

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>.

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

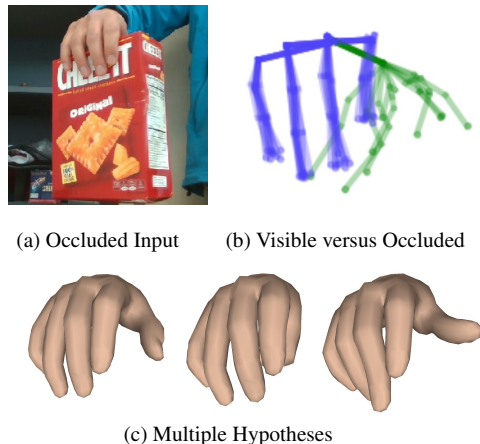


Figure 1. Our method estimates diverse and feasible hypotheses for occluded joints while preserving accuracy for visible joints.

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 :

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

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:

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

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} :

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

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:

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

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:

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

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

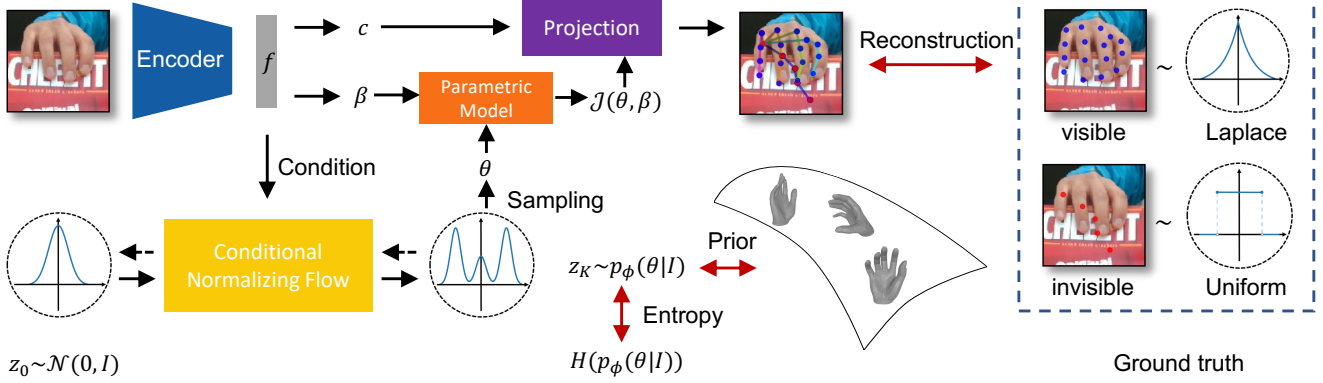


Figure 2. Framework overview. The framework is optimized by considering three components: reconstruction, prior, and entropy (see the red double-headed arrows). The distribution of the feasible pose parameters is captured by an NF model conditioned on image features. The reconstruction applies different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

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

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:

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

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:

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

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

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

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:

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

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.*,

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

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

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

Entropy. We use a negative log-likelihood loss:

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

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:

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

where the λ s 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 joints¹. 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:

$$\text{RD} = \frac{\text{PJD}_{2\text{d vis}}}{\text{PJD}_{3\text{d occ}}}, \quad (15)$$

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.

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

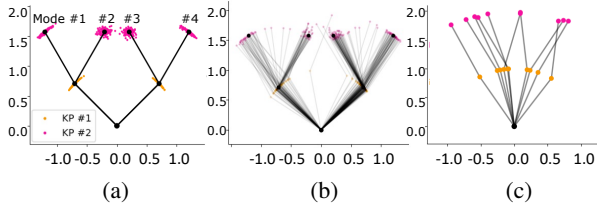


Figure 3. (a) Toy problem setting featuring four modes. (b) Our probabilistic method in a weakly supervised setting recovers all the modes. (c) The *deterministic* method, even under strong supervision, predicts wrong modes as it easily overfits ubiquitous observation noise.

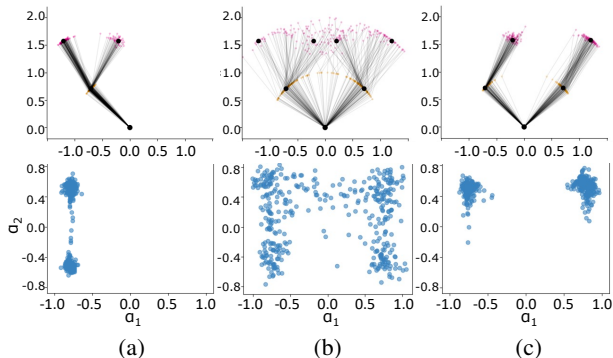


Figure 4. (a) Lowering the weight of the entropy term can impair multi-mode learning while (b) increasing the weights hurts the fitting accuracy, although it achieves diversity. (c) Choosing different feasibility priors will also make the final learned modes different. The second row shows the mode distribution 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}, α) 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 underlined. Our method achieves competitive SOTA results; occluded keypoints are diverse while visible ones are relatively accurate, *i.e.*, have a lower RD.

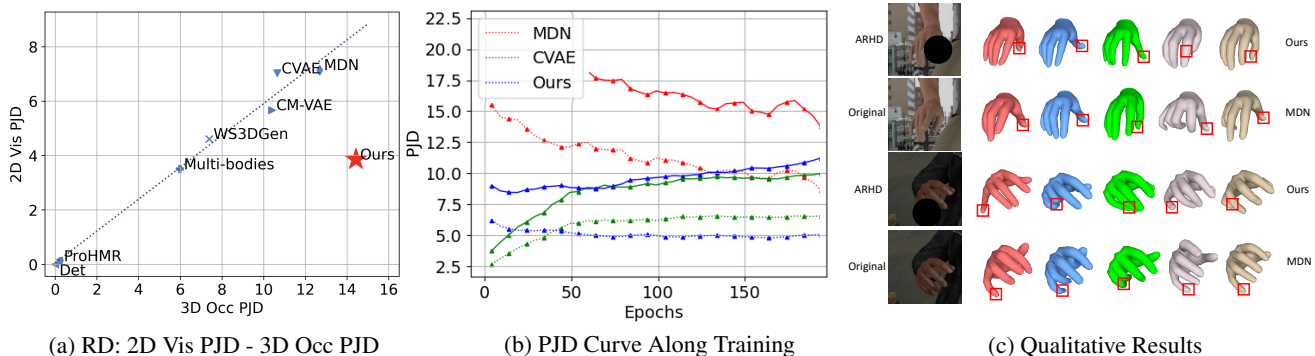


Figure 5. Illustration on ARHD. (a) Deviation from the diagonal (dashed line) towards the lower-right indicates a better trade-off between the 2D visible accuracy and 3D occluded diversity. (b) Comparison of PJD during learning for occluded (solid line) and visible (dashed line) keypoints. Our approach 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)).

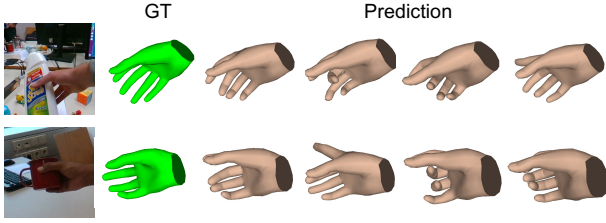


Figure 6. Visualization of our multiple hypotheses on HO3D.

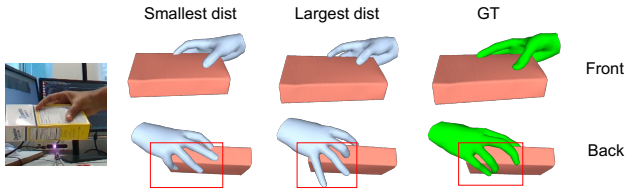


Figure 7. Visualization of hypotheses with the smallest and largest HOI Chamfer distance in two views. Differences are highlighted with red boxes. Hypothesis selection leads to more feasible grasps.

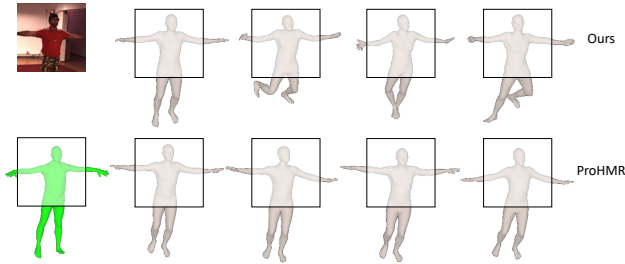


Figure 8. Qualitative results on AH36M. The legs of our hypotheses are much more diverse, while all trunks are consistent with the image. The green mesh is the ground-truth.

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	60.1
	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.*,

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

where $c(\theta|e)$ 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

MHEntropy: Entropy Meets Multiple Hypotheses for Pose and Shape Recovery

Supplementary Material

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

Contents

A Problem Formulation Illustration	1
B Objective Derivation	1
B.1. KL Divergence	1
B.2. Specific Forms	2
C Implementation Details	3
C.1. Architectures	3
C.2. Training & Hyper-Parameters	3
D Data Processing Details	3
E Evaluation Details	4
F SOTAs	4
G More Experimental Results	5
G.1 Toy	5
G.2 ARHD	5
G.3 HO3D	6

Note that all notations and abbreviations here are consistent with the main manuscript.

A. Problem Formulation Illustration

Here, we provide more insight into the formulation of the ambiguity problem in this task (Fig. a). When occlusion occurs, there are multiple joints, $\mathbf{j}^{(1)}$ and $\mathbf{j}^{(2)}$, that match image I 's evidence (1st and 2nd columns); when estimating poses from monocular 2D images, multiple poses $\theta^{(1,1)}$ and $\theta^{(1,2)}$ have similar 2D joint projections $\mathbf{j}^{(1)}$ (2nd and 3rd columns). The data itself $(I, \mathbf{j}^{(1)})$ may not have complete labels (the missing annotation is indicated by dashed lines), *i.e.*, all 2D joints $\mathbf{j}^{(1)}$ and $\mathbf{j}^{(2)}$ corresponding to the image I and their corresponding poses θ . Our objective is to use only these incomplete $(I, \mathbf{j}^{(1)})$ in the data to find all the $(I, \{\theta^{(1,1)}, \theta^{(1,2)}, \theta^{(2,1)}, \theta^{(2,2)}\})$. To this end, we use prior and weakly-supervised reconstruction conditions to define

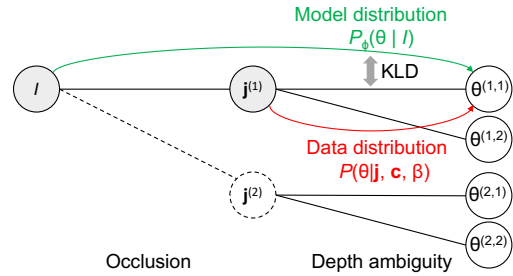


Figure a. An illustration of our problem formulation. One image I corresponds to multiple feasible 2D joints \mathbf{j} , while one joint \mathbf{j} corresponds to multiple poses θ . Shaded nodes represent observations, and white nodes represent those not observed from the data.

the data distribution rather than use the available data samples themselves purely. The visibility we propose naturally considers ambiguities in occlusion, and 2D weak annotations consider depth ambiguity. The figure also shows the conditional independence of θ and I given $\mathbf{j}, \mathbf{c}, \beta$ (red arrow) in Eq. (6), while θ can be predicted directly from I (green arrow) reflected in Eq. (7).

B. Objective Derivation

B.1. KL Divergence

We here derive Eq. (8) in the main text. Given $I, \mathbf{j}, \mathbf{c}$, and β ,

$$\begin{aligned}
 & KL(p_\phi(\theta|I, \mathbf{j}, \mathbf{c}, \beta) || p(\theta|I, \mathbf{j}, \mathbf{c}, \beta)) \\
 &= \int_\theta p_\phi(\theta|I, \mathbf{j}, \mathbf{c}, \beta) \log \frac{p_\phi(\theta|I, \mathbf{j}, \mathbf{c}, \beta)}{p(\theta|I, \mathbf{j}, \mathbf{c}, \beta)} d\theta. \quad (\text{a})
 \end{aligned}$$

By plugging the definitions of the data (Eq. (6)) and model (Eq. (7)) distribution, we have Eq. (a) equal to,

$$\begin{aligned}
 & \int_\theta p_\phi(\theta|I) \log \frac{p_\phi(\theta|I)}{p(\theta|\mathbf{j}, \mathbf{c}, \beta)} d\theta \\
 &= \int_\theta p_\phi(\theta|I) \log \frac{p_\phi(\theta|I)}{\frac{p(\mathbf{j}|\mathbf{c}, \beta, \theta)p(\theta)}{p(\mathbf{j}|\mathbf{c}, \beta)}} d\theta, \quad (\text{b})
 \end{aligned}$$

*Equal contribution.

where $p(\mathbf{j}|\mathbf{c}, \beta) = \int_{\theta} p(\mathbf{j}|\mathbf{c}, \beta, \theta)p(\theta)d\theta$ by Bayes' rule in Eq. (6). Since $p(\mathbf{j}|\mathbf{c}, \beta)$ is constant w.r.t. θ and our learnable parameters ϕ , we can ignore it. Thus, Eq. (b) becomes Eq. (8),

$$\begin{aligned} & - \left(\int_{\theta} p_{\phi}(\theta|I) \log p(\mathbf{j}|\mathbf{c}, \beta, \theta)d\theta + \int_{\theta} p_{\phi}(\theta|I) \log p(\theta)d\theta \right. \\ & \left. - \int_{\theta} p_{\phi}(\theta|I) \log p_{\phi}(\theta|I)d\theta \right) \\ = & - \underbrace{\left(\frac{E}{p_{\phi}(\theta|I)} [\log p(\mathbf{j}|\mathbf{c}, \beta, \theta)] \right)}_{\text{reconstruction}} + \underbrace{\left(\frac{E}{p_{\phi}(\theta|I)} [\log p(\theta)] \right)}_{\text{prior}} + \underbrace{H(p_{\phi}(\theta|I))}_{\text{entropy}} \end{aligned} \quad \text{e}$$

B.2. Specific Forms

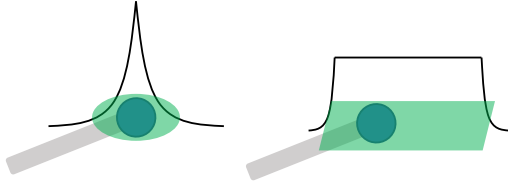


Figure b. Laplace and uniform distributions are used for visible and occluded joints respectively. The green is the valid region.

We assume that all keypoints are conditionally independent, *i.e.*, $p(\mathbf{j}|\mathbf{c}, \beta, \theta) = \prod_k p(\mathbf{j}_k|\mathbf{c}, \beta, \theta)$, where k indexes the keypoint, and $\mathbf{j}_k \in \mathbb{R}^2$.

Reconstruction. For *visible keypoints*, we expect them to be accurate estimates, so we define the reconstruction $p(\mathbf{j}|\mathbf{c}, \beta, \theta)$ in Eq. (8) as,

$$\text{Laplace}(\mathbf{j}_k|\hat{\mathbf{j}}_k, b\mathbf{I}) = \frac{1}{(2b)^2} \exp\left(-\frac{\|\mathbf{j}_k - \hat{\mathbf{j}}_k\|_1}{b}\right), \quad \text{(c)}$$

where b is the scale hyper-parameter, $\hat{\mathbf{j}}_k = \text{proj}(\mathbf{c}, \beta, \theta) = s\Pi(\mathbf{R}\mathcal{J}(\theta, \beta)) + \mathbf{t}$ in Eq. (1) of the main text.

Prior. We follow standard practices to use pose priors for human hands and bodies. Specifically, β are PCA coefficients for both human hands and bodies and are predicted deterministically. We follow Eq. (2) and use an l_2 regularization on β . For bodies, θ are axis-angle rotations, and the adversarial prior [9] is used. For hands, θ are PCA coefficients and can be restricted simply as a uniform distribution $\mathcal{U}(\theta|[-2, 2]^{45})$ [21]. Specifically, we use a softening uniform for optimization [19, 26], *i.e.*, penalizing the out-of-range part along each component, and get the following loss.

$$\mathcal{L}_{\theta} = -\log \text{SoftU}(\theta|[-2, 2]^{45}) \stackrel{c}{=} \sum_{i=1}^{45} \max(0, |\theta_i| - 2)^2, \quad \text{(d)}$$

where $\text{SoftU}(x|[-a, a]) \stackrel{c}{=} \exp(-\max(0, |x| - a)^2)$.

Finally, we obtain the final training objective Eq. (12).

Remarks. Eq. (8) is derived for visible keypoints regardless of occluded ones. Here, we show that the occluded keypoints do not contribute to the final loss with the assumption that the occluded region is large enough relative to the hand/human. The assumption is reasonable as we focus on the cases of large object occlusion (*i.e.*, HO3D) and image truncation (*i.e.*, AH36M) in this paper.

We consider a data distribution integrating possible *underlying* $\bar{\mathbf{j}}$ including both visible and occluded joints as follows,

$$p(\theta|I, \mathbf{c}, \beta) = \int_{\bar{\mathbf{j}}} p(\theta|\bar{\mathbf{j}}, \mathbf{c}, \beta)p(\bar{\mathbf{j}}|I)d\bar{\mathbf{j}}. \quad \text{(e)}$$

Similar to the derivation of Eq. (8), the only difference is the term $p(\mathbf{j}|\mathbf{c}, \beta, \theta)$ of reconstruction term becomes,

$$\int_{\bar{\mathbf{j}}} \frac{p(\bar{\mathbf{j}}|\mathbf{c}, \beta, \theta)p(\bar{\mathbf{j}}|I)}{p(\bar{\mathbf{j}}|\mathbf{c}, \beta)} d\bar{\mathbf{j}}. \quad \text{(f)}$$

For the term $p(\bar{\mathbf{j}}|\mathbf{c}, \beta, \theta)$, we simply assume it is a deterministic projection and we get $p(\bar{\mathbf{j}}|\mathbf{c}, \beta, \theta) = \delta(\bar{\mathbf{j}}|\hat{\mathbf{j}})$. For $p(\bar{\mathbf{j}}|I)$, we assume occluded joints have tolerance to locations and uniformly distribute around feasible locations inside the occluded region and we get $p(\bar{\mathbf{j}}|I) \stackrel{c}{=} \mathcal{U}(\bar{\mathbf{j}}|\Omega(I))p(\bar{\mathbf{j}}|\mathbf{c}, \beta)$. Here, $\Omega(I)$ denotes the occluded region. With the assumption that all keypoints are conditionally independent, Eq. (f) can be reformulated as,

$$\int_{\bar{\mathbf{j}}_k} \delta(\bar{\mathbf{j}}_k|\hat{\mathbf{j}}_k)\mathcal{U}(\bar{\mathbf{j}}_k|\Omega(I))d\bar{\mathbf{j}}_k = \mathcal{U}(\bar{\mathbf{j}}_k = \hat{\mathbf{j}}_k|\Omega(I)), \quad \text{(g)}$$

where $\hat{\mathbf{j}}_k$ is the 2D projected keypoint. For the uniform distribution, we also use a softening version to penalize the out-of-range part similar to Eq. (d), which gives,

$$\text{SoftU}(\epsilon|[-a, a]^2) \stackrel{c}{=} \exp\left(-\sum_{d=1}^2 \max(0, |\epsilon_d| - a)^2\right), \quad \text{(h)}$$

where the occluded region Ω is approximated by a square $S(\mathbf{o}, 2a)$ centered at \mathbf{o} with a width of $2a$ (Fig. b), the deviation from the joint to the center $\epsilon = \hat{\mathbf{j}}_k - \mathbf{o} \sim \mathcal{U}([-a, a]^2)$, ϵ_d indicates the d^{th} dimension of ϵ . We can see that when a is *large enough* relative to the hand scale, *i.e.*, $|\epsilon_d| < a$, this term becomes 0. For example, in ARHD, a is around 50 pixels and a projection is seldomly out of the occluded region.

Omitting constant terms (*i.e.*, additive and multiplicative terms), we combine Eqs. (c) and (h) and have the reconstruction term for both visible and occluded joints,

$$\begin{cases} \|\mathbf{j}_k - \text{proj}(\mathbf{c}, \theta, \beta)\|_1, & v_k = 1, \\ \sum_{d=1}^2 \max(0, |\epsilon_d| - a)^2 = 0, & v_k = 0. \end{cases} \quad \text{(i)}$$

Thus, the reconstruction loss \mathcal{L}_{rec} is summed over joints as,

$$\mathcal{L}_{rec} = \sum_k v_k \|\mathbf{j}_k - \hat{\mathbf{j}}_k\|_1. \quad (\text{j})$$

C. Implementation Details

C.1. Architectures

Feature Extractors. For the toy problem, we use a 3-layer MLP. For ARHD, we use an ImageNet [3] pre-trained ResNet-18 with $\mathbf{f} \in \mathbb{R}^{512}$ [23]. For HO3D and (A)H36M, we use a ResNet-50 with $\mathbf{f} \in \mathbb{R}^{2048}$ [7, 9]. We use the same backbones for all methods.

Normalizing Flows. For hands, we use a lightweight and concise implementation of the Real NVP [4]¹. In particular, it mainly includes affine coupling layers [4] and does not include random permutation [4] or multi-scale structures [4]. This is because their effects may not be so significant in non-image generation tasks. Our NF network contains 12 coupling layers, and each coupling layer consists of 3 linear layers with 256 hidden units. For humans, we follow ProHMR [14] to use Glow [11].

C.2. Training & Hyper-Parameters

Training. The Adam optimizer is used with default parameters [10]. The learning rate of each parameter group is decayed from the initial $2e^{-4}$ by $\gamma = 0.1$ twice. The batch size is set to 64. We clip the gradient norm of iterable parameters for more stable training. We train all the models to converge, typically for 260 epochs. Generative models like NFs usually take longer to converge than discriminative models [11]. We apply standard random scale, translation, rotation, and color jitter data augmentation. For hands, we set hyper-parameters with $\lambda_{rec} = \frac{1}{0.02} = 50$, $\lambda_\theta = \frac{50}{4} = 12.5$, $\lambda_H = -1$. The loss is averaged across batches. Effects of the hyper-parameters are shown in Tab. c. For humans, we set hyper-parameters with $\lambda_{rec} = \frac{1}{0.01} = 100$, $\lambda_\theta = \lambda_\beta = 10$, $\lambda_H = -1$.

As shown in the pipeline overview in Fig. 2, during training, we have the following steps:

1. Extract the image feature \mathbf{f} from I ;
2. Predict \mathbf{c} and β based on \mathbf{f} ;
3. Sample $S \mathbf{z}_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, transform it to θ through $\theta = \mathcal{F}(\mathbf{z}_0|\mathbf{f})$, and compute \mathcal{L}_H in Eq. (11) and (4);
4. Compute and optimize the final objective Eq. (12).

It is standard to optimize log-likelihood and entropy with SGD by taking one [12] or more samples. We find that

¹Based on <https://github.com/senya-ashukha/real-nvp-pytorch/blob/master/real-nvp-pytorch.ipynb>

taking more samples helps entropy optimization and convergence (Fig. c); we choose $S = 10$ samples to balance performance with the computational expense.

During testing, for sampling, we similarly do the first three steps of training.

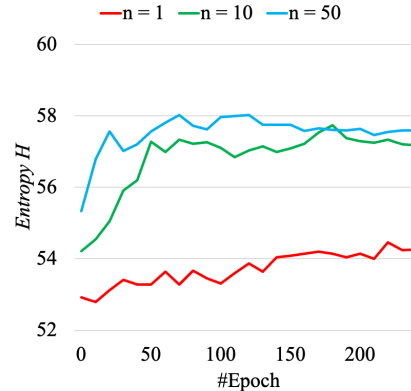


Figure c. Entropy (1K samples) curves with different MC sampling numbers $S = 1, 10, 50$ on ARHD.

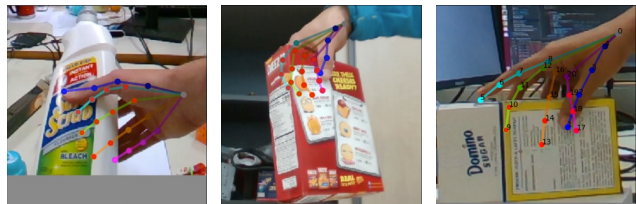
Training Strategy on H36M. We follow [14, 2, 13]’s mixed data training with MPII [1], MPI-INF-3DHP [20], UP-3D [15], and MS-COCO [18].

D. Data Processing Details

Toy. We take 4 Gaussians centered at $(\alpha_1, \alpha_2) = (\pm \frac{\pi}{4}, \pm \frac{\pi}{6})$ with a standard deviation of 0.05 and draw 512 samples from them. We compute the y -projection from the poses α , and add Gaussian noise with $\sigma = 0.01$ to create a toy dataset.

ARHD. We are motivated by [2] to consider constructing occlusion. Instead, we simulate the hand occluded by an object. For each image of ARHD, we fixedly select 1 of the 5 DIPs as the center and add a black circular patch with a radius of 50 pixels. That is, we change the data before training, which will not change anymore during training. It can be determined whether each keypoint is occluded knowing the range of the added patch.

HO3D. HO3D V3 itself does not release ground truths for the test dataset officially. We split the test set from the annotated training dataset to evaluate our metric, including



(a) Bleach cleanser (b) Cracker box (c) Box of sugar

Figure d. Some of our HO3D test samples and their visibility annotation.

BH. We select all frames of the ABF14, MC5, SB14, and ShSu13 clips from the dataset as the test set (Fig. d). They cover the actor’s hands and objects seen in the training set as well as unseen poses and perspectives. For visibility, if the difference between the depth calculated from the 3D coordinates of the keypoint and the depth on its 2D projection position is greater than a threshold (40 mm, the thickness of the wrist), it is considered occluded [6]. We also perform the manual verification of visibility annotations (Fig. d). **AH36M**. The visibility of out-of-view keypoints is set to 0.

E. Evaluation Details

Visible & Occluded EPE of BH are also separately reported in Tab. a supplementary to Tab. 1(a).

PJD & Gaussian Entropy. Standard deviation in PJD is closely related to Gaussian entropy which is tractable as,

$$H(\mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})) = \frac{1}{2} \sum_d (\log \sigma_{k,d}^2 + \log 2\pi + 1) \quad (k)$$

$$= \log \prod_d \sigma_{k,d} + C, \quad (l)$$

where k and d index keypoints and dimensions, respectively. The root is not included because the standard deviation after the centralization is 0.

Sampling & Runtime. For the computation of BH and PJD, we draw 200 samples, following previous work [22, 24]. Results are consistent across multiple runs of training and evaluation. STD across BH evaluations is 0.013. The impact of sample sizes/hypothesis numbers on BH is also shown in Fig. e. BH improves and our advantage is more pronounced with increasing sample sizes, up to 5.65mm lower than Det (2D Vis). It takes 0.023s per image on A5000, 0.014s for 10, 0.028s for 1000 samples. Besides, faithful standard deviation (PJD) requires some amount of samples.

	BH (mm)↓	
	Vis	Occ
Det (3D) [27]	22.44	21.88
Det (2D Vis)	25.03	28.13
Multi-bodies [2]	<u>21.97</u>	<u>21.53</u>
MDN [16]	22.63	22.61
CVAE [22]	22.05	22.43
ProHMR [14]	24.25	27.36
CM-VAE [23]	23.22	23.29
WS3DPG [17]	24.23	27.55
Ours	21.91	20.40

Table a. EPE of the best hypothesis on separately visible and occluded joints on ARHD, except all.

F. SOTAs

We briefly introduce some recent state-of-the-art methods, comparing them to ours as well as their connection to our method in the following.

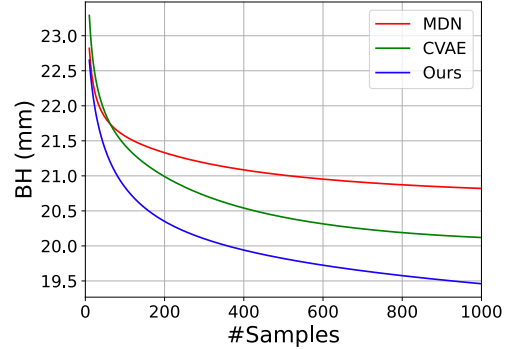


Figure e. BH on ARHD for an increasing number of sample sizes.

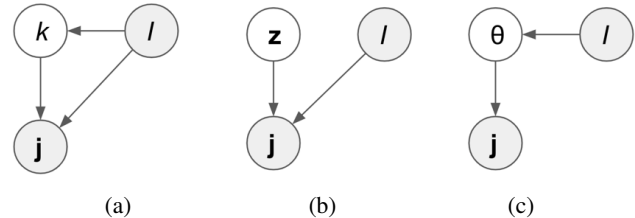


Figure f. Probabilistic Graphical Models (PGMs) of state-of-the-art methods, specifically, (a) MDNs [16], (b) CVAEs [22], (c) ours, CM-VAEs [23], and WS3DPG [17]. Shaded nodes represent observed variables, while white ones represent latent variables.

MDN [16] is designed based on Fig. f(a). It explicitly models k modes for each input. In our experiments, k is set to 10. It is optimized by exactly calculating the likelihood rather than sampling.

CVAE [22] is often used to do conditional generation tasks. The network encodes conditional inputs I and multiple outputs j into the bottleneck latent space \mathcal{Z} (Fig. f(b)). The latent variable z represents the uncertainty when ambiguity occurs, similar to k in MDNs. The optimization objective Evidence Lower Bound (ELBO) is also divided into a reconstruction and KLD term like ours. Nonetheless, for computation, an additional encoder needs to be introduced during training but not used during test sampling. Moreover, they do not directly attach the entropy maximization objective to the concerned θ as we do. Instead, they apply KLD constraints on the \mathcal{Z} space. Under our setting, we report the BH results, for which oracle ground truths are used. **Multi-bodies [2]** is similar to MDNs but based on a deterministic framework to generate multiple hypotheses. Under the weak supervision setting, we change the corresponding best-of-M losses; for the best and other modes, we optimize only visible keypoints. The single point k they obtain with arg min is similar to the k and z found in MDNs and CVAEs, respectively. However, they do not explicitly incentivize diversity to avoid the convergence of generated modes. We use all hypotheses generated by 200 heads for evaluation without requiring quantization.

ProHMR [14] also uses NFs to model θ , instead of point

prediction as deterministic in HMR [9]. In a weakly-supervised setting, the objective is almost equivalent to our objective without the entropy term, *i.e.*, only the reconstruction and prior term. Note that the mode loss in the original paper optimizes the predictive ability instead of diversity.

CM-VAE [23] & WS3DPG [17] all just use different model choices under our framework. From the PGM in Fig. f(c), we can merge θ and \mathbf{z} (in other PGMs), *i.e.*, directly treat the parameters θ as a latent variable. The Cross-Modal VAE (CM-VAE) [23] uses a Single Gaussian Network (SGN) to predict from one modal I to another \mathbf{j} , and the WS3DPG uses an implicit Latent Variable Model (LVM), Generative Adversarial Nets (GANs) [5] when we use NFs. For GANs, the computation of entropy is known to be intractable. A mutual information lower bound [8] and some empirical losses [25] can usually be used to approximately optimize entropy.

Furthermore, as much as possible in our experiments, we use architectures, hyper-parameters, and training strategies similar to the original paper.

G. More Experimental Results

G.1. Toy

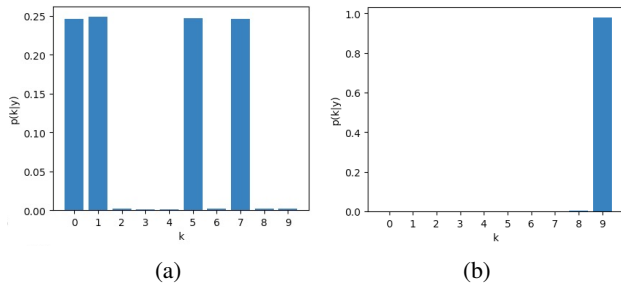


Figure g. Modes learned by MDNs under (a) strong and (b) weak supervision, respectively.

MDNs under Different Supervision. In Fig. g, we show the modes learned by MDNs under strong and weak supervision, respectively, in Sec. 5.2. The MDN learns all the modes given complete strong supervision (*i.e.*, all 4 modes) while only fitting one of them under incomplete strong (partial modes) or weak supervision (*i.e.*, 1D projection \mathbf{y}). This indicates that MDNs have the ability to fit modes explicitly existing in the data but not in other cases.

G.2. ARHD

Visibility Setting Ablation. We also discuss the use of visibility in Tab. b. We show the baseline ‘Det w/ all’ and ‘Ours w/ all’ using all 2D keypoints as weak labels for training. Compared to ‘Ours’ with only visible keypoints, ‘Ours w/ all’ requires more labor to obtain labels without the benefits of BH. Moreover, the occluded labels harm the diversity of occluded keypoints.

	BH (mm)↓	AH (pix)↓	PJD		RD↓
			2D Vis	3D Occ	
Det w/ all	24.33	16.48	-	-	-
Ours w/ all	21.83	16.10	3.55	6.09	0.58
Ours	20.35	13.42	3.86	14.42	0.27

Table b. Ablation study on the influence of keypoint visibility on ARHD.

b	λ_θ	BH (mm)↓	AH (pix)↓	PJD		RD ↓
				2D Vis	3D Occ	
0.02	5	21.59	14.58	3.72	15.87	0.23
	50	21.76	14.91	3.27	12.44	0.26
	500	20.66	13.96	3.54	11.76	0.30
0.01	50	22.21	14.60	2.51	9.57	0.26
		21.54	16.16	4.73	15.73	0.30

Table c. Effects of loss weights. Models are trained for 260 epochs.

Trade-Off among Accuracy, Feasibility, & Diversity.

Tab. c demonstrates a similar trend compared to those in toy experiments. Smaller b leads to better evidence reconstruction (*i.e.*, lower AH and 2D Vis PJD) sacrificed with diversity (*i.e.*, lower 3D Occ PJD) while smaller λ_θ with less feasibility constraint favors diversity as well.

	BH (mm)↓	3D Occ	RD↓	Err↓
Ours	20.35	14.42	0.27	0.09
Ours+PS	19.39	12.93	0.28	0.00
Ours+ L_{rec}	20.38	12.72	0.30	0.06

Table d. Error rate of out-of-occlusion keypoints (Err) and 3D Occ (PJD) of ours with Post-Selection (PS) and reconstruction loss (L_{rec}), respectively. Note that the Error of MDN [16] is 0.10.

Meaningful Diversity. See Tab. d supplementary to the text described in the remarks of Sec. 5.3. The consistency of our framework with additional information improves without much loss of diversity. Though they are experiments on ARHD, they may be more readable in Sec. 5.3’s context.

Det (2D Vis)	Multi-bodies [2]	MDN [16]	CVAE [22]	Ours
<u>25.11</u>	25.60	27.37	27.10	25.05

Table e. LH ($n = 1$) in mm on ARHD. A lower score is better.

Most Likely Hypothesis (LH). As per [17], based on the hypothesis with the highest probability (Tab. e), which is quantized from 200 normally sampled samples using K-Means [17, 2, 14]. Ours improves over baselines in a single prediction.

	Consistency↑	Diversity↑	Similarity
MDN [16]	3.44	3.00	2.80
Ours	3.64	3.80	

Table f. User perceptual study on ARHD. Each score ranges from 1 to 5.

User Perceptual Study. We surveyed 15 people to evaluate 5 hypotheses from our method vs. 5 hypotheses from MDN [16] for 20 images from ARHD. Our hypotheses are rated more diverse and consistent with the images (Tab. f).

	BH (mm)↓	2D Vis
Det	22.63	-
MDN [16]	<u>20.24</u>	6.79
ProHMR [14]	22.23	0.13
Ours	19.27	3.45

Table g. Generalization from ARHD to RHD.

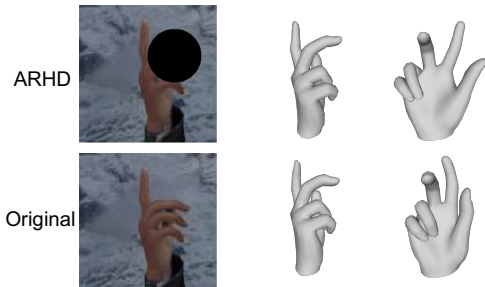


Figure h. Depth ambiguity of index fingers. Our two hypotheses are the same from the front view but different from the side view.

Generalization to the original RHD. Table g shows that all methods generalize, though we maintain a clear advantage.

Depth Ambiguity Visualizations. We visualize two hypotheses and show that our method can handle depth ambiguity (Fig. h).

G.3. HO3D

Single-View		Multi-View	
Det (2D Vis)	Ours	MDN [16]	Ours
24.87	26.49	22.30	22.15

Table h. EPE in mm of hypothesis selection with multi-view images on HO3D. A lower score is better.

Multi-View Hypothesis Selection. Apart from hypothesis selection based on grasp feasibility in the manuscript Fig. 7, we also show hypothesis selection using the multi-view images from the set of calibrated cameras. Tab. h shows that ours disambiguates with the help of multi-view images and improves the EPE from 26.49 mm to 22.15 mm. Moreover, with multi-view hypothesis selection, ours outperforms MDN.

References

- [1] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2D human pose estimation: New benchmark and state of the art analysis. In *CVPR*, 2014. 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. 3, 4, 5
- [3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 3
- [4] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real NVP. In *ICLR*, 2017. 3
- [5] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 5
- [6] Kerui Gu, Linlin Yang, and Angela Yao. Removing the bias of integral pose regression. In *ICCV*, 2021. 4
- [7] Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. HONnotate: A method for 3D annotation of hand and object poses. In *CVPR*, 2020. 3
- [8] 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. 5
- [9] 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, 5
- [10] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *ICLR*, 2015. 3
- [11] Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In *NeurIPS*, 2018. 3
- [12] Diederik P. Kingma and Max Welling. Auto-encoding variational Bayes. In *ICLR*, 2014. 3
- [13] 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. 3
- [14] Nikos Kolotouros, Georgios Pavlakos, Dinesh Jayaraman, and Kostas Daniilidis. Probabilistic modeling for human mesh recovery. In *ICCV*, 2021. 3, 4, 5, 6
- [15] Christoph Lassner, Javier Romero, Martin Kiefel, Federica Bogo, Michael J. Black, and Peter V. Gehler. Unite the people: Closing the loop between 3D and 2D human representations. In *CVPR*, 2017. 3
- [16] Chen Li and Gim Hee Lee. Generating multiple hypotheses for 3D human pose estimation with mixture density network. In *CVPR*, 2019. 4, 5, 6
- [17] Chen Li and Gim Hee Lee. Weakly supervised generative network for multiple 3D human pose hypotheses. In *BMVC*, 2020. 4, 5
- [18] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, 2014. 3
- [19] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. In *NeurIPS*, 2020. 2
- [20] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3D human pose estimation in the wild using improved cnn supervision. In *3DV*, 2017. 3
- [21] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied Hands: Modeling and capturing hands and bodies together. *TOG*, 36(6):1–17, 2017. 2
- [22] 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. 4, 5
- [23] Adrian Spurr, Jie Song, Seonwook Park, and Otmar Hilliges. Cross-modal deep variational hand pose estimation. In *CVPR*, 2018. 3, 4, 5
- [24] Tom Wehrbein, Marco Rudolph, Bodo Rosenhahn, and Bastian Wandt. Probabilistic monocular 3D human pose estimation with normalizing flows. In *ICCV*, 2021. 4
- [25] Dingdong Yang, Seunghoon Hong, Yunseok Jang, Tianchen Zhao, and Honglak Lee. Diversity-sensitive conditional generative adversarial networks. In *arXiv:1901.09024*, 2019. 5
- [26] Junbo Zhao, Michael Mathieu, and Yann LeCun. Energy-based generative adversarial network. In *arXiv:1609.03126*, 2016. 2
- [27] 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. 4