







(Supplemental Materials) 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 

¹MPI Informatics & Saarland Informatics Campus, Germany

²Google Inc

³University of Bonn, Germany

⁴Universidad Rey Juan Carlos, Spain

⁵University of Freiburg, Germany

1. Network Training Settings

We use the PyTorch framework (v1.9) for our implementation. The ResNet-50 [HZRS16] backbone we used is pre-trained on the InterHand2.6M dataset using the weights from [MYW*20]. This ensures that the feature vector to the subsequent normalizing flow network contains relevant features for 3D hand pose estimation when starting to train the complete *HandFlowNet*.

For the optimization algorithm, we used AdamW, a version of Adam [KB14] with Decoupled Weight Decay Regularization [LH19]. Default parameters were used except for a learning rate of 10^{-4} and a weight decay of 10^{-4} .

For loss weights, we used $\lambda_{\text{Joint3D}} = 10^2$, $\lambda_{\text{Joint2D}} = 10^{-1}$, $\lambda_{\text{DetMag}} = 10^{-3}$, $\lambda_{\text{Param}} = 1.25 \times 10^{-3}$, $\lambda_{\text{NLL}} = 10^{-3}$, $\lambda_{\text{RotReg}} = 10^{-1}$.

2. MultiHands Dataset

In this section, we provide additional details of the algorithm we used to generate plausible annotations for our MultiHands dataset (See Algorithm 1 for overview.).

Overall, our method perturb the ground truth pose and check for the four plausibility criteria to generate new annotation. However, doing so naively (e.g. adding Gaussian noise to the pose space) would result mostly in samples that does not fit the plausibility criteria. Our method use several heuristics to speed up the discovery of plausible poses, which are explained in the following sections.

2.1. Translation Sampling:

We start by sampling the translation perturbations to the initial ground truth provided in InterHands2.6M. The goal is to constrain the range of plausible translation samples so that visible joints are *image consistent*, and resulting pose is *collision free*.

For this, we used binary search to find the *image consistent range* of each hand; the range of depth offsets that limits 2D position change under 3.5 pixels for visible joints.

We then calculate the *collision free range*; the range of valid left hand depth translations that avoid collision with the right hand. Collision is detected using sphere proxies obtained from the volumetric Gaussian approximation of the MANO model [MDB*19, WMB*20].

Given these ranges, we obtain the final translation change by first sampling a global depth offset from the overlapping *image consistent ranges* of both hands. The left hand is then offset from the right by sampling from the overlap between the *collision free range* and *image consistent range* of the left hand.

2.2. Articulation Sampling:

To find a plausible articulation, we observe that only occluded joints can change its position and the resulting position must also project on to an occluded pixel. Thus we iteratively select occluded joints and propose new occluded positions to cut down on the search space. As articulations are propagated down a kinematic chain, all joints on the same finger are considered together.

Select_finger(ψ_i): For each iterations, we first select a finger to perturb for the pose ψ_i . A finger can be selected if there exist an occluded joint whose child joints are all occluded.

For the selected finger, its joint locations $\{P_j^{3D}\}$ are updated from the base to the tip starting with the first occluded joint.

Sample_joint(P_j^{3D}): Given the current joint location, we want to find a new 3D position that both preserves the bone length and results in occlusion. To maintain the bone length, we define points on a 3D sphere centered at the parent joint with radius equal to the bone length as *bone length consistent points*.

We then try to eliminate points that would allow the joint to become visible. This is done by first projecting the sphere on to the image plane to obtain a set of pixels that lie within the projection, and then checking the occlusion status of each pixel based on a depth rendering. The occluded pixels locations are unprojected to form rays, and the intersections between the pixel rays and the sphere becomes the *preliminary joint proposals* Q^{3D} .

Update_pose(ψ'_i, Q^{3D}): For each *preliminary joint proposal*, the rotation needed to transform the current joint from the original position to the proposed location is calculated. This rotation update is used to update the current pose parameter ψ'_i .

Update_child($\{P^{3D}\}, \psi'_i$): The remaining finger joints $\{P_j^{3D}\}$ are updated using the pose parameter ψ'_i .

is_plausible(ψ'_i): After all child joints in a finger has been updated, we need to ensure the current pose parameter ψ'_i is anatomically plausible. This is done by converting the pose rotation parameters to the MANO pose PCA parameters. This allows us to estimate the pose likelihood under Gaussian assumptions. Hand proposals with low log likelihood (less than -60) are rejected.

To generate a single accepted new plausible annotation, we run pose perturbation for 100 iterations. For MultiHands, 100 new plausible annotations were obtained per image.

Algorithm 1: Pseudocode for sampling additional annotations

Data: Initial MANO Pose ψ_0
Result: Additional annotations ψ_i

```

// Ensure: sampled  $t \leftarrow \{t_{right}, t_{left}\}$ 
// 1.  $\Delta P^{2D}(\psi_i(t)) < T$ 
// 2.  $\neg \text{collision}(\psi_i(t))$ 
 $t \leftarrow \text{sample\_translation}(\psi_0)$ ;
 $\psi_i \leftarrow \text{update\_translation}(\psi_0, t)$ ;
for  $N$  iterations do
    // Get finger with occluded joints:
    //  $\forall P_j^{3D} \in \{P^{3D}\}$ 
    // child( $P_j^{3D}$ ) are occluded
     $\{P^{3D}\} \leftarrow \text{select\_finger}(\psi_i)$ ;
     $\psi'_i \leftarrow \psi_i$ 
    for  $P_j^{3D}$  in  $\{P^{3D}\}$  do
        // Ensure: sampled  $Q^{3D}$ 
        // 1.  $\text{bone}(Q^{3D}) = \text{bone}(P_j^{3D})$ 
        // 2.  $Q^{3D}$  is occluded
         $Q^{3D} \leftarrow \text{sample\_joint}(P_j^{3D})$ ;
         $\psi'_i = \text{update\_pose}(\psi'_i, Q^{3D})$ ;
         $\{P^{3D}\} = \text{update\_child}(\{P^{3D}\}, \psi'_i)$ ;
    end
    // likelihood from pose PCA
    if  $\text{is\_plausible}(\psi'_i)$  then
        |  $\psi_i \leftarrow \psi'_i$ 
    end
end

```

3. Evaluation Metrics

In this section, we provide additional details of the metrics used to evaluate our method.

3.1. Pose Alignments

Considering $\hat{P}^{3D} = \mathcal{J}(\psi)$ as the 3D joint positions calculated from our estimated hand parameters ψ , the **Global MPJPE (Global)** metric is defined as

$$\text{MPJPE}_{\text{Global}} = \frac{1}{N} \sum_{i=1}^N \|\hat{P}_i^{3D} - P_i^{3D}\|_2, \quad (1)$$

where N is the total number of annotated joints and P^{3D} are the ground-truth 3D joint positions. Note that this metric is computed *without any alignment*.

If we consider a *right root alignment*, the joints of both hands are aligned to the right hand root joint before computing the error. Let

$$\mathcal{R}_r(P_i^{3D}) = P_i^{3D} - P_{\text{right_root}}^{3D}, \quad (2)$$

be the function that calculates the joint position relative to the right hand root. Then the **Right-Root-Relative MPJPE (RRR)** metric is defined by

$$\text{MPJPE}_{\text{RRR}} = \frac{1}{N} \sum_{i=1}^N \|\mathcal{R}_r(\hat{P}_i^{3D}) - \mathcal{R}_r(P_i^{3D})\|_2. \quad (3)$$

We also evaluate the error for each hand individually. For this, we represent the joint position relative to the corresponding hand root joint by

$$\mathcal{R}(P_i^{3D}) = P_i^{3D} - \text{root}(P_i^{3D}), \quad (4)$$

where $\text{root}(\cdot)$ is a function that returns the right/left root joint position if P_i^{3D} belongs to the right/left hand. Then, the **Root-Relative MPJPE (RR)** is defined by:

$$\text{MPJPE}_{\text{RR}} = \frac{1}{N} \sum_{i=1}^N \|\mathcal{R}(\hat{P}_i^{3D}) - \mathcal{R}(P_i^{3D})\|_2. \quad (5)$$

3.2. Maximum Mean Discrepancy (MMD)

Given the set of all joint position of predicted pose samples $\hat{\mathcal{P}}^{3D} = \{\hat{P}_i^{3D}\}_{i=1}^n$, and the set of ground truth joint positions $\mathcal{P}^{3D} = \{P_j^{3D}\}_{j=1}^m$, we can estimate the Maximum Mean Discrepancy (MMD) with kernel κ with:

$$\begin{aligned} \text{MMD}^2(\hat{\mathcal{P}}^{3D}, \mathcal{P}^{3D}) &= \frac{1}{n(n-1)} \sum_{i \neq j}^n \kappa(\hat{P}_i^{3D}, \hat{P}_j^{3D}) \\ &+ \frac{1}{m(m-1)} \sum_{i \neq j}^m \kappa(P_i^{3D}, P_j^{3D}) \\ &- \frac{2}{nm} \sum_{j=1}^m \sum_{i=1}^n \kappa(\hat{P}_i^{3D}, P_j^{3D}) \end{aligned} \quad (6)$$

We used a Gaussian kernel for κ and averaged the MMD across $[1-100mm]$ distance range sampled at $1mm$ intervals.

MMD_{RR}^2 and MMD_{RRR}^2 , are similarly defined with $\mathcal{R}(P_i^{3D})$, $\mathcal{R}(\hat{P}_i^{3D})$ or with $\mathcal{R}_r(P_i^{3D})$, $\mathcal{R}_r(\hat{P}_i^{3D})$ respectively.

Method	Global ↓	RRR ↓	RR ↓
Ours	22.8	21.0	16.1
Ours (Mode)	51.4	30.7	18.2
InterNet [MYW*20]	83.3	29.1	19.4
Fan et al. [FSK*21]	81.7	32.1	17.2
InterShape* [ZWD*21]	-	33.7	18.7

Table 1: Our method produces samples that are on-par or better than the state-of-the-art methods. All results are in mm.

4. Experiment Results: Baseline Details

For implementing the baselines, we reuse existing HandFlowNet components as much as possible to ensure fair comparison in terms of network capacity. When using the normalizing flow network as a feed forward network in the baseline, we set $\mathbf{z} = \mathbf{0}$.

To implement MC-dropout, we use the existing dropout layers (with dropout probability of 0.5) in normalizing flow network during inference time to obtain samples. For the Gaussian baseline, we model the pose distribution as Gaussian aleatoric uncertainty $\mathcal{N}(\mu, \Sigma)$ inspired by Kendall *et al.* [KG17]. The normalizing flow network is trained to estimate μ and Σ directly from the extracted image feature \mathbf{v} . To implement the VAE baseline, we added a fully connected layer to the image feature extractor to act as the encoder for the image. The normalizing flow network then acts as the decoder to recover the hand pose. We empirically found that latent code size of 256 and KL divergence weight of 4×10^{-4} works best as hyper-parameters.

For the MC-dropout and VAE baselines, \mathcal{L}_{nll} and $\mathcal{L}_{\text{DetMag}}$ can not be used in their formulation and is thus omitted. Otherwise, all loss terms were used during training.

5. Experiment Results: Single Annotation

In sec.5.1.1, we shown that the commonly used MPJPE on a single annotation is not suitable for capturing the uncertainty present in the highly ambiguous task of monocular two-hands reconstruction. However, we still provide this comparison to the current state-of-the-art two-hand pose estimation methods for reference.

We compare our method to current : InterNet [MYW*20], InterShape [ZWD*21], and Fan et al [FSK*21]. Notice that InterNet and Fan et al. both require an additional network to explicitly estimate the global hand position, and InterShape only estimates relative hand position. In contrast, we directly output global hand position. Additionally, InterShape requires the ground-truth bone lengths to scale their results while our method does not.

Evaluation of Samples. Table 1 show the comparison on InterHand2.6M using MPJPE in mm. To evaluate whether our predicted distribution well captures the ground truth, we follow the established convention [YK18, WRRW21] to sample 100 poses and report the values of the *best sample* according to each metric.

We additionally report the metrics on just the mode sample to provide a baseline of our method as a traditional deterministic pose estimator. We see that our *HandFlowNet* produce samples that are

is significantly closer to the ground truth, while still being competitive even as a single pose estimator. As such, our method better captures the recoverable 3D information from the input.

6. Details on the Tzionas Dataset

The Tzionas dataset [TBS*16] has 1,307 RGB images across 7 sequences captured from a single view. Of these images, 264 has 2D joint annotations. For each hand, 14 joints annotation were given (with no annotations on the 5 fingertips, the wrist, or the carpometacarpal (CMC) joint of the thumb). The provided annotations only exist for joints that are visible. Thus we can apply our 2D visible joint loss even without MANO annotations.

7. More Results

In Figure 1, we show more renderings of our individual samples. In Figure 2, we use the skeleton visualization to show the spread of the predicted distribution. In Figure 3, we show annotations from our MultiHands dataset.

References

- [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). 3
- [HZRS16] HE K., ZHANG X., REN S., SUN J.: Deep residual learning for image recognition. In *CVPR* (June 2016). 1
- [KB14] KINGMA D. P., BA J.: Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014). 1
- [KG17] KENDALL A., GAL Y.: What uncertainties do we need in bayesian deep learning for computer vision? In *NIPS* (2017). 3
- [LH19] LOSHCHILOV I., HUTTER F.: Decoupled weight decay regularization. In *ICLR* (2019). 1
- [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
- [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). 1, 3, 5
- [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). 3, 5
- [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. 1
- [WRRW21] WEHRBEIN T., RUDOLPH M., ROSENHAHN B., WANDT B.: Probabilistic Monocular 3D Human Pose Estimation with Normalizing Flows. In *ICCV* (2021). 3
- [YK18] YE Q., KIM T.-K.: Occlusion-aware hand pose estimation using hierarchical mixture density network. In *ECCV* (September 2018). 3
- [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). 3

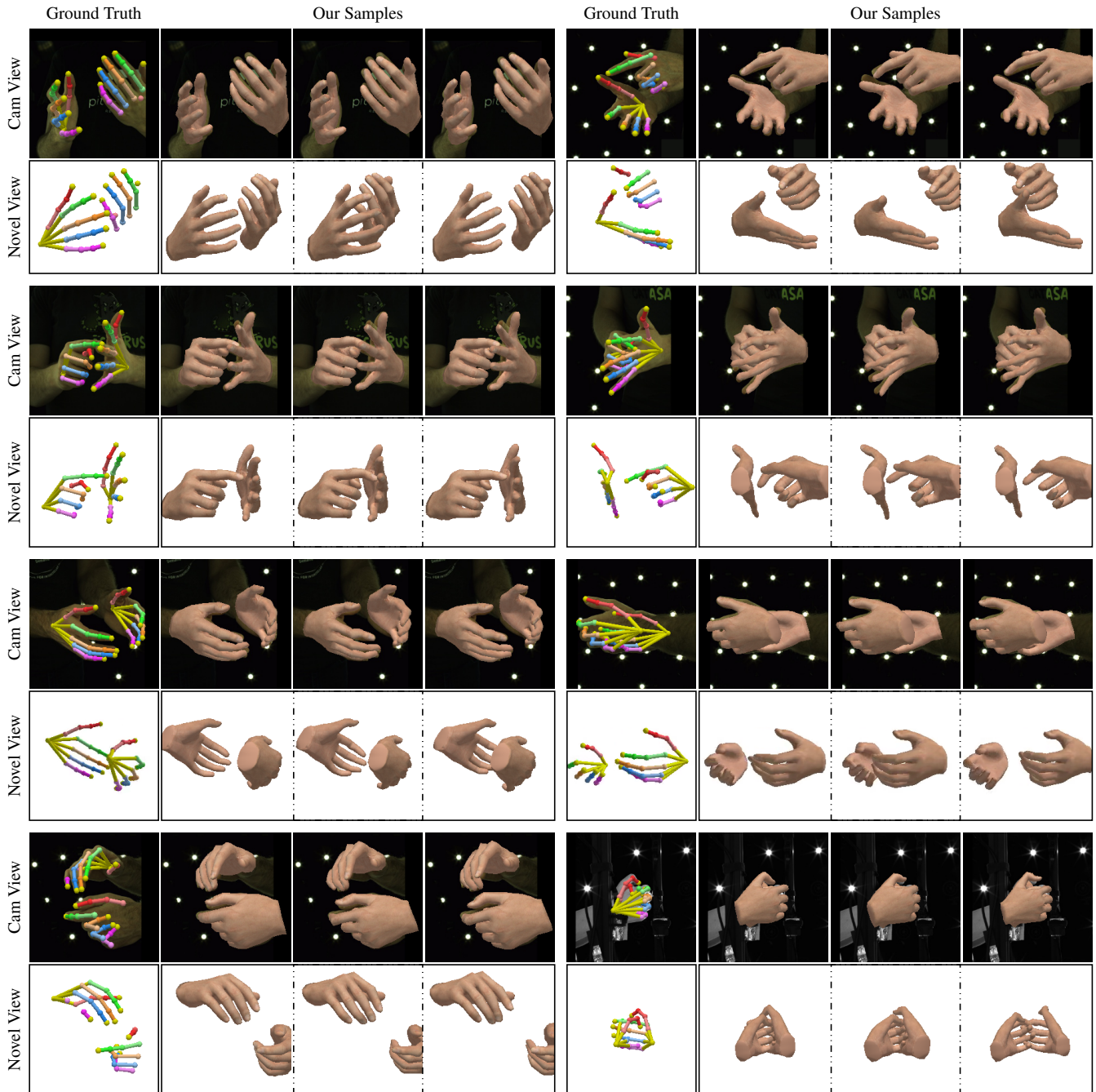


Figure 1: We show more individual mesh samples from the camera view and from a novel view. Note that not all joints are annotated in the ground truth. This shows up as missing segments in the skeleton.

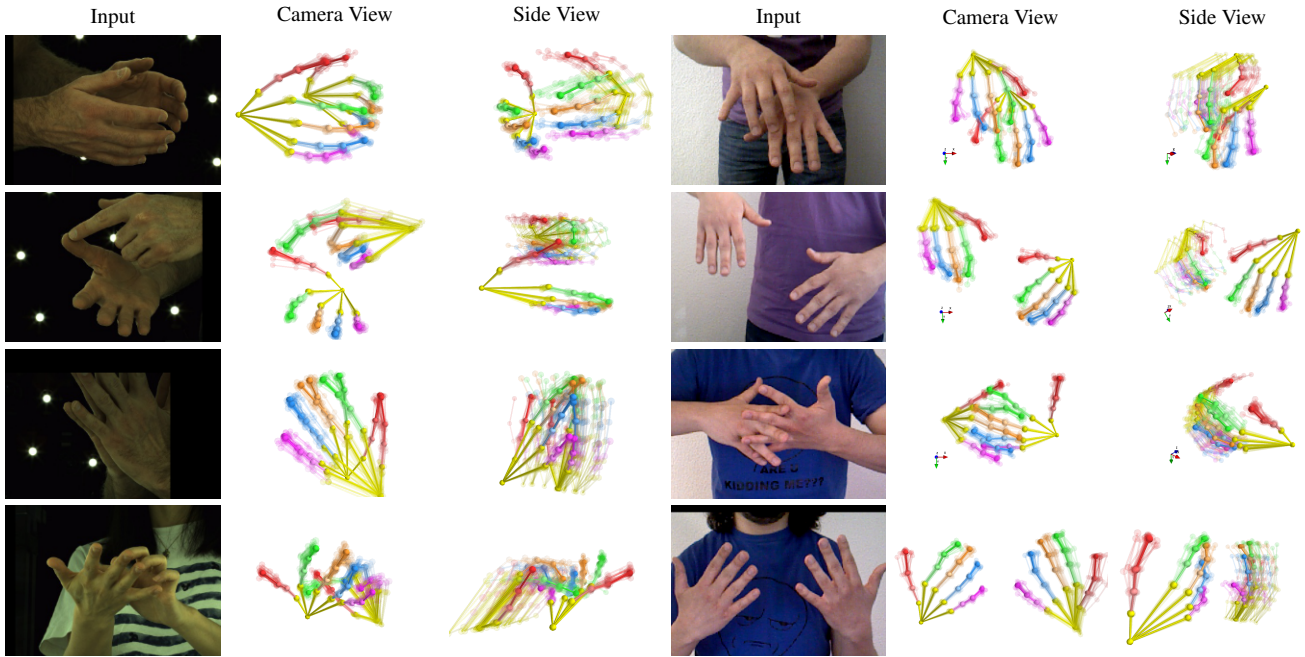


Figure 2: We show 30 samples from the estimated distribution, rendered as semi-transparent skeletons, superimposed on a single image. These Samples are aligned to the root joint of one hand and the mode of the distribution is made opaque for ease of visualization. Here we show more results from both the InterHand2.6M [MYW* 20] and the Tzionas [TBS* 16] datasets.



Figure 3: More visualizations of the our MultiHands dataset. Note the diversity of 3D poses that can be seen in the novel view.