*

$$p(\boldsymbol{R}|\boldsymbol{d}) \sim \mathcal{MF}(\boldsymbol{F} + \kappa\boldsymbol{d}\boldsymbol{l}^T).$$

Proof.

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

□

Theorem 2.2. *(General form) The analytical probability of rotation $\boldsymbol{R} \in \mathcal{SO}(n)$ conditioned on directional observations $\boldsymbol{d} \in \mathcal{S}^{n-1}$ and rotational observations $\boldsymbol{D} \in \mathcal{SO}(n)$ satisfies:*

$$p(\boldsymbol{R}|\{\boldsymbol{d}_i, \boldsymbol{D}_j\}) \sim \mathcal{MF}(\boldsymbol{F} + \sum_i \kappa_i\boldsymbol{d}_i\boldsymbol{l}_i^T + \sum_j \boldsymbol{D}_j\boldsymbol{K}_j^T).$$

Proof. Similar to the bone direction, the rotational observations from other sensors can also become the observation variables $\boldsymbol{D}$ of the latent variable $\boldsymbol{R}$:

$$\begin{aligned} p(\boldsymbol{D}|\boldsymbol{R}) &= \frac{1}{c}exp(tr[(\boldsymbol{R}\boldsymbol{K})^T\boldsymbol{D}]) = \frac{1}{c}exp(tr[\boldsymbol{K}^T\boldsymbol{R}^T\boldsymbol{D}]) \\ &= \frac{1}{c}exp(tr[\boldsymbol{D}^T\boldsymbol{R}\boldsymbol{K}]) = \frac{1}{c}exp(tr[\boldsymbol{K}\boldsymbol{D}^T\boldsymbol{R}]) \\ &= \frac{1}{c}exp(tr[(\boldsymbol{D}\boldsymbol{K}^T)^T\boldsymbol{R}]) \sim \mathcal{MF}(\boldsymbol{D}\boldsymbol{K}^T), \end{aligned}$$

where the parameter is decomposed into a mean term $\boldsymbol{R}$ and a concentration term $\boldsymbol{K}$ at first, and we use some properties of square matrices: (i) $tr(A) = tr(A^T)$; (ii) $tr(AB) = tr(BA)$. Therefore, the theorem can be derived as follows:

$$\begin{aligned} p(\boldsymbol{R}|\{\boldsymbol{d}_i, \boldsymbol{D}_j\}) &= \frac{p(\boldsymbol{R})\cdot p(\{\boldsymbol{d}_i, \boldsymbol{D}_j\}|\boldsymbol{R})}{p(\{\boldsymbol{d}_i, \boldsymbol{D}_j\})} \\ &\propto p(\boldsymbol{R})\cdot\prod_i p(\boldsymbol{d}_i|\boldsymbol{R})\cdot\prod_j p(\boldsymbol{D}_j|\boldsymbol{R}) \\ &= \frac{1}{c}exp(tr[(\boldsymbol{F} + \sum_i \kappa_i\boldsymbol{d}_i\boldsymbol{l}_i^T + \sum_j \boldsymbol{D}_j\boldsymbol{K}_j^T)^T\boldsymbol{R}]) \\ &\sim \mathcal{MF}(\boldsymbol{F} + \sum_i \kappa_i\boldsymbol{d}_i\boldsymbol{l}_i^T + \sum_j \boldsymbol{D}_j\boldsymbol{K}_j^T) \end{aligned}$$

□

Theorem 2.1 and 2.2 provide the proofs of Eq. (6), (9) in the main manuscript, respectively.

2.2. Property

Here we provide more details about the proof of the property (Eq. (7) in the main manuscript). First, we state a lemma for the later proof.

Lemma 2.3. *(Interlacing theorem) Suppose $n \geq 2$, $\boldsymbol{K} \in \mathbb{R}^{n\times n}$ is Hermitian, and $\boldsymbol{l} \in \mathbb{R}^n$ is nonzero. Then the eigenvalues satisfy the following inequality:*

$$\begin{aligned} &\lambda_1(\boldsymbol{K}) \leq \lambda_1(\boldsymbol{K} + \boldsymbol{l}\boldsymbol{l}^*) \leq \lambda_2(\boldsymbol{K}) \leq \cdots \leq \\ &\lambda_n(\boldsymbol{K}) \leq \lambda_n(\boldsymbol{K} + \boldsymbol{l}\boldsymbol{l}^*), \end{aligned}$$

which means the eigenvalues for a Hermitian perturbation with rank 1 of a Hermitian matrix are larger than the corresponding original eigenvalues in an interlaced manner [3].

Back to the proof of the property. Since the mean term $\boldsymbol{M}$ is orthogonal and satisfies $\boldsymbol{M}\boldsymbol{l} = \boldsymbol{d}$, we can decompose the posterior parameter $\boldsymbol{F}'$ as follows:

$$\begin{aligned} \boldsymbol{F}' &= \boldsymbol{F} + \kappa\boldsymbol{d}\boldsymbol{l}^T = \boldsymbol{M}\boldsymbol{K} + \kappa\boldsymbol{M}\boldsymbol{l}\boldsymbol{l}^T = \boldsymbol{M}(\boldsymbol{K} + \kappa\boldsymbol{l}\boldsymbol{l}^T) \\ &= \boldsymbol{M}\boldsymbol{K}', \end{aligned}$$

where $\boldsymbol{K} = \boldsymbol{V}\Delta\boldsymbol{S}\boldsymbol{V}^T$ is a real symmetric matrix considering that both Δ and $\boldsymbol{S}$ are diagonal. Besides, $\boldsymbol{l}\boldsymbol{l}^T$, the

outer product of $\boldsymbol{l}$ with itself, is a symmetric matrix with rank 1. Thus $\boldsymbol{K}'$ is also real symmetric, *i.e.*, Hermitian. From Lemma 2.3, we can get the conclusion that $\boldsymbol{K}'$ has larger eigenvalues than $\boldsymbol{K}$, therefore the singular values of the posterior parameter $\boldsymbol{F}'$ is larger than those of the prior parameter $\boldsymbol{F}$.

Fig. 1 illustrates the convergence performance regarding the accuracy of SMPL poses for two different settings, respectively, estimating prior $\boldsymbol{F}$ without a keypoint branch and posterior $\boldsymbol{F}'$ with a keypoint branch. The latter clearly converges faster and better.

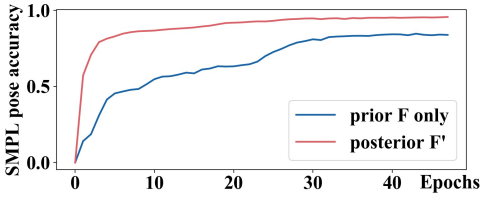

Figure 1. Convergence comparison.

3. Sampling

As we stated in the main manuscript, because of the equivalence between Bingham distribution on $\mathcal{S}^3$ and matrix Fisher distribution on $\mathcal{SO}(3)$, we sample the Bingham distribution instead via differentiable rejection sampling [4, 7]. Algorithm 1 lists the detailed procedure. The proposal distribution in rejection sampling is an angular central Gaussian (ACG) distribution, the sampling of which is implemented with the reparameterization trick.

Algorithm 1: $\mathcal{MF}$ Rejection Sampling

Input: $\boldsymbol{F} \in \mathbb{R}^{3\times3}$
Output: $\boldsymbol{R}_i \sim \mathcal{MF}(\boldsymbol{F})$
$\boldsymbol{U}, \boldsymbol{S}, \boldsymbol{V}^T = \text{SVD}(\boldsymbol{F})$, where $\boldsymbol{S} = \text{diag}(s_1, s_2, s_3)$
▷ Sample matrix Fisher via Bingham
$\boldsymbol{A} = \text{diag}(0, 2(s_2+s_3), 2(s_1+s_3), 2(s_1+s_2))$
Solve b, $s.t.$ $\sum_{i=1}^{4} \frac{1}{b+2\boldsymbol{A}_i} - 1 = 0$
$\boldsymbol{\Omega} = \boldsymbol{I}_4 + \frac{2}{b}\boldsymbol{A}$
$M = (\frac{4}{b})^2 \exp(\frac{b-4}{2})$
▷ Sample Bingham via ACG as proposal distribution
repeat
  Sample $w \sim \mathcal{U}(0,1)$
  Sample $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}_4, \boldsymbol{I}_4)$ ▷ reparameterization
  $\boldsymbol{y} = (\boldsymbol{\Omega}^{-1})^{\frac{1}{2}}\boldsymbol{\epsilon}$
  Get a proposal $\boldsymbol{x} = \frac{\boldsymbol{y}}{\|\boldsymbol{y}\|}$ s.t. $\boldsymbol{x} \in \mathcal{S}^3$
until $w < \frac{\exp(-\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x})}{M(\boldsymbol{x}^T\boldsymbol{\Omega}\boldsymbol{x})^{-2}}$;
$\boldsymbol{R}_i = \boldsymbol{U}\boldsymbol{X}\boldsymbol{V}^T$, where $\boldsymbol{X} = quat\_to\_mat(\boldsymbol{x})$
return $\boldsymbol{R}_i$

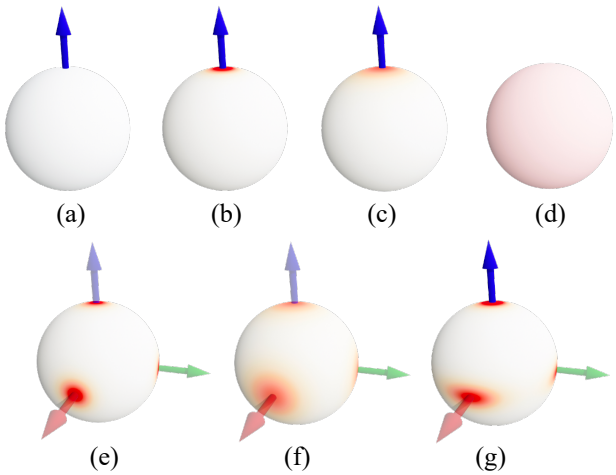

Figure 2. **Comparison of different models.**

4. Figure illustration

In this part, we illustrate Fig. 3 in the main manuscript thoroughly. As shown in Fig. 2, relevant spheres represent the following underlying models: (a) deterministic direction without probabilistic modeling (*e.g.*, IK estimations); (b) directional distribution with large confidence (*e.g.*, gravity sensors); (c) directional distribution with small confidence; (d) uniform distribution on the sphere; (e) rotational distribution with large confidence (*e.g.*, gyroscopes); (f) rotational distribution with small confidence; (g) rotational posterior distribution conditioned on directional information.

The eigenvalues of (e) are larger than those of (f), thus, (e) shows a more concentrated distribution. For (g), since the directional information is provided and fused, the distribution is more concentrated on the blue arrow, so the regions around the red and green arrows show oval shapes (not circles). (g) is easier to learn compared with directly regressing parameters (validated by Fig. 1), because the prior and the keypoints not only mutually narrow down the region of solutions but also have the potential to recover the ground-truth even for noisy cases.

5. Others

In this section, we provide some unmentioned details.

Error definition: Given the estimated rotation $\hat{\boldsymbol{R}}$, ground-truth $\boldsymbol{R}^*$, and canonical unit vector $\boldsymbol{l}$, the rotation error Err_R and direction error Err_d in the simulation experiment are as follows:

$$Err_R = \arccos(\frac{tr(\boldsymbol{R}^*\hat{\boldsymbol{R}}^T) - 1}{2}),$$
$$Err_d = \arccos(< \boldsymbol{R}^*\boldsymbol{l}, \hat{\boldsymbol{R}}\boldsymbol{l} >).$$

Module	Block	Input	Output
Conv(R)	ResNet	Image $(256, 256, 3)$	$(f_c, 8, 8)$
Deconv(R)	{Deconv+BN+ReLU} $\times 3$	$(f_c, 8, 8)$	$(f_c, 64, 64)$
	1×1 Conv	$(f_c, 64, 64)$	$(29 \times 64, 64, 64)$
	Soft-argmax	$(29 \times 64, 64, 64)$, $\boldsymbol{s}$	$\boldsymbol{J}(29, 3)$
Conv(H)	HRNet	Image $(256, 256, 3)$	$(f_c, 64, 64)$
Deconv(H)	1×1 Conv	$(f_c, 64, 64)$	$(29 \times 64, 64, 64)$
	Soft-argmax	$(29 \times 64, 64, 64)$, $\boldsymbol{s}$	$\boldsymbol{J}(29, 3)$
MLP feature	AvgPool + Fc1 + Dropout	$(f_c, 64, 64)$	(f'_c)
	Fc2 + Dropout	(f'_c)	(f'_c)
MLP($\boldsymbol{\beta}$)	Fc($\boldsymbol{\beta}$)	(f'_c)	$\boldsymbol{\beta}(10)$
MLP($\boldsymbol{F}$)	Fc($\boldsymbol{F}$)	(f'_c)	$\boldsymbol{F}(216)$
MLP($\boldsymbol{s}$)	Fc($\boldsymbol{s}$)	(f'_c)	$\boldsymbol{s}(1)$
Output	shape $\boldsymbol{\beta}(10)$, parameter $\boldsymbol{F}(216)$, 3D keypoints $\boldsymbol{J}(29, 3)$		

Table 1. **Network architecture.**

Differential entropy: For the matrix Fisher distribution, the differential entropy H_F can be a representation of its uncertainty [2, 8] and calculated via Bingham distribution:

$$H_F = \log c_B - \sum_{i=1}^{3} k_i \frac{\partial c_B}{\partial k_i} - \log(2\pi^2),$$

where c_B, $\boldsymbol{K} = [k_1, k_2, k_3]$ are the normalizing constant and the concentration parameters of Bingham distribution, respectively. We normalize the entropy from multiple sensors, since the differential entropy has not a certain range. Note that other metrics also deserve exploration.

Network architecture: Table 1 lists the structure of our network and the corresponding feature dimensions. f_c is the number of feature channel relevant to different backbones. f'_c is the MLP feature dimension. Fc1 and Fc2 are shared by MLP branches.

Model size and running time. The total numbers of trainable parameters and the running time on an RTX 2080ti GPU (batch size 1) are 27.8M / 19ms (R-34 backbone) and 73.4M / 46ms (H-48 backbone), respectively. The model size and speed are of the same level as other frameworks.

References

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

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

[3] Roger A Horn and Charles R Johnson. *Matrix analysis*. Cambridge university press, 2012. 2

[4] 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. 3

[5] 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

[6] Kanti V Mardia, Peter E Jupp, and KV Mardia. *Directional statistics*, volume 2. Wiley Online Library, 2000. 1

[7] 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. 3

[8] 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