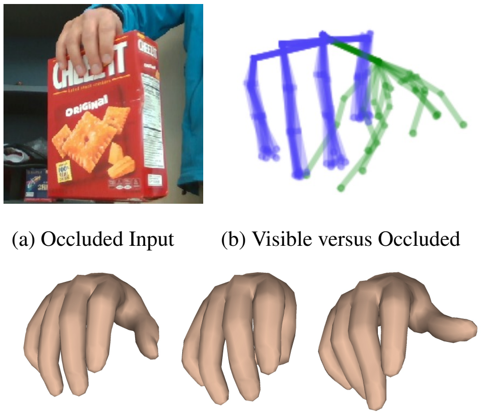
MHEntropy: Entropy Meets Multiple Hypotheses for Pose and Shape Recovery

Rongyu Chen[*] Linlin Yang[*] Angela Yao
National University of Singapore
{rchen, yangll, ayao}@comp.nus.edu.sg

Abstract

For monocular RGB-based 3D pose and shape estimation, multiple solutions are often feasible due to factors like occlusions and truncations. This work presents a multi-hypothesis probabilistic framework by optimizing the Kullback–Leibler divergence (KLD) between the data and model distribution. Our formulation reveals a connection between the pose entropy and diversity in the multiple hypotheses that has been neglected by previous works. For a comprehensive evaluation, besides the best hypothesis (BH) metric, we factor in visibility for evaluating diversity. Additionally, our framework is label-friendly – it can be learned from only partial 2D keypoints, such as visible keypoints. Experiments on both ambiguous and real-world benchmarks demonstrate that our method outperforms other state-of-the-art multi-hypothesis methods. The project page is at https://gloryyrolg.github.io/MHEntropy.



(a) Occluded Input (b) Visible versus Occluded

(c) Multiple Hypotheses

Figure 1. Our method estimates diverse and feasible hypotheses for occluded joints while preserving accuracy for visible joints.

1. Introduction

Pose and shape estimation is a core component of augmented and virtual reality applications. The majority of monocular 3D pose and shape estimation approaches [28, 31, 7, 3, 15] are designed to predict only a single solution, yet 3D recovery from a monocular input is an inverse problem. Multiple solutions are feasible, especially under settings with occlusions, truncations, low image quality, or other ambiguities. It is therefore meaningful and desirable to make multi-hypothesis predictions. Multiple hypotheses are also useful in downstream tasks such as 2D keypoint fitting [21] and multi-view fusion [21, 25].

Previous multi-hypothesis works use various 2D [2, 21, 37], 3D [39, 25, 2, 21, 37, 35], and mesh [2] reconstruction losses to facilitate kinematically feasible poses and shapes, while encouraging diversity in the solution set. Yet the learning and evaluation for multi-hypothesis works are underdeveloped. Learning-wise, many existing works are not label-friendly and often require one-to-many labeled data to achieve diversity [21, 2, 35, 25], *i.e.*, similar ambiguous observations with multiple distinct ground-truth poses. Such labels are challenging to obtain, especially under occlusion or out-of-view scenarios for which multi-hypothesis predictions are the most meaningful. While empirical efforts have been made to avoid mode collapse while encouraging diversity [26, 27], these methods often do not fully explore the solution space. It is, therefore, non-trivial to achieve feasible and as diverse as possible solutions.

Evaluation-wise, multi-hypothesis works use the best hypothesis (BH) as the metric of choice [25, 2, 21, 37, 26]. BH measures the closest distance between a hypothesis set and the ground truth. It emphasizes the accuracy of the closest hypothesis while ignoring the set holistically. Few works explicitly and quantitatively evaluate the diversity of the predicted hypotheses. When they do so, diversity is evaluated independently from the input [26]. Yet joints under occlusion or a lack of evidence feature more uncertainty, and as such should correspond to more diversity, while joints should be less diverse on unambiguous parts. Current multi-hypothesis evaluation schemes do not make such distinctions. Unwanted diversity on unambiguous joints also contributes to the calculation of overall diversity.

[*] Equal contribution.

This paper addresses these shortcomings and presents a new multi-hypothesis framework for estimating 3D poses and shapes. At the heart of our method is a simple yet essential underlying criterion - hypotheses should be diverse, but meaningfully so, and correlate with an observation's ambiguity (see Fig. 1 (b)). To that end, we take a probabilistically principled approach and minimize the Kullback-Leibler (KL) divergence between the model distribution and the underlying data distribution. Our objective results in three terms: a reconstruction accuracy, a pose prior, and most interestingly, a model entropy term. This additional entropy term echoes the principle of maximum entropy [10], where the probability distribution is designed to align with observations but otherwise be as unbiased as possible. For pose estimation, this means that visible and unambiguous keypoints should remain fixed across the hypotheses, while occluded and ambiguous keypoints should be feasible yet diverse (see Fig. 1 (b) and (c)).

Unlike existing works [35, 25, 21], our formulation for diversity does not require one-to-many training data. Instead, our model is trained by explicitly encouraging it to explore the feasible solution space via entropy maximization while remaining consistent with the evidence. For the implementation of our framework, we learn a conditional distribution of parameters with a normalizing flow model. The parameters modelled come from parametric 3D models like MANO [33] and SMPL [30].

To be label-friendly, we advocate using weak labels; specifically, we recommend 2D keypoints from only *visible* joints. Using 2D keypoint labels is well established in the literature [3, 22]. However, considering only visible 2D keypoints is not well-studied, even though it is highly intuitive from an annotation point of view and a natural fit for multiple hypotheses.

For a comprehensive evaluation beyond the best hypothesis we introduce a Per-Joint Diversity (PJD) to measure the diversity of visible and occluded keypoints. Based on PJD, we further propose a Relative Diversity (RD) ratio to measure the reasonable diversity of the hypotheses. A low ratio indicates that observed keypoints are deterministic while occluded keypoints are diverse. In experiments, our method finds highly accurate BH and achieves the best RD ratio compared to other multi-hypothesis approaches. To summarize our contributions,

- We formulate multi-hypothesis estimation as a minimization of the KL divergence between the model distribution and the data distribution. This formulation is probabilistically principled and naturally yields an entropy term that encourages diversity in the solution set.
- From our KL formulation, we propose a framework to estimate multiple hypotheses of 3D pose and shape that favours feasible and diverse solutions by design.
- We emphasize visibility in multi-hypothesis frameworks by exploring visible 2D keypoints, *i.e.* partial weak labels for estimating 3D pose and shape and factoring in visibility for evaluating diversity.
- Experiments on toy, ambiguous, and real-world data demonstrate that our method achieves excellent diversity and the best BH compared to other state-of-the-art multi-hypothesis methods.

2. Related Work

2.1. 3D Human/Hand Pose & Shape Recovery

Existing works estimate 3D pose and shape either directly [3, 16, 20, 19], or indirectly [5, 28, 31] via a parametric model like SMPL and MANO. Parametric models serve as priors to encourage feasible solutions [3, 16] and reduce the reliance on labels [22, 41]. The parameters of parametric models are typically learned under the supervision of 3D meshes and poses [3, 16, 20, 19]. Without parametric models, works like [42, 5, 28, 31, 22] focus on the representations of 3D surfaces and the design of architectures. They convert the surfaces into different 3D representations like mesh vertices [31, 28], UV mapping [42, 5, 40], and implicit representations [6], and then fit the surfaces based on architectures like GCNs [22] and Transformers [28].

2.2. Multi-Hypothesis Methods

These methods predict diverse and feasible predictions from ambiguous input evidence, often via deep generative models. The work in [35] proposes a conditional VAE to model the distribution of a 3D pose sample set that is consistent with the 2D pose, which helps to tackle the inherent ambiguity in 2D-to-3D lifting. MDN-based works [39, 25] introduce mixture density models to estimate multiple hypotheses by minimizing the negative log-likelihood of a multi-modal mixture of Gaussians.

More recent works [37, 2, 21] apply normalizing flow (NF) models. The work in [2] directly employs NFs as a prior on the distribution of plausible poses at test time. Differently, other works, like [37, 21], propose using a conditional NF. Specifically, [21] uses the conditional NF to model the distribution of SMPL pose parameters conditioned on the 2D image, while [37] adopts predicted 2D keypoints as a condition for the NF and specifically considers 2D poses in the latent space to model the distribution from 3D poses to 2D poses and its reverse.

Despite their use of probabilistic modeling, few existing works use explicit distributional optimization objectives. Our proposed method is derived directly from KL divergence, which also brings an entropy term and explicitly encourages the estimation of multiple hypotheses.

3. Preliminaries

3.1. Overview of Parametric Models

MANO [33] and SMPL [30] are commonly-used parametric 3D models for human hands and bodies with pose parameters $\theta \in \mathbb{R}^{N_\theta}$ and shape parameters $\beta \in \mathbb{R}^{N_\beta}$. Usually, θ and β are expressed as axis-angle rotations and PCA coefficients learned from pose data and registered shapes, respectively, though θ can also be expressed as PCA coefficients for MANO. Together, θ and β fully determine the surface mesh $\mathcal{M}(\theta, \beta) \in \mathbb{R}^{N_m \times 3}$ and joint coordinates $\mathcal{J}(\theta, \beta) \in \mathbb{R}^{N_j \times 3}$ in the 3D space.

Given camera parameters $\mathbf{c} = \{R, \mathbf{t}, s\}$, where $R \in \mathbb{R}^{3\times 3}$ is a global rotation matrix, $\mathbf{t} \in \mathbb{R}^2$ is the translation, and s is a scaling factor, the 3D pose $\mathcal{J}(\theta, \beta)$ can be projected into 2D joints $\mathbf{j}$ with an orthographic projection Π:

$$\mathbf{j} = s \cdot \Pi(R \cdot \mathcal{J}(\theta, \beta)) + \mathbf{t}. \tag{1}$$

3.2. 2D Keypoint Supervision

One weakly-supervised variant of monocular 3D pose and shape estimation is learned from only 2D keypoint annotations. A common approach [3] is to estimate the MANO or SMPL parameters $(\hat{\theta}, \hat{\beta})$ for a given image and project the resulting 3D pose back into 2D joints $\hat{\mathbf{j}}$, as per Eq. (1). The parameters can be learned with ground-truth 2D joints $\mathbf{j}$ by minimizing the following objective:

$$\mathcal{L} = ||\mathbf{j} - \hat{\mathbf{j}}||_1 + \lambda_\theta \mathcal{R}(\hat{\theta}) + \lambda_\beta ||\hat{\beta}||_2^2, \tag{2}$$

featuring a 2D reconstruction loss, a prior term $\mathcal{R}(\cdot)$ on θ to encourage feasible poses, an l_2 regularization on β, and weighting hyperparameters λ_θ and λ_β. The pose prior $\mathcal{R}(\cdot)$ could be adversarial prior for rotation representations [16] or an l_2 regularization for PCA coefficients [33].

3.3. Normalizing Flow

Normalizing Flows [32] are generative models with strong modeling capacity for complex, multi-modal distributions. Let $\mathbf{X}$ denote a d-dimensional random variable under distribution $P(\mathbf{X})$. The normalizing flow model represents $\mathbf{X}$ as a series of invertible mappings $\{f_l\}_{l=1}^L : \mathbb{R}^d \mapsto \mathbb{R}^d$ on d-dimensional random variable $\mathbf{Z}$:

$$\mathbf{X} = \mathcal{F}(\mathbf{Z}) = f_L \circ ... \circ f_2 \circ f_1(\mathbf{Z}). \tag{3}$$

Typically, the base distribution $P(\mathbf{Z})$ is simple, *e.g.*, a normal distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$. By some specially designed structures of flow blocks [8] and the change-of-variable rule [8], we can get the log-probability density of $\mathbf{X}$ as:

$$\log P(\mathbf{X}) = \log P(\mathbf{Z}) - \sum_{l=1}^{L} \log \left| \det \frac{\partial f_l}{\partial \mathbf{Z}_{l-1}} \right|, \tag{4}$$

where, $\mathbf{Z}_l = f_l(\mathbf{Z}_{l-1})$, $\mathbf{Z}_0 = \mathbf{Z}$ and $\mathbf{Z}_L = \mathbf{X}$. Normalizing flows estimate the likelihood with the reverse flow $\mathcal{F}^{-1}(\mathbf{X})$ transforming $\mathbf{X}$ to $\mathbf{Z}$. For sampling, it first samples $\mathbf{z}$ from $P(\mathbf{Z})$, and passes $\mathbf{z}$ through the flow $\mathcal{F}$ to get $\mathbf{x}$. Normalizing flows are favoured as generative models because they can tractably estimate the exact likelihood and be optimized through Maximum Likelihood Estimation (MLE). Furthermore, they can also be optimized by sampling through the Law of the Unconscious Statistician (LOTUS).

3.4. Principle of Maximum Entropy

The entropy of random variable $\mathbf{X}$ taking values in $\mathcal{X}$, $H(\mathbf{X})$, quantifies the uncertainty of $\mathbf{X}$. It is defined as:

$$H(\mathbf{X}) = -\int_{\mathcal{X}} p(\mathbf{x}) \log p(\mathbf{x}) \mathbf{dx}. \tag{5}$$

Methods such as heuristic objectives [26] and mutual information [13, 23] have been proposed to estimate and optimize entropy.

Under the principle of maximum entropy [10], the probability distribution that most accurately reflects the current state of the system is the one with the highest entropy. In the context of 3D pose and shape estimation, the distribution should be compatible with complete observations, *i.e.* visible joints, but otherwise be as unbiased as possible for incomplete or ambiguous observations to maximize the entropy. This prevents unnecessary information from being assumed inadvertently. Especially, entropy maximization has garnered significant attention on efficient learning *e.g.*, self-supervised learning [1, 29] and semi-supervised learning [24]. It is used to remove inadvertent assumptions and encourage the model to explore the full set of prototypes.

4. Methodology

4.1. Framework

We target multi-hypothesis 3D pose and shape recovery from RGB inputs based on visible 2D keypoints. Consider training instances $\{I, \mathbf{j}, \mathbf{v}\}$, where I is the RGB image, $\mathbf{j}$ is the corresponding visible 2D keypoints, and $\mathbf{v}$ is an indicator variable for 2D keypoint visibility. In line with previous works [22, 25, 26], we treat the shape parameter β and camera parameters $\mathbf{c}$ deterministically and assume that they can be estimated reasonably from I.

Our main interest then is to model the distribution of the pose parameter θ, conditioned on the input image I with associated $\mathbf{j}$, β and $\mathbf{c}$, *i.e.*, the conditional distribution $p(\theta|I, \mathbf{j}, \mathbf{c}, \beta)$, which we refer to as the data distribution. To model the data distribution, we learn a model ϕ in the form of a neural network. Similarly, the model ϕ has the distribution $p_\phi(\theta|I, \mathbf{j}, \mathbf{c}, \beta)$, which we term as the model distribution. The model ϕ can be learned by minimizing the

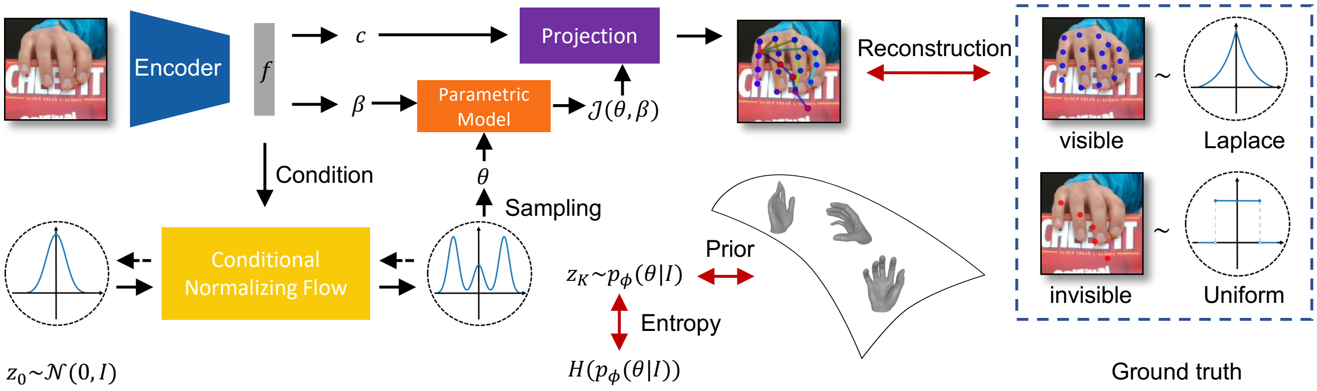


Figure 2. Framework overview. The framework is optimized by considering three components: reconstruction, prior, and entropy (see the red double-headed arrows). The distribution of the feasible pose parameters is captured by an NF model conditioned on image features. The reconstruction applies different distributions for the visible keypoints (blue dots) and occluded keypoints (red dots).

Kullback-Leibler (KL) divergence between the data and the model distribution, *i.e.*,

$$KL(p_\phi(\theta|I,\mathbf{j},\mathbf{c},\beta)\|p(\theta|I,\mathbf{j},\mathbf{c},\beta)). \tag{6}$$

Data Distribution. Inspired by the existing 2D-to-3D lifting works [7, 25, 26] that a 3D pose could be accurately estimated by its corresponding 2D pose, β and camera information, we assume that once $\{\mathbf{j},\mathbf{c},\beta\}$ are given, θ and I are conditionally independent. As such, I can be omitted as a conditioning variable. With Bayes' rule, the data distribution can be split as:

$$p(\theta|I,\mathbf{j},\mathbf{c},\beta) = p(\theta|\mathbf{j},\mathbf{c},\beta) \propto p(\mathbf{j}|\mathbf{c},\beta,\theta)\cdot p(\theta). \tag{7}$$

The first decomposed term in Eq. (7), the likelihood $p(\mathbf{j}|\mathbf{c},\beta,\theta)$, is a projection consistency term that reflects the reconstruction accuracy. The second term, $p(\theta)$, serves as a general pose prior in a probabilistic perspective [16, 21].

Model Distribution. Like previous works [22, 16, 21, 2], we estimate θ from an image I, as the image contains sufficient information to infer $\mathbf{c}$, β and the keypoints $\mathbf{j}$. The model distribution can be simplified as:

$$p_\phi(\theta|I,\mathbf{j},\mathbf{c},\beta) = p_\phi(\theta|I). \tag{8}$$

Based on Eqs. (7) and (8), the KL divergence between the model and data distributions can be expressed as:

$$KL(p_\phi(\theta|I,\mathbf{j},\mathbf{c},\beta)\|p(\theta|I,\mathbf{j},\mathbf{c},\beta)) = \\ -\Bigg(\underbrace{\underset{p_\phi(\theta|I)}{E}[\log p(\mathbf{j}|\mathbf{c},\beta,\theta)]}_{\text{reconstruction}} + \underbrace{\underset{p_\phi(\theta|I)}{E}[\log p(\theta)]}_{\text{prior}} + \underbrace{H(p_\phi(\theta|I))}_{\text{entropy}}\Bigg), \tag{9}$$

where $H(p_\phi(\theta|I)) = -E_{p_\phi(\theta|I)}[\log p_\phi(\theta|I)]$ is the entropy of θ given input image I. See Suppl. B for the full derivation. Minimizing the KL divergence in Eq. (9) maximizes the reconstruction accuracy and the conditional entropy of the pose under pose prior $p(\theta)$; this can be used directly to supervise the neural network ϕ.

Based on the above derivation, we propose a weakly-supervised multi-hypothesis framework as illustrated in Fig. 2. We consider only *visible* 2D keypoints as supervisory signals for the reconstruction. Like [21, 26], we assume that visible keypoints follow a Laplace distribution for sharpness. In line with the principle of maximum entropy (Sec. 3.4), the occluded keypoints each follow a uniform distribution of feasible locations. The intuition behind such an assumption is that, for occluded keypoints, we relax the supervision using prior knowledge $p(\theta)$ to generate feasible solutions for the image. This prevents an overconfident model that tries to fit all labels regardless of visibility and also compensates for the lack of one-to-many data-label pairs. As θ is derived from a parametric model, the prior can be applied simply as a uniform distribution [33] or adversarially [16] based on its representation. We choose to model the distribution $p_\phi(\theta|I)$ using a conditional NF model, *i.e.*, $p_\phi(\theta|I) = \mathcal{F}^{-1}(\theta|I)$, as it is more feasible to calculate the entropy term $H(p_\phi(\theta|I))$ via Monte Carlo (MC) sampling. Therefore, all three terms in Eq. (9) can be maximized by MC sampling and SGD [18].

An important part of our formulation involves the explicit maximization of entropy. The link between entropy and diverse hypotheses is highly intuitive, yet this has been overlooked in previous work. Even without one-to-many labels, the entropy term encourages the model ϕ to generate hypotheses that are diverse; the reconstruction and prior term ensure that the hypotheses respect the observed labels while remaining feasible.

4.2. Implementation Details

In Eq. (9), we represent image I with image features extracted from a ResNet-50 [12] backbone. To estimate $\mathbf{c}$ and β, we append a 512-hidden unit MLP to the backbone. For the normalizing flow model, we use the Real NVP[8].

First, to obtain θ for an image, according to LOTUS, we sample $\mathbf{z}_0$ from a Gaussian distribution and feed it together with image feature of I to the invertible flow network $\mathcal{F}(\mathbf{z}|I)$ (the solid line in the yellow block in Fig. 2).

Reconstruction. We use a constant scale b for the Laplace and assume the joints do not exceed the large occlusion range (Suppl. B.2); the reconstruction loss simplifies to:

$$\mathcal{L}_{\text{rec}} = \sum_{k=1}^{K} v_k ||\mathbf{j}_k - \hat{\mathbf{j}}_k||_1, \tag{10}$$

where K is the number of joints and $\mathbf{j}_k$ represents k-th joint; the loss is effective only on visible joints based on visibility indicator v_k.

Prior. To encourage feasible poses, we introduce a prior term $\mathcal{R}(\cdot)$ on θ. For MANO, θ is given as PCA coefficients. We empirically place a uniform distribution $\mathcal{R}(\theta) = \mathcal{U}(-2, 2)$ on these coefficients, which covers most feasible solutions while avoiding invalid poses [33], *i.e.*,

$$\mathcal{L}_\theta = \sum_i \max(0, |\theta_i| - 2)^2. \tag{11}$$

For SMPL, θ represents axis-angle rotations and can be restricted by an adversarial prior [16], *i.e.*,

$$\mathcal{L}_\theta = \text{Adv}(\theta). \tag{12}$$

Entropy. We use a negative log-likelihood loss:

$$\mathcal{L}_H = -\log p_\phi(\theta|I), \tag{13}$$

where θ is sampled from the normalizing flow. The reverse path of the NF, $\mathcal{F}^{-1}(\theta|I)$, maps θ back to $\mathbf{z}_0$ in the latent space conditional on the image features to compute Eq. (4) [8, 17] (dashed lines in the yellow block in Fig. 2).

The losses in Eqs. (10)-(13) each covers a term in Eq. (9). With the β regularization in Eq. (2), all losses sum into the final training objective:

$$\mathcal{L} = \lambda_{\text{rec}}\mathcal{L}_{\text{rec}} + \lambda_\theta\mathcal{L}_\theta + \lambda_H\mathcal{L}_H + \lambda_\beta\mathcal{L}_\beta, \tag{14}$$

where the λs are the trade-off hyperparameters and $\mathcal{L}_\beta = ||\beta||_2^2$. See Suppl. B for the full derivations and more training details.

5. Experiments

5.1. Datasets, Metrics, & Baselines

Datasets. We synthesize a 2-joint toy setting in 2D to highlight components of our model. For the human body, we experiment on Human3.6M (**H36M**) [14] and its ambiguous version **AH36M** [2] with randomly truncated images of H36M to hide keypoints. Following [2, 21], we train with subjects S1 and S5-9 and test with S11, training with H36M and AH36M jointly, while evaluating separately.

Inspired by AH36M, we construct Ambiguous RHD (**ARHD**) from the synthetic hand pose dataset RHD [43] by adding circular patches with a predefined radius to the fingers' DIP joints[1]. The visibility in the scene is affected depending on the finger and circle radius (see Fig. 5(c)). We also use **HO3D** [11], a real-world hand-object dataset that features severe occlusions. To evaluate the multi-hypothesis metrics, similar to previous work [38], we split a test subset from the training dataset. We estimate the visibility of a joint [9] by thresholding the difference between captured surface depth and the true keypoint position. More details of the datasets are provided in Suppl. D.

Evaluation Metrics. Mean End-Point Error (EPE) is the average Euclidean distance between predicted and ground-truth joints, from which we consider the Best Hypothesis (BH) [25, 2] and our newly proposed All Hypothesis (AH).

BH is a standard multi-hypothesis metric that selects the hypothesis with the lowest pose or mesh EPE. To evaluate the accuracy of *all* hypotheses, we propose **AH**, which is the mean EPE of *all* hypotheses on 2D visible joints to measure consistency to the image evidence.

As we highlighted, multiple hypotheses should be diverse, but the diversity should only be on uncertain joints *e.g.* under occlusion. However, existing BH and diversity metrics [34, 26] do not capture this target because undesirable diversity on the visible joints may also contribute to the diversity metrics. Therefore, we propose to complement the evaluation of multi-hypothesis methods with Per-Joint Diversity (PJD) and a Relative Diversity (RD) ratio.

PJD measures the standard deviation per joint and can be used to show the diversity of both visible and occluded joints in 2D and 3D spaces. To highlight the source of diversity, we propose a ratio:

$$\textbf{RD} = \frac{\text{PJD}_{\text{2d vis}}}{\text{PJD}_{\text{3d occ}}}, \tag{15}$$

to account for the diversity of both the certain (*i.e.*, 2D visible keypoints) and the uncertain parts (*i.e.*, 3D occluded keypoints). A lower RD means more diversity on the occluded keypoints relative to visible keypoints. We follow [35, 37] and sample 200 hypotheses for evaluation.

[1] The distal interphalangeal (DIP) joint is the one closest to the fingertip.

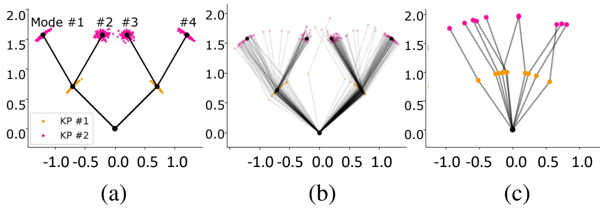


Figure 3. **(a)** Toy problem setting featuring four modes. **(b)** *Our* probabilistic method in a weakly supervised setting recovers all the modes. **(c)** The *deterministic* method, even under strong supervision, predicts wrong modes as it easily overfits ubiquitous observation noise.

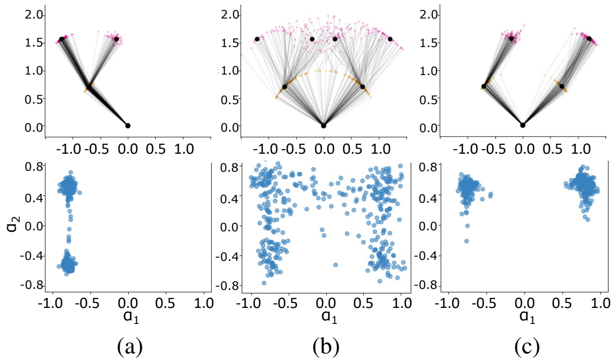


Figure 4. **(a)** Lowering the weight of the entropy term can impair multi-mode learning while **(b)** increasing the weights hurts the fitting accuracy, although it achieves diversity. **(c)** Choosing different feasibility priors will also make the final learned modes different. The second row shows the mode distribution in α space.

Baselines & SOTAs. We compare our method with two deterministic methods, Det (2D Vis) and Det (3D), which use visible 2D and all 3D keypoint positions as supervision, respectively. For more details, refer to a similar pipeline in [3]. Moreover, we compare our method with the state-of-the-art multi-hypothesis methods, including MDN [25], conditional VAE [35], Multi-bodies [2], ProHMR [21], CM-VAE [36] and WS3DPG [26].

Some of these approaches are designed with 3D pose supervision; we replace their corresponding supervised losses with the 2D version in Eq. (2). Unless specified, we also use visibility labels as weights while remaining as faithful as possible to the original method. See Suppl. F for method introductions and details.

5.2. Toy Experiments & Ablation Studies

Settings. We perform the toy experiment under a simple setting (depth ambiguity), as shown in Fig. 3(a). Consider a single chain with two keypoints plus a root keypoint. The bones are fixed to length 1 and the root is at the origin. The data $(\mathbf{y}, \alpha)$ consists of the 1D projection y_k of the keypoint on the y-axis and the angle α_k between the chain and the y-axis. From α_k, we can get 2D coordinates $\mathbf{j}_k = (\sin\alpha_k, \cos\alpha_k)$ relative to their parent. There are in total four Gaussian modes for the complete data, *i.e.*, each joint can swing left and right. The model is trained to predict α based on $\mathbf{y}$. For weak supervision, only 1D projections $\mathbf{y}$ are given. For strong supervision, all 2D coordinates $\mathbf{j}$ are provided. Our model uses an MLP and Real NVP as the backbone and optimizes an objective similar to Eq. (14). For the prior loss, we add an L_2 norm constraint on α.

Deterministic vs. Multi-Hypothesis. The deterministic model trained under both weak and strong supervision can learn only one of the four modes. Moreover, under strong supervision, it is sensitive to similar input data and predicts wrong modes if the inputs are corrupted with small perturbations (Fig. 3(c)). In contrast, existing multi-hypothesis methods can recover all modes under strong supervision.

Ours vs. Existing Multi-Hypothesis Methods. Existing methods require similar observations with multiple distinct ground-truth poses. In the weakly supervised setting, we compare with MDN and observe that it finds only one of the modes (Suppl. G.1) while our proposed method can successfully recover all the modes (Fig. 3(b)).

Reconstruction vs. Entropy. The entropy term encourages the set of predictions to cover diverse solutions while maintaining a low reconstruction error. As the weight of the entropy term λ_H decreases, the objective emphasizes reconstruction at the cost of entropy, leading to missed modes (Fig. 4(a)). The extreme is the degradation to a deterministic model. On the other hand, with the increase in λ_H, the model pays less attention to reconstructing observed evidence, hence the modes become dispersed (Fig. 4(b)).

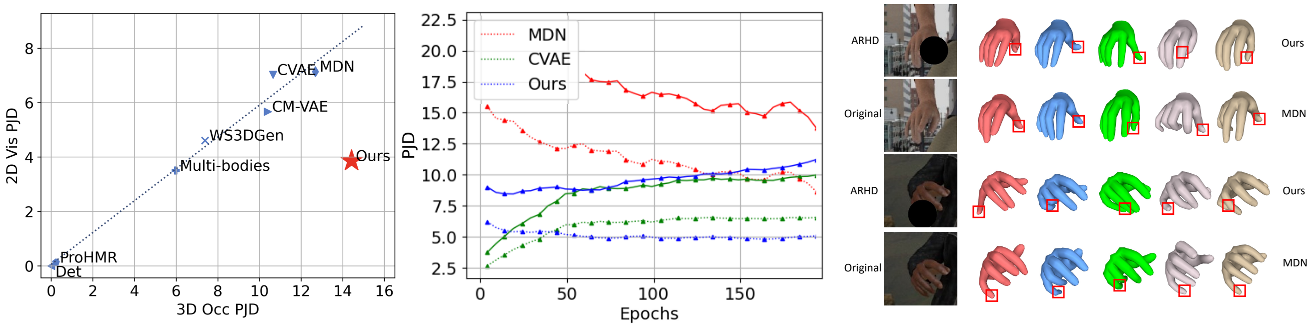
Angle Prior. The prior term determines the distribution over the feasible solution space. When we lower the weight of the prior loss, the model may fit the evidence better but consider less feasible poses. On the other hand, prior knowledge defines the solution space, and the entropy term will encourage the predictions to cover all possible solutions based on the prior. For example, when we add a prior $\sin\alpha_k \geq 0$ for the top keypoints, some previous modes become infeasible and the model will only find two of the four original poses (Fig. 4(c)).

5.3. Synthetic Ambiguous RHD

Deterministic vs. Multi-Hypothesis. Table 1(a) shows that on ARHD, all the multi-hypothesis methods [25, 35, 26, 21, 37] outperform the deterministic 'Det (2D Vis)' on the BH metric. Some methods even outperform the 'Det (3D)' baseline with 3D supervision.

	ARHD					HO3D					
	BH (mm)↓	AH (pix)↓	PJD		RD↓	BH (mm)↓		AH (pix)↓	PJD		RD↓
	Joint		2D Vis	3D Occ		Joint	Vert		2D Vis	3D Occ	
Det (3D) [44]	20.98	-	-	-	-	23.88	25.18	-	-	-	-
Det (2D Vis)	25.11	14.39	-	-	-	24.10	25.40	16.85	-	-	-
Multi-bodies [2]	<u>20.52</u>	15.76	3.50	5.98	0.59	22.07	23.56	19.57	1.92	3.09	0.62
MDN [25]	21.33	18.47	7.14	12.69	0.56	21.28	22.67	18.81	3.48	6.25	<u>0.56</u>
CVAE [35]	20.99	19.95	7.02	10.66	0.66	<u>21.04</u>	<u>22.62</u>	18.90	4.07	6.64	0.61
ProHMR [21]	24.44	**13.37**	0.13	0.22	0.59	24.05	25.41	<u>17.19</u>	0.16	0.25	0.64
CM-VAE [36]	21.99	16.59	5.67	10.40	<u>0.55</u>	22.20	23.54	18.39	4.94	8.64	0.57
WS3DPG [26]	24.34	17.79	4.61	7.39	0.62	23.67	24.99	18.34	3.06	4.73	0.65
Ours	**20.35**	<u>13.42</u>	3.86	14.42	**0.27**	**20.55**	**21.83**	**16.95**	3.30	11.93	**0.28**

Table 1. Quantitative results on ARHD and HO3D V3. The best is marked in **bold**; the second best is <u>underlined</u>. Our method achieves competitive SOTA results; occluded keypoints are diverse while visible ones are relatively accurate, *i.e*., have a lower RD.



(a) RD: 2D Vis PJD - 3D Occ PJD

(b) PJD Curve Along Training

(c) Qualitative Results

Figure 5. Illustration on ARHD. (a) Deviation from the diagonal (dashed line) towards the lower-right indicates a better trade-off between the 2D visible accuracy and 3D occluded diversity. (b) Comparison of PJD during learning for occluded (solid line) and visible (dashed line) keypoints. Our approach has a descending trend for visible keypoints and an ascending trend for the occluded ones. (c) Multi-hypothesis meshes. The red boxes highlight the concerned occluded keypoints. Our method predicts diverse and feasible poses under occlusion while MDN's predictions are inconsistent with image evidence on visible keypoints.

Comparisons with SOTA. Methods like MDN [25] and CVAE [35] all have low BH (Tab. 1(a)). However, their diversity (PJD) for visible and occluded keypoints is high, *i.e*., the entire hand is diverse. Together with a large AH, this suggests that their hypotheses as a whole do not fit the image evidence well. ProHMR [21] degenerates to an almost deterministic method, and has the lowest PJD for visible and occluded keypoints. We speculate that it is because of its reliance on the strong supervision for θ labels. More comparisons can be found in Sec. 5.5.

In contrast, our framework obtains lower BH and AH and higher occluded diversity (higher PJD_{occ} and thus significantly lower RD). Fig. 5(a) shows that we strike a good balance between reconstruction and diversity; qualitatively, Fig. 5(c) shows that our recovered meshes are diverse and consistent with the observation.

Connection to Existing Works & Ablation Studies. ProHMR [21], CM-VAE [36], and WS3DPG [26] are closely related to our method. Among them, ProHMR is a variant without entropy optimization; CM-VAE [36] applies a single Gaussian distribution instead of our NF; WS3DPG [26] uses a GAN whose entropy is intractable. Tab. 1(a) shows that our entropy term increases the PJD significantly, especially for occluded joints in 3D (ProHMR's 0.22 vs. our 14.42). Moreover, we outperform CM-VAE and WS3DPG on all metrics on ARHD, showing the powerful ability of NFs to model complex distributions and estimate entropy. Additional ablations on visibility in Suppl. G.2 shows that adding additional supervision on occluded joints hurts diversity.

PJD During Training. Fig. 5(b) shows the test PJD scores throughout training. The PJD of MDN [25] for both visible and occluded keypoints decreases as training progresses. Meanwhile, CVAE [35] has the opposite trend, with increasing PJD scores for both types of keypoints. In contrast, our approach has a descending trend on the PJD of visible keypoints and an ascending trend on the occluded ones (Tab. 1(a)).

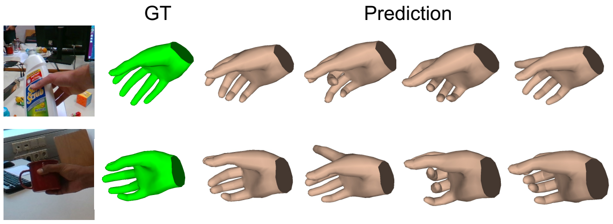


Figure 6. Visualization of our multiple hypotheses on HO3D.

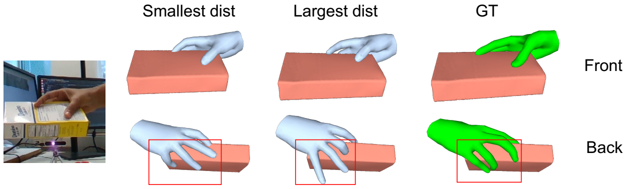


Figure 7. Visualization of hypotheses with the smallest and largest HOI Chamfer distance in two views. Differences are highlighted with red boxes. Hypothesis selection leads to more feasible grasps.

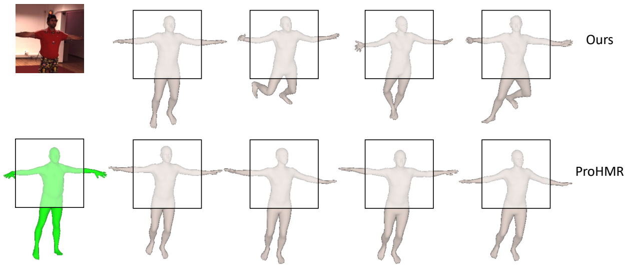


Figure 8. Qualitative results on AH36M. The legs of our hypotheses are much more diverse, while all trunks are consistent with the image. The green mesh is the ground-truth.

5.4. Real-World Data: HO3D

When faced with real-world ambiguous settings, such as objects occluding hands, our method is state-of-the-art compared to other multi-hypothesis methods. The results in Tab. 1(b) and Fig. 6 are consistent with the trends on the synthetic ARHD – our method has the lowest BH and RD.

Hypothesis Selection. With additional input information, the multiple hypotheses can be filtered for an improved set of solutions. We show an example of hypothesis selection in which we select feasible grasps based on the object interaction. We simply use a post-processing strategy to select samples. Specifically, we use a hand-object interaction (HOI) feasibility metric, *e.g.*, the widely used Chamfer distance [4]. Based on the value of the Chamfer distance, we can select hypotheses. We found it useful for picking more plausible poses from hypotheses using task-related constraints. We visualize hypotheses with the lowest and highest HOI Chamfer distance in Fig. 7. Our hypotheses are all consistent with the image cues in the front camera view. From a different view, by incorporating the HOI constraint for hypotheses, we can select more feasible grasp poses.

Meaningful Diversity. The purpose of our method is to encourage meaningful diversity; hence we treat visible and occluded joints differently. The quality of diversity can be further improved by incorporating more information about the observed ambiguity. One example is to add mask information about the occlusions; it can be incorporated into our framework as a post-hoc hypothesis selection during inference or as a reconstruction loss in Eq. (9) during training. In both ways, the error rate for out-of-occlusion can therefore be reduced without much loss in diversity (Suppl. G.2).

5.5. State-of-The-Art on H36M & AH36M

Supervision		MH	H36M	AH36M
2D Vis	HMR		67.4	85.2
	ProHMR	✓	64.3	82.6
	Ours	✓	**51.3**	**66.4**
3D	HMR		56.8	-
	SPIN		41.1	-
	MDN	✓	42.7	69.5
	CVAE	✓	46.2	75.1
	Multi-bodies	✓	42.2	64.2
	ProHMR	✓	**36.8**	<u>60.1</u>
	Ours	✓	**36.8**	**50.6**

Table 2. PA-MPJPE (mm) of BH results on H36M and its ambiguous version AH36M under the supervision of visible 2D keypoints (2D Vis) and 3D keypoints (3D) with $n = 25$.

Supervision		AH (pix)↓	PJD		RD↓
			2D Vis	3D Occ	
2D Vis	ProHMR	10.92	0.06	0.26	0.23
	Ours	**9.75**	4.56	64.05	**0.07**
3D	ProHMR	13.38	3.98	24.27	0.16
	Ours	**10.73**	4.23	47.95	**0.09**

Table 3. Diversity metrics on AH36M under the supervision of visible 2D keypoints (2D Vis) and 3D keypoints (3D).

Our framework is also effective for human pose estimation. We follow [2, 21] and evaluate the accuracy and diversity of hypotheses on the benchmark H36M and its ambiguous version AH36M. Besides using 2D visible keypoints as supervision, we also test a variant using 3D keypoints. For supervision with 3D keypoints, we follow ProHMR [21] and supervise the predicted 3D poses from SMPL directly.

Tab. 2 shows that we achieve the best BH on both datasets with different supervision settings. Our method has an impressive 9.5mm improvement on AH36M with 3D keypoints as labels; furthermore, our performance with 2D visible keypoints as weak labels is comparable to using 3D keypoints as labels. Comparing diversity metrics in Tab. 3 and Fig. 8, we outperform ProHMR with respect to AH and RD. Our method with the entropy term predicts highly diverse results in the weakly-supervised setting.

	H36M		AH36M
	Multi-View	Fitting	Fitting
ProHMR	34.5	34.8	61.4
Ours	**34.2**	**34.4**	**53.5**

Table 4. PA-MPJPE (mm) of downstream tasks multi-view refinement and fitting humans with 2D ground-truth. For these tasks, a trained model distribution outputs only one hypothesis by observing more evidence.

Downstream Tasks. Our method, by estimating an accurate yet diverse set of hypotheses, excels at providing inputs for downstream tasks. We verify the hypotheses through multi-view refinement and fitting humans with 2D ground-truth on H36M and AH36M. We follow ProHMR to find the solution that best matches the evidence among many possible hypotheses by optimization, *i.e.*,

$$\max_{\theta} \log p_{\phi}(\theta|I) + c(\theta|\mathbf{e}), \tag{16}$$

where $c(\theta|\mathbf{e})$ is consistency with additional information, *e.g.*, multi-view and 2D projection consistency. Lower PA-MPJPE in Tab. 4 verifies our effectiveness, which may imply ours obtains a better representation with a more accurate distribution than ProHMR.

6. Discussion & Conclusion

Our proposed multi-hypothesis framework is flexible, efficient, and label-friendly. We emphasize that diversity in the hypotheses should not be arbitrary; instead, it should come from ambiguity present in the image itself. To that end, we also propose a more comprehensive evaluation scheme based on visibility. In our work, we have only considered ambiguity from object occlusions and image truncations, but additional factors such as image quality and light are interesting directions to investigate for future work.

We note that by virtue of evaluating on the provided annotations of the current datasets, the diversity concept remains vague and limited to the dataset. For instance, in-distribution testing aims to learn the pose diversity captured in that dataset. It may not require learning more diverse outdoor poses to achieve a good BH on indoor datasets and thus not conducive to generalization to other scenarios. It is suggested to focus more on the improvement and evaluation of the diversity in the generalization scenario.

Acknowledgments. This research/project is supported by the Ministry of Education, Singapore, under its MOE Academic Research Fund Tier 2 (STEM RIE2025 MOE-T2EP20220-0015). We would also like to thank the ACs, reviewers, Dr. Chen Li, and Junpeng Hu for their valuable suggestions.

References

[1] Mahmoud Assran, Mathilde Caron, Ishan Misra, Piotr Bojanowski, Florian Bordes, Pascal Vincent, Armand Joulin, Mike Rabbat, and Nicolas Ballas. Masked siamese networks for label-efficient learning. In *ECCV*, 2022. 3

[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. 1, 2, 4, 5, 6, 7, 8

[3] Adnane Boukhayma, Rodrigo de Bem, and Philip HS Torr. 3D hand shape and pose from images in the wild. In *CVPR*, 2019. 1, 2, 3, 6

[4] Zhe Cao, Ilija Radosavovic, Angjoo Kanazawa, and Jitendra Malik. Reconstructing hand-object interactions in the wild. In *ICCV*, 2021. 8

[5] Ping Chen, Yujin Chen, Dong Yang, Fangyin Wu, Qin Li, Qingpei Xia, and Yong Tan. I2UV-HandNet: Image-to-UV prediction network for accurate and high-fidelity 3D hand mesh modeling. In *ICCV*, 2021. 2

[6] Zerui Chen, Yana Hasson, Cordelia Schmid, and Ivan Laptev. AlignSDF: Pose-aligned signed distance fields for hand-object reconstruction. In *ECCV*, 2022. 2

[7] Hongsuk Choi, Gyeongsik Moon, and Kyoung Mu Lee. Pose2Mesh: Graph convolutional network for 3D human pose and mesh recovery from a 2D human pose. In *ECCV*, 2020. 1, 4

[8] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real NVP. In *ICLR*, 2017. 3, 5

[9] Kerui Gu, Linlin Yang, and Angela Yao. Removing the bias of integral pose regression. In *ICCV*, 2021. 5

[10] Silviu Guiasu and Abe Shenitzer. The principle of maximum entropy. *Math. Intell.*, 7(1):42–48, 1985. 2, 3

[11] Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. HOnnotate: A method for 3D annotation of hand and object poses. In *CVPR*, 2020. 5

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

[13] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. In *ICLR*, 2019. 3

[14] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. *TPAMI*, 36(7):1325–1339, 2013. 5

[15] Umar Iqbal, Pavlo Molchanov, Thomas Breuel Juergen Gall, and Jan Kautz. Hand pose estimation via latent 2.5D heatmap regression. In *ECCV*, 2018. 1

[16] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *CVPR*, 2018. 2, 3, 4, 5

[17] Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In *NeurIPS*, 2018. 5

[18] Diederik P. Kingma and Max Welling. Auto-encoding variational Bayes. In *ICLR*, 2014. 4

[19] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. VIBE: Video inference for human body pose and shape estimation. In *CVPR*, 2020. 2

[20] 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

[21] Nikos Kolotouros, Georgios Pavlakos, Dinesh Jayaraman, and Kostas Daniilidis. Probabilistic modeling for human mesh recovery. In *ICCV*, 2021. 1, 2, 4, 5, 6, 7, 8

[22] Dominik Kulon, Riza Alp Guler, Iasonas Kokkinos, Michael M Bronstein, and Stefanos Zafeiriou. Weakly-supervised mesh-convolutional hand reconstruction in the wild. In *CVPR*, 2020. 2, 3, 4

[23] Rithesh Kumar, Sherjil Ozair, Anirudh Goyal, Aaron Courville, and Yoshua Bengio. Maximum entropy generators for energy-based models. In *arXiv:1901.08508*, 2019. 3

[24] Jogendra Nath Kundu, Siddharth Seth, Pradyumna YM, Varun Jampani, Anirban Chakraborty, and R Venkatesh Babu. Uncertainty-aware adaptation for self-supervised 3d human pose estimation. In *CVPR*, 2022. 3

[25] Chen Li and Gim Hee Lee. Generating multiple hypotheses for 3D human pose estimation with mixture density network. In *CVPR*, 2019. 1, 2, 3, 4, 5, 6, 7

[26] Chen Li and Gim Hee Lee. Weakly supervised generative network for multiple 3D human pose hypotheses. In *BMVC*, 2020. 1, 3, 4, 5, 6, 7

[27] Wenhao Li, Hong Liu, Hao Tang, Pichao Wang, and Luc Van Gool. MHFormer: Multi-hypothesis transformer for 3D human pose estimation. In *CVPR*, 2022. 1

[28] Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-end human pose and mesh reconstruction with transformers. In *CVPR*, 2021. 1, 2

[29] Xin Liu, Zhongdao Wang, Ya-Li Li, and Shengjin Wang. Self-supervised learning via maximum entropy coding. 2022. 3

[30] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. SMPL: A skinned multi-person linear model. *TOG*, 34(6):1–16, 2015. 2, 3

[31] 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*, 2020. 1, 2

[32] George Papamakarios, Eric T Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *JMLR*, 22(57):1–64, 2021. 3

[33] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied Hands: Modeling and capturing hands and bodies together. *TOG*, 36(6):1–17, 2017. 2, 3, 4, 5

[34] 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. 5

[35] Saurabh Sharma, Pavan Teja Varigonda, Prashast Bindal, Abhishek Sharma, and Arjun Jain. Monocular 3D human pose estimation by generation and ordinal ranking. In *ICCV*, 2019. 1, 2, 5, 6, 7

[36] Adrian Spurr, Jie Song, Seonwook Park, and Otmar Hilliges. Cross-modal deep variational hand pose estimation. In *CVPR*, 2018. 6, 7

[37] Tom Wehrbein, Marco Rudolph, Bodo Rosenhahn, and Bastian Wandt. Probabilistic monocular 3D human pose estimation with normalizing flows. In *ICCV*, 2021. 1, 2, 5, 6

[38] Lixin Yang, Xinyu Zhan, Kailin Li, Wenqiang Xu, Jiefeng Li, and Cewu Lu. CPF: Learning a contact potential field to model the hand-object interaction. In *ICCV*, 2021. 5

[39] Qi Ye and Tae-Kyun Kim. Occlusion-aware hand pose estimation using hierarchical mixture density network. In *ECCV*, 2018. 1, 2

[40] Ziwei Yu, Linlin Yang, You Xie, Ping Chen, and Angela Yao. UV-based 3D hand-object reconstruction with grasp optimization. In *BMVC*, 2022. 2

[41] Andrei Zanfir, Eduard Gabriel Bazavan, Hongyi Xu, William T Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Weakly supervised 3D human pose and shape reconstruction with normalizing flows. In *ECCV*, 2020. 2

[42] Wang Zeng, Wanli Ouyang, Ping Luo, Wentao Liu, and Xiaogang Wang. 3D human mesh regression with dense correspondence. In *CVPR*, 2020. 2

[43] Christian Zimmermann and Thomas Brox. Learning to estimate 3D hand pose from single RGB images. In *ICCV*, 2017. 5

[44] Christian Zimmermann, Duygu Ceylan, Jimei Yang, Bryan Russell, Max Argus, and Thomas Brox. FreiHAND: A dataset for markerless capture of hand pose and shape from single RGB images. In *ICCV*, 2019. 7