

HandFlow: Quantifying View-Dependent 3D Ambiguity in Two-Hand Reconstruction with Normalizing Flow

J. Wang¹ , D. Luvizon¹ , F. Mueller² , F. Bernard³, A. Kortylewski^{1,5} , D. Casas⁴ , and C. Theobalt¹

¹ MPI Informatics & Saarland Informatics Campus, Germany

² Google Inc

³ University of Bonn, Germany

⁴ Universidad Rey Juan Carlos, Spain

⁵ University of Freiburg, Germany

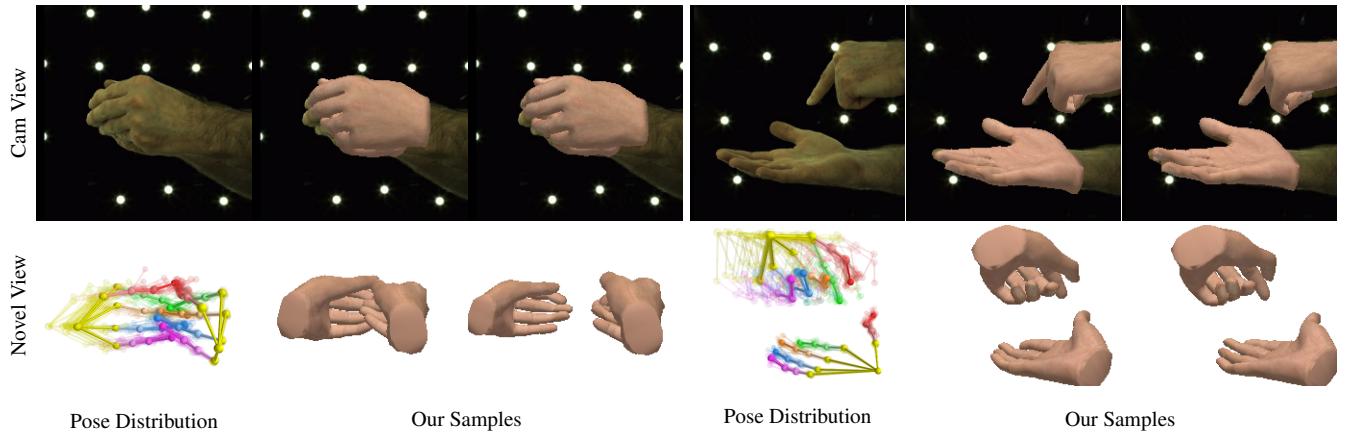


Figure 1: Given a single RGB image of two-hand interaction, our method predicts a distribution of plausible 3D hand poses. We show that samples from our method project well into the image (top) while being diverse in 3D (bottom), which demonstrate our method’s ability to capture the inherent ambiguity in the task.

Abstract

Reconstructing two-hand interactions from a single image is a challenging problem due to ambiguities that stem from projective geometry and heavy occlusions. Existing methods are designed to estimate only a single pose, despite the fact that there exist other valid reconstructions that fit the image evidence equally well. In this paper we propose to address this issue by explicitly modeling the distribution of plausible reconstructions in a conditional normalizing flow framework. This allows us to directly supervise the posterior distribution through a novel determinant magnitude regularization, which is key to varied 3D hand pose samples that project well into the input image. We also demonstrate that metrics commonly used to assess reconstruction quality are insufficient to evaluate pose predictions under such severe ambiguity. To address this, we release the first dataset with multiple plausible annotations per image called MultiHands. The additional annotations enable us to evaluate the estimated distribution using the maximum mean discrepancy metric. Through this, we demonstrate the quality of our probabilistic reconstruction and show that explicit ambiguity modeling is better-suited for this challenging problem.

CCS Concepts

- Computing methodologies → Tracking; Computer vision; Neural networks;

1. Introduction

Reconstructing two interacting hands in 3D is an actively researched topic, as it enables applications in various areas of vision

and graphics, including augmented and virtual reality, robotics, or sign language translation. While earlier methods leverage multi-camera setups [BTG*12, SOT13] or depth sensors [TTT*17,

[MDB*19](#), recent works focus on using monocular RGB cameras to enable potential applications in mobile or wearable settings.

However, hand pose estimation from monocular RGB images is a very challenging problem. Hand interactions lead to severe occlusions; and monocular color images exhibit an inherent depth and scale ambiguity. Existing methods [[WMB*20](#), [MYW*20](#), [ZWD*21](#)] aim to *deterministically* estimate the relative depth between the two hands directly. However, this is prone to error in heavily occluded situations due to the ill-posed nature of the problem. For example, a small error in hand scale or depth can cause a significant difference in touch points and hence semantics of the interaction. As a result, most methods evaluate each hand pose independently using the root-relative pose error which discards important information regarding the positioning of the hands.

Given these extreme challenges, we take a different approach and propose to explicitly model the ambiguities instead (see Fig. 1). Inspired by previous work on reconstruction of human body and face [[KPJD21](#), [WRRW21](#), [KWMF*18](#), [SEMFV17](#)], we propose to predict a distribution over likely two-hand poses. To this end, we adopt normalizing flow [[RM15](#)] as a way to parameterize the posterior distribution that enables not only fast sampling but also differentiable likelihood estimates. This allows us to formulate a novel loss to supervise the shape of the distribution. Our proposed regularization term encourages diversity in distribution without sacrificing image consistency, which is key to model the severe ambiguities in our setting.

We quantitatively demonstrate that our sampled reconstructions capture the range of plausible articulations better than existing state-of-the-art methods. This is facilitated by our new dataset, MultiHands, the first to provide multiple plausible annotations per image for measuring the accuracy of distribution predictions.

In summary, our main contributions are:

- A method for reconstructing two-hand interactions that can generate diverse 3D poses which match the observed image.
- A new regularization term for training conditional normalizing flow to encourage diversity of samples.
- The first dataset to account for pose ambiguity by providing multiple pose annotations.

Finally, we demonstrate that the estimated pose distribution can be leveraged for unambiguous view-point selection, a downstream application not possible with deterministic approaches.

2. Related Work

The majority of existing works investigate the reconstruction of a single hand in free air or with a rigid object. These methods use input data that ranges from multi-camera setups [[BTG*12](#), [SOT13](#)], over depth sensors [[SMOT15](#), [MAE*20](#)], to monocular color images [[MBS*18](#), [BBT19](#), [HTB*20](#)]. However, estimating hand poses during interaction with another hand is a significantly greater challenge due to, for example, occlusion and similarity between hands. Thus we focus our discussion on methods tackling reconstruction of two interacting hands.

Multiple Hands. Few existing methods reconstruct two interacting

hands in 3D. Oikonomidis *et al.* [[OKA12](#)] first leveraged a multi-camera setup to mitigate some of the inherent challenges like strong occlusions. While some recent methods still use multi-view setups and markers [[SJMS17](#), [HLW*18](#)], most works have moved to employing single depth sensors for the flexibility of the capture setup [[TTT*17](#), [MDB*19](#)].

Considering a monocular color image as input, Wang *et al.* [[WMB*20](#)] and Moon *et al.* [[MYW*20](#)] simultaneously developed the first two methods for 3D pose estimation of interacting hands. While the former combines machine-learning-based pixel-to-model correspondence prediction with optimization-based model fitting, the latter uses a neural network to predict 3D joint positions directly. Recent extensions make use of visibility [[KKB21](#)] or hand part segmentation [[FSK*21](#)] to help the network to take into account occlusion information. Others were developed to additionally estimate hand surfaces, either as a parametric model [[ZWD*21](#)] or as a mesh [[LAZ*22](#)]. These methods have in common that they deterministically reconstruct the hand interaction. However, during interactions, we often observe heavy occlusions between hands or ambiguous semantics (see Fig. 1). In contrast, our work focuses on tackling the specific challenges of such interactions through a probabilistic approach.

Probabilistic Methods. Explicitly accounting for ambiguities in monocular RGB images is an important problem that has received recent attention in the field of 3D human poses estimation and face reconstruction [[ZBX*20](#), [KPJD21](#), [WRRW21](#), [KWMF*18](#), [SEMFV17](#)]. However, the only existing method for hands is designed for a single hand and uses depth images as input [[YK18](#)]. Our approach is the first to address the more ambiguous case of RGB images with challenging two-hand interactions.

Some recent probabilistic solutions for estimating 3D pose of a single body [[KPJD21](#), [WRRW21](#)] are also based on normalizing flows [[RM15](#)], which can construct a complex distribution from a simple probability density using invertible operations. In our approach, we build upon normalizing flows to address severe monocular depth ambiguities and occlusions in two interacting hands scenarios. To this end, we introduce a new regularization term for the pose distribution and show that it is crucial for encouraging the sample diversity needed to model the ambiguities.

Two-Hand Datasets. Although several datasets exist that contain images of synthetic [[MDB*19](#), [WMB*20](#), [LWM21](#)] or real [[TBS*16](#), [SJMS17](#), [WMB*20](#), [MYW*20](#)] images of two interacting hands, they all provide only a single annotation per image. This is insufficient for monocular RGB reconstruction since multiple plausible poses can fit the image equally well. We argue that these plausible poses should also be considered correct, and propose our new MultiHands dataset to extend the existing InterHand2.6M [[MYW*20](#)] with 100 additional annotations per image. With MultiHands, we are able to quantify the pose ambiguity in an image, and to use a new metric for measuring the distance between predicted and ground-truth pose distributions.

3. Method

The goal of our method is to estimate a distribution of 3D hand poses that are plausible to explain a given monocular color image.

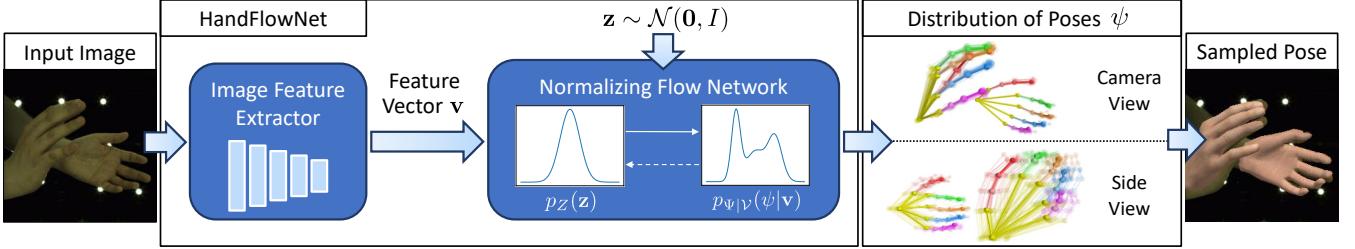


Figure 2: Our HandFlowNet first extracts an image feature vector \mathbf{v} from 2D cues in the input image. The feature vector is then used as conditioning input to a normalizing flow network to output a distribution of 3D hand poses that plausibly explain the monocular input.

To this end, we propose *HandFlowNet*. Our method first extracts a feature encoding from the input image, which we then use to generate the desired output pose distribution from a normalizing flow network (Section 3.2). The estimated 3D hand poses are parameterized using the MANO hand model (Section 3.1).

3.1. Hand Model

We use the MANO hand model [RTB17] to represent the hand surface as well as additional parameters for the rigid transformation. We will first describe the parameterization of a single hand, which is then readily expanded to two hands.

Given 15 joint rotations $R \in \mathbb{R}^{15 \times 3 \times 3}$ represented as stacked rotation matrices and shape parameters $\beta \in \mathbb{R}^{10}$, the MANO model computes the hand surface as mesh and the 3D hand keypoint positions. In order to place the hand correctly relative to the camera, we additionally estimate global rotation parameters $r \in \mathbb{R}^{3 \times 3}$, the hand root position in image coordinates $t \in \mathbb{R}^2$, and the perspective scale factor $s \in \mathbb{R}$. This enables the recovery of the global pose when the focal length is known at inference [BBT19]. The combined global and joint rotations $\{r, R\}$ are parameterized using the 6 DOF representation $\theta \in \mathbb{R}^{16 \times 3 \times 2}$ as proposed in [ZBJ^{*}19]

Therefore, the full set of parameters for a single hand is defined as $\psi = \{\theta, \beta, t, s\} \in \Psi$, where Ψ denotes the parameter space, and the full set of parameters for both hands is defined as $\psi_{\text{both}} = [\psi_{\text{right}}, \psi_{\text{left}}]$. In the following, we will refer to ψ_{both} simply as ψ .

3.2. HandFlowNet

Given a monocular input image, our *HandFlowNet* regresses a distribution of 3D hand poses corresponding to plausible hand poses that could be observed in the image (see Fig. 2). *HandFlowNet* can be divided into two parts, an image feature extractor and a conditional normalizing flow network that produces a 3D pose distribution and is conditioned on the extracted image feature vector.

3.2.1. Image Feature Extractor

The image feature extractor summarizes the visible, unambiguous features that the sampled poses should reconstruct. We use ResNet-50 [HZRS16] as the backbone architecture. From an input image with resolution 224×224 , we extract the 2048-dimensional feature vector $\mathbf{v} \in \mathcal{V}$ from the average pooling of the last residual block, and use it as the conditional vector for the next step.

3.2.2. Normalizing Flow Network

To predict a range of plausible poses, we must first choose a way to parameterize a distribution $p_Y(\mathbf{y})$.

Normalizing flow [RM15] does this by learning an invertible transformation $f : \mathbb{R}^d \rightarrow \mathbb{R}^d$ of a simple distribution $p_Z(\mathbf{z})$, i.e.

$$p_Y(\mathbf{y}) = p_Z(\mathbf{z}) \left| \det \frac{\partial f(\mathbf{z})}{\partial \mathbf{z}} \right|^{-1}. \quad (1)$$

where $\mathbf{y} = f(\mathbf{z})$. This invertible parameterization allows for both differentiable sampling and likelihood estimation. As a result, we can apply losses on each sample to improve reconstruction quality, while supervising the entire distribution using negative log likelihood loss and multiple annotations (see Section 3.2.3).

Since we want to estimate a distribution over the space of 3D hand poses Ψ given an image feature vector \mathbf{v} , we are interested in finding the conditional distribution $p_{\Psi|\mathcal{V}}(\psi|\mathbf{v})$. To this end, normalizing flow can be extended to conditional normalizing flow [WWHW19] by using transformations $f_v : \mathbb{R}^d \rightarrow \mathbb{R}^d$ parameterized by \mathbf{v} , so that we have

$$p_{\Psi|\mathcal{V}}(\psi|\mathbf{v}) = p_{Z|\mathcal{V}}(\mathbf{z}|\mathbf{v}) \left| \det \frac{\partial f_v(\mathbf{z})}{\partial \mathbf{z}} \right|^{-1}. \quad (2)$$

For our implementation, we use the conditional GLOW architecture for f_v which has been successfully used in previous work [KPDJ21] due to its quick sampling and probability estimation. For a more detailed overview of different architectures, we refer to [KPB21].

By setting $p_{Z|\mathcal{V}} = p_Z \sim \mathcal{N}(\mathbf{0}, I)$, the mode of the distribution $p_{\Psi|\mathcal{V}}(\psi|\mathbf{v})$ can be obtained as $f_v(\mathbf{0})$. We choose this design to provide easy access to the mode sample for use in our losses.

3.2.3. Training Losses

In the following, we detail the losses used for training. The entire loss is given by

$$\mathcal{L} = \mathcal{L}_{\text{nll}} + \mathcal{L}_{\text{DetMag}} + \mathcal{L}_\Psi + \mathcal{L}_{\mathcal{J}_{3D}} + \mathcal{L}_{\mathcal{J}_{2D}} + \mathcal{L}_\theta. \quad (3)$$

Here, \mathcal{L}_{nll} and $\mathcal{L}_{\text{DetMag}}$ are used to supervise the likelihood of the annotations, and \mathcal{L}_Ψ , $\mathcal{L}_{\mathcal{J}_{3D}}$, $\mathcal{L}_{\mathcal{J}_{2D}}$, and \mathcal{L}_θ are used to supervise the quality of the sampled reconstructions. For network training parameters and loss weights, please refer to the supplemental document.

Maximum Likelihood Estimation. Given images and their 3D annotation, we want to ensure that the probability of the pose annota-

tion ψ^* is maximized. Hence, we minimize the negative log likelihood (NLL) loss

$$\begin{aligned}\mathcal{L}_{\text{nll}} &= -\ln p_{\Psi|\mathcal{V}}(\psi^*|\mathbf{v}) \\ &= -\ln p_Z(f_{\mathbf{v}}^{-1}(\psi^*)) \left| \det \frac{\partial f_{\mathbf{v}}(\mathbf{z})}{\partial \mathbf{z}} \right|^{-1}.\end{aligned}\quad (4)$$

When multiple annotations $\{\dots, \psi_n^*\}$ are available, the NLL loss is minimized over all annotated poses.

Enhancing Pose Variety. We observe that training the network using just the term \mathcal{L}_{nll} quickly collapses the variety in the output pose distribution. To explain this, we note that \mathcal{L}_{nll} maximizes $\left| \det \frac{\partial f_{\mathbf{v}}(\mathbf{z})}{\partial \mathbf{z}} \right|^{-1}$, which describes the compression factor between the two spaces for density conservation. Therefore, the network can trivially optimize the conditional distribution by concentrating the density in the pose space, leading to the collapse in $p_{\Psi|\mathcal{V}}(\psi|\mathbf{v})$. To prevent this, we add the regularization term

$$\mathcal{L}_{\text{DetMag}} = -\ln \left| \det \frac{\partial f_{\mathbf{v}}(\mathbf{z})}{\partial \mathbf{z}} \right|. \quad (5)$$

Since this term aims at increasing variation in the output distribution of the normalizing flow network only, we do not backpropagate it into the image feature extractor. Otherwise, the extraction network might be hindered in learning pose-relevant features.

Mode Supervision. While \mathcal{L}_{nll} encourages the probability of the pose annotations to be maximized, we also want the mode sample $f_{\mathbf{v}}(\mathbf{0})$ to be a valid reconstruction. We use the loss

$$\mathcal{L}_{\Psi} = \|f_{\mathbf{v}}(\mathbf{0}) - \psi^*\|_2^2. \quad (6)$$

Note that \mathcal{L}_{Ψ} is complementary to \mathcal{L}_{nll} and both together form a two-sided loss that ensures plausible pose predictions. When multiple annotations are available, a single annotation is randomly chosen to act as the mode sample for the entire training procedure.

Although data with MANO parameter annotation exists, the amount is limited compared to the amount of data with joint position annotations. To make use of all available data, we impose the additional 3D joint position loss

$$\mathcal{L}_{\mathcal{J}_{\text{3D}}} = \sum_{i=1}^{N_J} \|\mathcal{J}(\psi)_i - P_i^{\text{3D}}\|_2^2, \quad (7)$$

where \mathcal{J} is a function defined by the hand model that calculates the 3D joint positions given pose parameters ψ , and P^{3D} are the 3D joint position annotations.

2D Consistency. Our *HandFlowNet* aims to provide a distribution of poses that all correspond to the same input image. Hence, the 2D position of visible joints should be the same for the mode and the samples of the distribution, and should thus match the annotation. We employ

$$\mathcal{L}_{\mathcal{J}_{\text{2D}}} = \sum_{i=1}^{N_J} \eta_i \|\Pi(\mathcal{J}(\psi)_i) - P_i^{\text{2D}}\|_2^2, \quad (8)$$

where Π is the known camera projection, P^{2D} are the 2D joint position annotations, and $\eta_i = 1$ if the joint i is visible and 0 otherwise. These visibility scores are computed from the meshes of MANO

pose annotations. We calculate $\mathcal{L}_{\mathcal{J}_{\text{2D}}}$ on the mode of the distribution $f_{\mathbf{v}}(\mathbf{0})$ and on two samples from the estimated distribution $p_{\Psi|\mathcal{V}}(\psi|\mathbf{v})$.

Rotation Regularization. As explained in Section 3.1, we use the continuous 6-dimensional representation for 3D rotations proposed by Zhou et al. [ZBJ^{*}19]. The representation is not unique, i.e., there are multiple $A \in \mathbb{R}^{3 \times 2}$ that represent the same 3D rotation $R \in \text{SO}(3)$. To encourage consistent output, we follow previous work [KPJD21] and add a regularizer that constrains all rotations in their 6-dimensional representation A to be orthonormal

$$\mathcal{L}_0 = \sum_{A \in \Theta} \|A^\top A - I\|_F^2. \quad (9)$$

4. Creating Additional Annotations

While there exists a single *ground-truth* pose, i.e. the one that forms a given image, recovering this exact pose from an image is ambiguous since there are multiple *plausible pose annotations*. Since our goal is to model this ambiguity with a distribution, the single ground truth found in most datasets is not sufficient for evaluating our predictions and more annotations are needed.

Here we describe how we obtain additional annotations from a provided MANO ground truth.

Plausible Pose Annotations: Given the ground-truth pose parameters ψ_{gt} and a camera projection Π , an annotation ψ_{annot} is plausible if the hand joints fit the observed image and the overall articulation is anatomically possible. To ensure this, we use the following criteria:

- The 2D locations of visible joints should be within a pixel threshold of the ground truth locations.
- Occluded joints in the original pose should remain occluded.
- The pose should be anatomically likely as measured using the pose PCA space of the MANO model [RTB17]. A likelihood threshold is used to eliminate extreme articulations.
- No collision between hands. Collisions are detected using Gaussian proxies [MDB^{*}19] attached to the MANO model. Collision occurs when the one-standard-deviation spheres of the Gaussian proxies intersect each other.

Annotation Generation: Starting from the ground-truth pose parameters ψ_{gt} , we perturb the hand pose parameters to generate hand pose proposals. These proposals are checked for plausibility as defined in the above criteria, and implausible pose annotations are rejected. The accepted plausible pose annotations will now serve as new starting poses for the next iteration. This perturbation and plausibility checking is repeated for a fixed number of iterations to obtain the final plausible pose annotations. For additional implementation details, please refer to the supplemental document.

5. Experimental Results

We evaluate our method on existing datasets (Sec. 5.1), and discuss the limitations of commonly used metrics in dealing with ambiguity (Sec. 5.2 and 5.3). To deal with this ambiguity, we propose to use an alternative metric (Sec. 5.3) to evaluate our method (Sec. 5.7, 5.6, 5.7). Finally, we show an application beyond pose estimation to demonstrate the advantages of distribution estimation (Sec. 5.8).

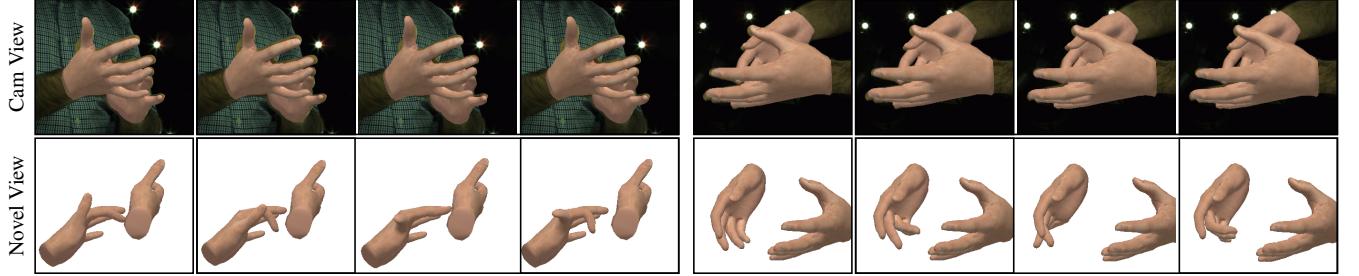


Figure 3: Our MultiHands dataset captures the ambiguities of monocular input with diverse 3D reconstructions that fit the input images.

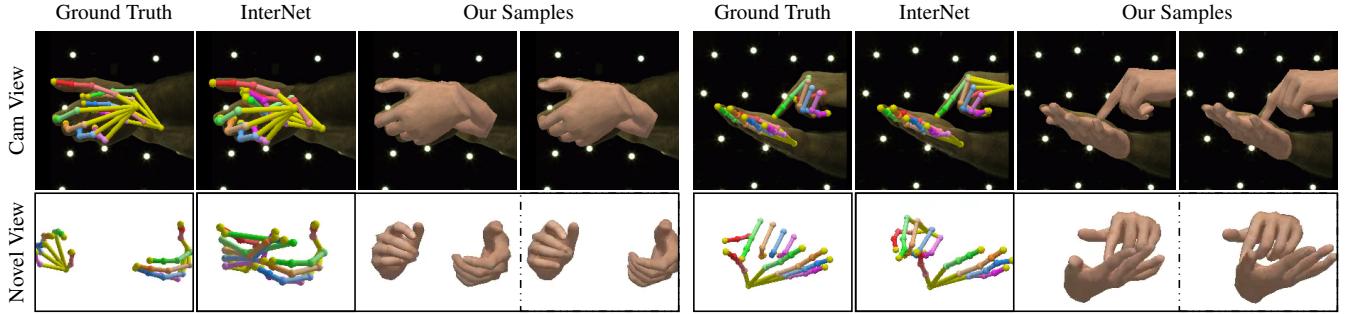


Figure 4: From the camera view, we show that both our samples and InterNet predictions are consistent with the input image (top row). From a novel view (bottom rows), the diversity in our samples is apparent. This diversity allows for samples that are close to the single InterHand2.6M ground truth, while the deterministic InterNet predictions is misled by the ambiguity or fails under heavy occlusions.

5.1. Datasets

Here we describe each dataset and practical considerations that we took into account to run the experiments.

InterHand2.6M Dataset [MYW^{*}20]. We use the 673,514 training frames labeled as interacting hands to train our method. Notice that the terms \mathcal{L}_{nll} , \mathcal{L}_{ψ} from Eq. 4, 6, respectively, require MANO parameter annotations. These losses are applied to the subset of 394,599 frames where these are available.

Following the method of InterNet [MYW^{*}20], we use RootNet [MCL19] for hand detection. A 334×334 crop centered around the provided bounding box is used for the image feature extractor.

MultiHands Dataset: Using the method described in Sec. 4, we propose to extend InterHand2.6M with 100 additional annotations for each of the 281,369 test and 394,599 training images with MANO annotations. Since our losses \mathcal{L}_{nll} and $\mathcal{L}_{\text{DetMag}}$ can use multiple annotations, we also use MultiHands for training. See Fig. 3 and the supplemental document for examples of annotations.

Tzionas Dataset [TBS^{*}16]. To demonstrate that the learned 3D pose distribution generalizes to other settings, we show qualitative results on the Tzionas Dataset. This dataset has seven sequences captured in an office environment with only 2D annotations.

Following Moon *et al.* [MYW^{*}20], we trained with mixed batches on 90% of the annotated 2D frames, and show results on the remaining 10%.

Method	Global MPJPE \downarrow	RRR MPJPE \downarrow	RR* MPJPE \downarrow
InterNet (min)	67.2	24.5	22.6
InterNet (max)	103.6	42.2	24.6
Fan et al. (min)	65.7	27.1	20.5
Fan et al. (max)	102.1	45.9	22.5

Table 1: MPJPE of deterministic estimates can vary wildly depending on the plausible pose annotation used. For RR*, the error is reported for occluded joints. All errors are in mm.

5.2. Pose Alignment

We use three different alignments to evaluate the mean per-joint position error (MPJPE) in mm. All equations can be found in the supplemental document.

Root-Relative MPJPE (RR) captures the errors in articulation, where each hand is individually root-aligned. **Right-Root-Relative MPJPE (RRR)** measures the accuracy of the two hands together, where both hands are aligned to just the root of the right hand. **Global MPJPE (Global)** captures the accuracy of the global pose estimate, *without any alignment*.

Although the RR metric is most commonly reported in the literature, it evaluates the two hands independently by ignoring the relative hand placements. Since this placement is vital for most applications, we show and focus on the RRR and Global metrics.

5.3. Problem with Traditional Metric

When the observed image is ambiguous, the choice of the target pose can greatly impact the MPJPE even though equally valid al-

Method	Global MMD ↓	RRR MMD ↓	RR MMD ↓
Ours	0.50	0.42	0.44
VAE	0.61	0.47	0.48
Gaussian	0.82	0.51	0.46
MCDropout	0.91	0.60	0.51
InterNet	1.12	0.59	0.56
Fan et al.	1.12	0.63	0.50

Table 2: Our method best captures the true distribution of plausible poses using the maximum mean discrepancy (MMD) [GBR*12].

ternative exist. To quantify this effect on InterHand2.6M, we evaluated the InterNet [MYW*20] and Fan et al. [FSK*21] predictions against the closest and farthest annotation in MultiHands (Table 1).

For the challenging Global and RRR metrics, the choice of plausible annotation accounts for a difference of 36mm and 18mm on average. Even when each hand is evaluated independently with the RR metric, the occluded joints differ by 2mm on average.

We argue that this sensitivity to the choice of annotation makes MPJPE unsuitable for the highly ambiguous monocular two-hand reconstruction task. Instead, a metric that measures the distances between pose distributions would better reflect prediction quality.

5.4. Maximum Mean Discrepancy (MMD)

We can measure how well the estimated distribution matches the annotation distribution using the maximum mean discrepancy (MMD) [GBR*12].

The empirical MMD can be estimated given sampled pose predictions, multiple pose annotations, and the selection of a kernel function. We used 100 samples and annotations, and choose Gaussian kernels for our evaluation. All reported MMD are averaged over different kernel distance scales.

5.5. Comparison to the State of the Art

Competing methods. We implement the widely applied probabilistic baselines Monte Carlo dropout (MC-dropout) [GG16], aleatoric uncertainty (Gaussian) [KG17], and Variational Auto Encoder (VAE) [KW14] for comparisons. The implementation details can be found in the supplemental document. As reference, we also compare against deterministic methods [MYW*20, FSK*21] by treating the estimates as a Dirac delta distribution. Given each method, 100 poses are sampled to find the MMD to ground-truth samples. MMD is computed for all alignment to better understand the sources of ambiguity.

Results. Overall, our method produces estimates that best match the ground-truth distribution (Table 2). This is especially notable for the challenging Global and RRR MMD metric, which demonstrates the benefits of our formulation under ambiguity. State-of-the-art deterministic methods fail to account for ground truth variability (Fig. 4). As a result, they have one of the worst MMD.

For reference, a comprehensive evaluation of our method using the single provided InterHand2.6M annotation can be found in the

Method	Global MMD ↓	RRR MMD ↓	RR MMD ↓
Ours	0.50	0.42	0.44
w/o MultiHands	0.53	0.44	0.46
w/o $\mathcal{L}_{\text{DetMag}}$	0.72	0.49	0.46
w/o $\mathcal{L}_{\mathcal{J}_{3D}}$	0.65	0.62	0.52
w/o $\mathcal{L}_{\mathcal{J}_{2D}}$	0.74	0.74	0.46
w/o \mathcal{L}_{ψ}	0.55	0.42	0.45
w/o \mathcal{L}_{θ}	0.61	0.46	0.45

Table 3: All losses and annotations are needed for the best results.

supplemental document. There, we show that our best sample outperforms the state-of-the-art methods while still remaining competitive as a single pose estimator.

5.6. Ablation Study

We show in Table 3 that every loss helps to make our samples match the ground-truth distribution. In particular, our proposed determinant magnitude regularization $\mathcal{L}_{\text{DetMag}}$ is vital for increasing the diversity of 3D samples. The mean standard deviation of the joint positions is improved from 18 to 31 mm while lowering the MMD. Lastly, we observe that using multiple annotations from MultiHands in the \mathcal{L}_{nll} and $\mathcal{L}_{\text{DetMag}}$ terms further improves the MMD, which demonstrate the advantage of the differentiable likelihood estimation in the normalizing flow formulation.

5.7. More Qualitative Results

In Fig. 5, we show qualitative results to demonstrate the diversity and accuracy of our learned pose distribution. Specifically, we show pose samples visualized as superimposed transparent kinematic skeletons. Note that pose variations well reflect the expected monocular ambiguity, and occlusions further increase variability. Hence, the standard deviation of our samples can serve as an indicator for the ambiguity in the input image and thus uncertainty in the pose prediction. See supplemental video for more results.

5.8. Application: View Selection

By using the sample standard deviations to estimate pose ambiguity, we can identify camera views that provide the most information for a given motion sequence. This information can be useful, for example, in a multi-view capture setup where uninformative cameras can be removed to reduce the hardware and data bandwidth requirements. We demonstrate this on the InterHand2.6M test set with over 100 images in the sequence. This consists of the 7 sequences in Capture0-1 with interacting hands, each with 140 camera views.

The view quality is evaluated using regret [BF85] in MPJPE: the difference between the MPJPE on the selected view and the lowest MPJPE. The best and worst views selected by our method have a regret of 3.1 and 15.9 mm respectively, while the average regret over the cameras is 10.7 mm. This shows that our method is able to eliminate cameras with ambiguous views where the monocular pose estimator is not expected to perform well, while keeping cameras views where the estimator is likely to succeed.

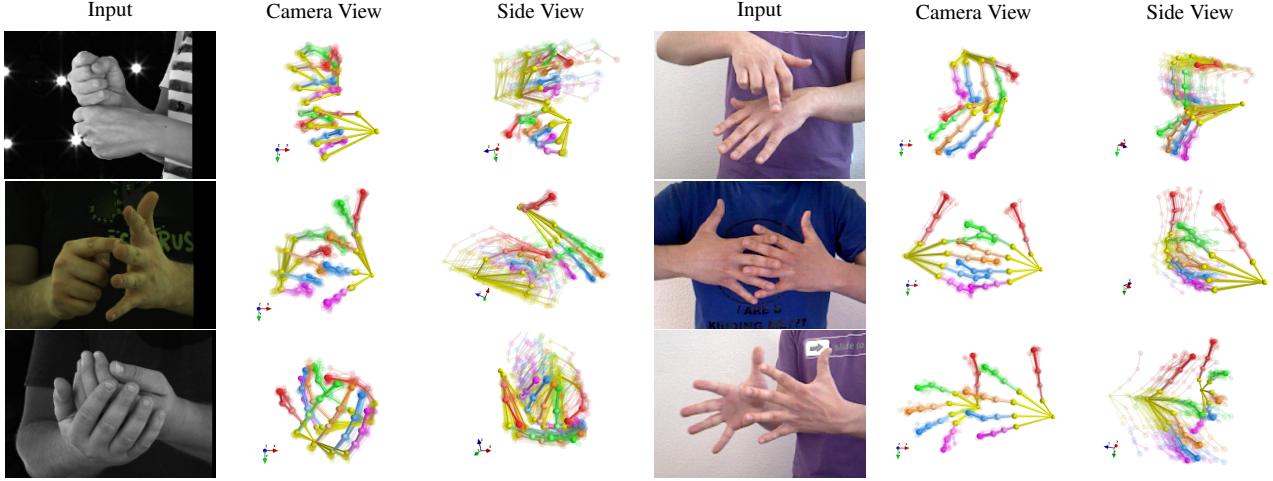


Figure 5: Here we show 30 pose samples superimposed as semi-transparent skeletons. Samples are aligned to the root joint of one hand and the mode of the distribution is made opaque for ease of visualization. The samples are consistent in the camera view, while showing larger variations in novel views. Examples are from the InterHand2.6M dataset (left) and the Tzionas dataset (right), where we transferred learned 3D ambiguity modeling using only 2D annotations.

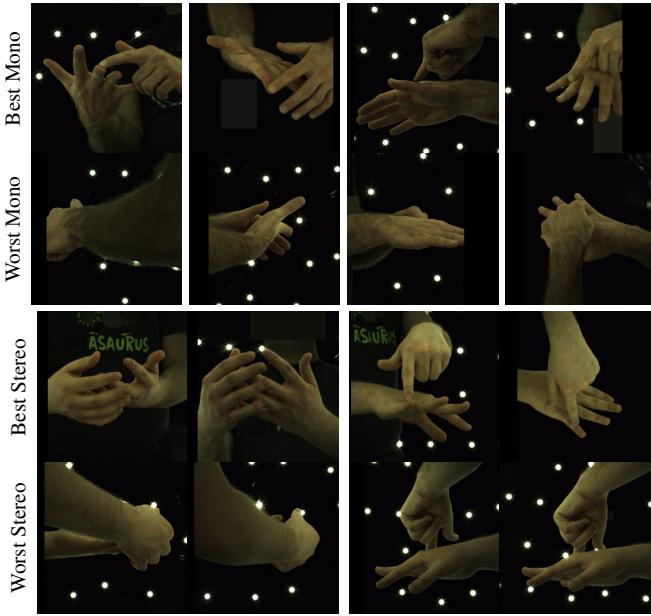


Figure 6: Each column shows the least (**Best Mono**) and most ambiguous views (**Worst Mono**) selected; note the sever occlusions in the worst view. Likewise, the **Best Stereo** pair disambiguate the pose, while the **Worst stereo** pair has similar views.

We can extend view selection to stereo camera pairs by combining two monocular pose distributions. By assuming conditional independence, we can approximating the pose samples from each view with normal distributions and combine them by taking their product. See Fig. 6 and supplemental video for a qualitative evaluation of the selected views.

6. Limitations and Future Work

Although we demonstrated promising results, there are some limitations that could be addressed in future work.

Currently, we do not penalize physically implausible intersections in our reconstructions. As demonstrated in related work [WMB^{*}20, HVT^{*}19], an explicit loss to prevent these intersections could be used to improve the results.

Although we showed promising generalization results on the Tzionas dataset, we did not tackle completely unconstrained in-the-wild images. We believe that in the future this can be solved with more data, especially 2D annotations for in-the-wild data.

While our experiments verified the need for probabilistic pose estimates in ambiguous scenarios, many applications can only make use of a single pose prediction. Future work could investigate ways to integrate additional observations (*e.g.*, temporal information, multi-view images, depth images, task-based priors) to disambiguate the output distribution for a given down-stream task.

7. Conclusion

We have presented the first two-hand reconstruction approach to explicitly model the inherent ambiguities that arise from using a single monocular input image. Given this challenging setting, our method produce a distribution of plausible reconstructions, from which diverse 3D pose samples can be drawn that all explain the observed image evidence. Additionally, we showed that existing evaluation schemes are problematic as they assume a single correct pose even though multiple solutions are equally valid. Along with our proposed dataset with multiple annotations and the distribution metric, we hope our work demonstrates the need for probabilistic approaches and provides a way to evaluate them.

8. Acknowledgments

The work was supported by the ERC Consolidator Grants 4DRepLy (770784) and TouchDesign (772738) and Spanish Ministry of Science (RTI2018-098694-B-I00 VizLearning). AK acknowledges support via his Emmy Noether Research Group funded by the German Science Foundation (DFG) under Grant No. 468670075.

References

- [BBT19] BOUKHAYMA A., BEM R. D., TORR P. H.: 3d hand shape and pose from images in the wild. In *CVPR* (June 2019). [2](#), [3](#)
- [BF85] BERRY D. A., FRISTEDT B.: Bandit problems: sequential allocation of experiments (monographs on statistics and applied probability). *London: Chapman and Hall* 5, 71-87 (1985), 7–7. [6](#)
- [BTG*12] BALLAN L., TANEJA A., GALL J., GOOL L. V., POLLEFEYS M.: Motion Capture of Hands in Action using Discriminative Salient Points. In *ECCV* (2012). [1](#), [2](#)
- [FSK*21] FAN Z., SPURR A., KOCABAS M., TANG S., BLACK M., HILLIGES O.: Learning to disambiguate strongly interacting hands via probabilistic per-pixel part segmentation. In *3DV* (2021). [2](#), [6](#)
- [GBR*12] GRETTON A., BORGWARDT K. M., RASCH M. J., SCHÖLKOPF B., SMOLA A.: A kernel two-sample test. *Journal of Machine Learning Research* 13, 25 (2012), 723–773. [6](#)
- [GG16] GAL Y., GHAHRAMANI Z.: Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *ICML* (2016), PMLR, pp. 1050–1059. [6](#)
- [HLW*18] HAN S., LIU B., WANG R., YE Y., TWIGG C. D., KIN K.: Online optical marker-based hand tracking with deep labels. *AMC TOG* 37, 4 (2018), 166. [2](#)
- [HTB*20] HASSON Y., TEKIN B., BOGO F., LAPTEV I., POLLEFEYS M., SCHMID C.: Leveraging photometric consistency over time for sparsely supervised hand-object reconstruction. In *CVPR* (June 2020). [2](#)
- [HVT*19] HASSON Y., VAROL G., TZIONAS D., KALEVATYKH I., BLACK M. J., LAPTEV I., SCHMID C.: Learning joint reconstruction of hands and manipulated objects. In *CVPR* (2019). [7](#)
- [HZRS16] HE K., ZHANG X., REN S., SUN J.: Deep residual learning for image recognition. In *CVPR* (June 2016). [3](#)
- [KG17] KENDALL A., GAL Y.: What uncertainties do we need in bayesian deep learning for computer vision? In *NIPS* (2017). [6](#)
- [KKB21] KIM D. U., KIM K. I., BAEK S.: End-to-end detection and pose estimation of two interacting hands. In *ICCV* (2021). [2](#)
- [KPB21] KOBYZEV I., PRINCE S. J., BRUBAKER M. A.: Normalizing flows: An introduction and review of current methods. *TPAMI* (2021). [3](#)
- [KPJD21] KOLOTOUROS N., PAVLAKOS G., JAYARAMAN D., DANIILIDIS K.: Probabilistic modeling for human mesh recovery. In *ICCV* (October 2021), pp. 11605–11614. [2](#), [3](#), [4](#)
- [KW14] KINGMA D. P., WELLING M.: Auto-encoding variational bayes. In *ICLR* (2014), Bengio Y., LeCun Y., (Eds.). [6](#)
- [KWMF*18] KORTYLEWSKI A., WIESER M., MOREL-FORSTER A., WIECZOREK A., PARBHOO S., ROTH V., VETTER T.: Informed mcmc with bayesian neural networks for facial image analysis. *Bayesian Deep Learning Workshop (NeurIPS 2018)* (2018). [2](#)
- [LAZ*22] LI M., AN L., ZHANG H., WU L., CHEN F., YU T., LIU Y.: Interacting attention graph for single image two-hand reconstruction. In *CVPR* (June 2022). [2](#)
- [LWM21] LIN F., WILHELM C., MARTINEZ T.: Two-hand global 3d pose estimation using monocular rgb. In *WACV* (2021). [2](#)
- [MAE*20] MALIK J., ABDELAZIZ I., ELHAYEK A., SHIMADA S., ALI S. A., GOLYANIK V., THEOBALT C., STRICKER D.: Handvoxnet: Deep voxel-based network for 3d hand shape and pose estimation from a single depth map. In *CVPR* (June 2020). [2](#)
- [MBS*18] MUELLER F., BERNARD F., SOTNYCHENKO O., MEHTA D., SRIDHAR S., CASAS D., THEOBALT C.: GANerated Hands for Real-Time 3D Hand Tracking from Monocular RGB. In *CVPR* (2018). [2](#)
- [MCL19] MOON G., CHANG J., LEE K. M.: Camera distance-aware top-down approach for 3d multi-person pose estimation from a single rgb image. In *ICCV* (2019). [5](#)
- [MDB*19] MUELLER F., DAVIS M., BERNARD F., SOTNYCHENKO O., VERSCHOOR M., OTADUY M. A., CASAS D., THEOBALT C.: Real-time pose and shape reconstruction of two interacting hands with a single depth camera. *ACM TOG* 38, 4 (2019), 49. [1](#), [2](#), [4](#)
- [MYW*20] MOON G., YU S.-I., WEN H., SHIRATORI T., LEE K. M.: InterHand2.6M: A dataset and baseline for 3D interacting hand pose estimation from a single RGB image. In *ECCV* (2020). [2](#), [5](#), [6](#)
- [OKA12] OIKONOMIDIS I., KYRIAZIS N., ARGYROS A. A.: Tracking the articulated motion of two strongly interacting hands. In *CVPR* (2012), IEEE, pp. 1862–1869. [2](#)
- [RM15] REZENDE D., MOHAMED S.: Variational inference with normalizing flows. In *ICML* (2015), pp. 1530–1538. [2](#), [3](#)
- [RTB17] ROMERO J., TZIONAS D., BLACK M. J.: Embodied hands: Modeling and capturing hands and bodies together. *ACM Trans. Graph.* 36, 6 (Nov. 2017), 245:1–245:17. [3](#), [4](#)
- [SEMFV17] SCHÖNBORN S., EGGER B., MOREL-FORSTER A., VETTER T.: Markov chain monte carlo for automated face image analysis. *IJCV* 123, 2 (2017), 160–183. [2](#)
- [SJMS17] SIMON T., JOO H., MATTHEWS I., SHEIKH Y.: Hand key-point detection in single images using multiview bootstrapping. In *CVPR* (2017). [2](#)
- [SMOT15] SRIDHAR S., MUELLER F., OULASVIRTA A., THEOBALT C.: Fast and Robust Hand Tracking Using Detection-Guided Optimization. In *CVPR* (2015). [2](#)
- [SOT13] SRIDHAR S., OULASVIRTA A., THEOBALT C.: Interactive markerless articulated hand motion tracking using RGB and depth data. In *ICCV* (2013), pp. 2456–2463. [1](#), [2](#)
- [TBS*16] TZIONAS D., BALLAN L., SRIKANTHA A., APONTE P., POLLEFEYS M., GALL J.: Capturing hands in action using discriminative salient points and physics simulation. *IJCV* (2016). [2](#), [5](#)
- [TTT*17] TAYLOR J., TANKOVICH V., TANG D., KESKIN C., KIM D., DAVIDSON P., KOWDLE A., IZADI S.: Articulated distance fields for ultra-fast tracking of hands interacting. *ACM Trans. Graph.* (2017). [1](#), [2](#)
- [WMB*20] WANG J., MUELLER F., BERNARD F., SORLI S., SOTNYCHENKO O., QIAN N., OTADUY M. A., CASAS D., THEOBALT C.: RGB2Hands: real-time tracking of 3d hand interactions from monocular rgb video. *ACM TOG* 39, 6 (2020), 1–16. [2](#), [7](#)
- [WRRW21] WEHRBEIN T., RUDOLPH M., ROSENHAHN B., WANDT B.: Probabilistic Monocular 3D Human Pose Estimation with Normalizing Flows. In *ICCV* (2021). [2](#)
- [WWHW19] WINKLER C., WORRALL D., HOOGEBOOM E., WELLING M.: Learning likelihoods with conditional normalizing flows. [3](#)
- [YK18] YE Q., KIM T.-K.: Occlusion-aware hand pose estimation using hierarchical mixture density network. In *ECCV* (September 2018). [2](#)
- [ZBJ*19] ZHOU Y., BARNES C., JINGWAN L., JIMEI Y., HAO L.: On the continuity of rotation representations in neural networks. In *CVPR* (June 2019). [3](#), [4](#)
- [ZBX*20] ZANFIR A., BAZAVAN E. G., XU H., FREEMAN W. T., SUKTHANKAR R., SMINCHISESCU C.: Weakly supervised 3d human pose and shape reconstruction with normalizing flows. In *ECCV* (2020). [2](#)
- [ZWD*21] ZHANG B., WANG Y., DENG X., ZHANG Y., TAN P., MA C., WANG H.: Interacting two-hand 3d pose and shape reconstruction from single color image. In *ICCV* (2021). [2](#)