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

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 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).

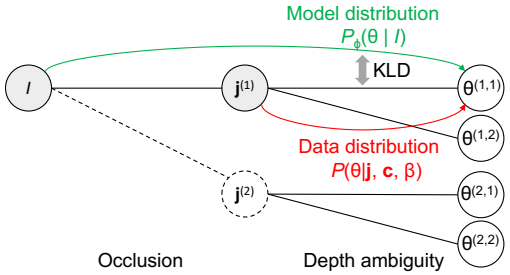


Figure a. An illustration of our problem formulation. One image I corresponds to multiple feasible 2D joints $\mathbf{j}$, while one joint $\mathbf{j}$ corresponds to multiple poses θ. Shaded nodes represent observations, and white nodes represent those not observed from the data.

B. Objective Derivation

B.1. KL Divergence

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

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

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

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

*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),

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

B.2. Specific Forms

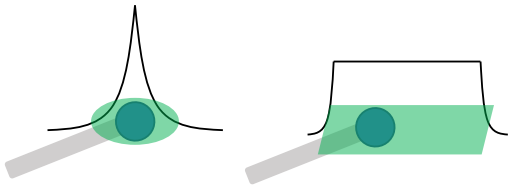


Figure b. Laplace and uniform distributions are used for visible and occluded joints respectively. The green is the valid region.

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, $\hat{\mathbf{j}}_k = proj(\mathbf{c},\beta,\theta) = s\Pi(\mathbf{R}\mathcal{J}(\theta,\beta)) + \mathbf{t}$ 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.

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

where $Soft\mathcal{U}(x|[-a,a]) \overset{c}{=} \exp(-\max(0,|x|-a)^2)$.

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,

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

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

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

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 $p(\bar{\mathbf{j}}|I) \overset{c}{=} \mathcal{U}(\bar{\mathbf{j}}|\Omega(I))p(\bar{\mathbf{j}}|\mathbf{c},\beta)$. Here, $\Omega(I)$ denotes the occluded region. With the assumption that all keypoints are conditionally independent, Eq. (f) can be reformulated as,

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

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,

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

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,

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

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 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.

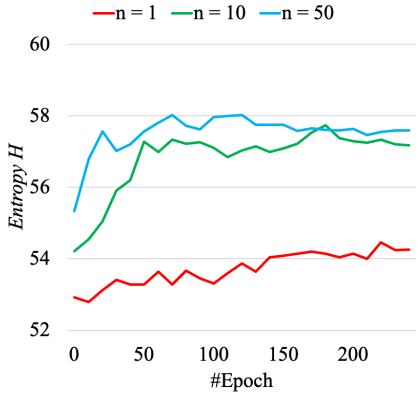


Figure c. Entropy (1K samples) curves with different MC sampling numbers $S = 1, 10, 50$ on ARHD.

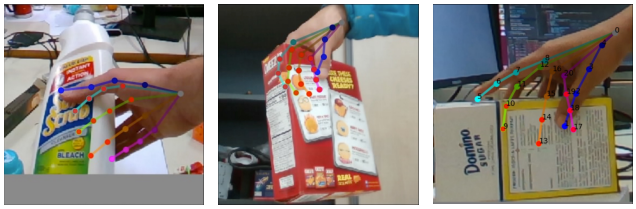
Training Strategy on H36M. We follow [14, 2, 13]'s mixed data training with MPII [1], MPI-INF-3DHP [20], UP-3D [15], and MS-COCO [18].

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.

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

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,

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

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

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.

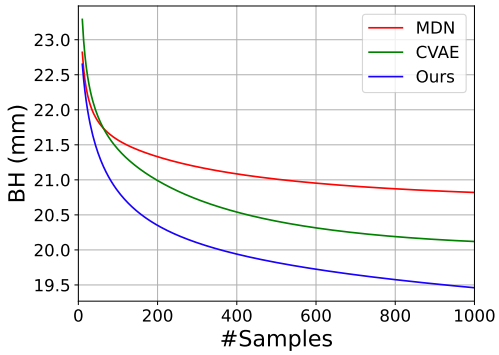
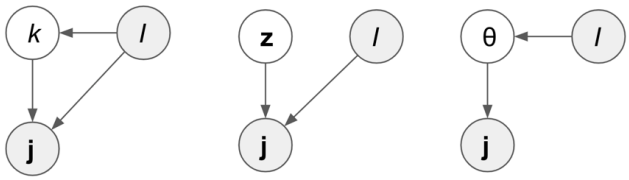


Figure e. BH on ARHD for an increasing number of sample sizes.



(a) (b) (c)

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 $\mathbf{j}$ into the bottleneck latent space $\mathcal{Z}$ (Fig. f(b)). The latent variable $\mathbf{z}$ 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 $\mathbf{z}$ 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

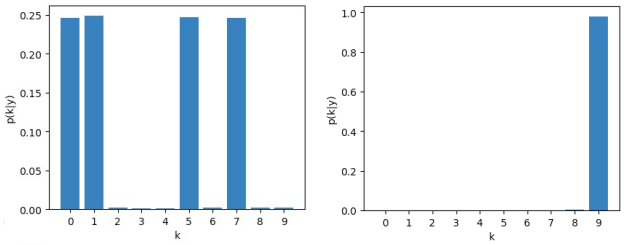
prediction as deterministic in HMR [9]. In a weakly-supervised setting, the objective is almost equivalent to our objective without the entropy term, *i.e.*, only the reconstruction and prior term. Note that the mode loss in the original paper optimizes the predictive ability instead of diversity.

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



(a) (b)

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
0.05		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
25.11	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]	20.24	6.79
ProHMR [14]	22.23	0.13
Ours	**19.27**	3.45

Table g. Generalization from ARHD to RHD.

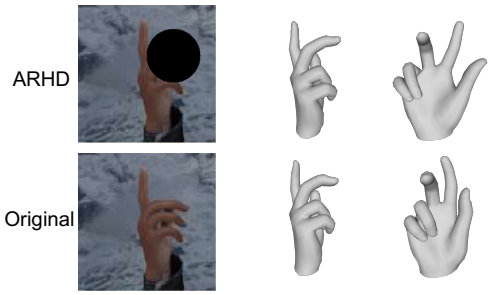


Figure h. Depth ambiguity of index fingers. Our two hypotheses are the same from the front view but different from the side view.

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