Weakly-Supervised 3D Hand Reconstruction with Knowledge Prior and Uncertainty Guidance

Yufei Zhang[1], Jeffrey O. Kephart[2], and Qiang Ji[1]

[1] Rensselaer Polytechnic Institute, [2] IBM Research
{zhangy76, jiq}@rpi.edu, kephart@us.ibm.com

Abstract. Fully-supervised monocular 3D hand reconstruction is often difficult because capturing the requisite 3D data entails deploying specialized equipment in a controlled environment. We introduce a weakly-supervised method that avoids such requirements by leveraging fundamental principles well-established in the understanding of the human hand's unique structure and functionality. Specifically, we systematically study hand knowledge from different sources, including biomechanics, functional anatomy, and physics. We effectively incorporate these valuable foundational insights into 3D hand reconstruction models through an appropriate set of differentiable training losses. This enables training solely with readily-obtainable 2D hand landmark annotations and eliminates the need for expensive 3D supervision. Moreover, we explicitly model the uncertainty that is inherent in image observations. We enhance the training process by exploiting a simple yet effective Negative Log-Likelihood (NLL) loss that incorporates uncertainty into the loss function. Through extensive experiments, we demonstrate that our method significantly outperforms state-of-the-art weakly-supervised methods. For example, our method achieves nearly a 21% performance improvement on the widely adopted FreiHAND dataset.

Keywords: Monocular 3D Hand Reconstruction · Weakly-Supervised Learning · Universal Hand Prior · Maximum Likelihood Estimation

1 Introduction

Reconstructing the 3D configuration of human hands has broad applications, especially for Virtual/Augmented Reality (VR/AR) [3, 22] and Human-Computer Interaction (HCI) [53,65]. Traditional approaches rely on depth sensors [4,19,45, 49,51,52,56,76,77] or multi-camera setups [6,66,79]. Due to their reliance on specialized equipment that is often expensive or unavailable, the practicality of such approaches is limited. We instead focus on monocular 3D hand reconstruction, reconstructing 3D hands from a single RGB image.

Due to the lack of depth information in recovering 3D geometry from its 2D observation, monocular 3D hand reconstruction poses an ill-posed problem. Recent methods tackle this issue using deep learning models that predict 3D hand joint positions [16, 33, 43, 64, 64, 71, 72, 86] or reconstruct a dense 3D hand

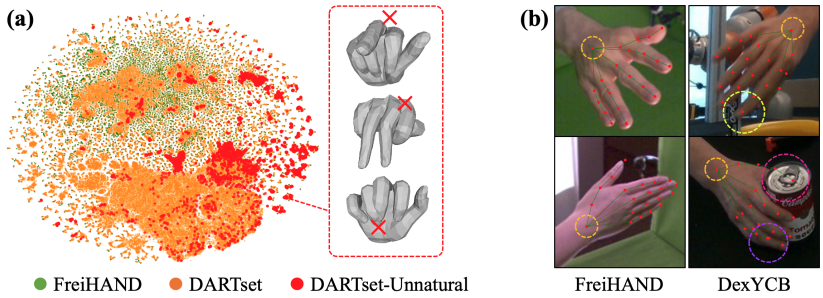

Fig. 1: Motivation of Studying Hand knowledge and Modeling Uncertainty. (a) T-SNE visualization [42] of hand poses from a real dataset (FreiHAND [87]) and a synthetic dataset (DARTset [17]). A large portion of synthetically generated hand poses can be unnatural (DARTset-Unnatural), such as the presence of invalid bending or penetration (marked by red crosses). (b) Images and 2D hand label from existing hand datasets. We mark regions with high uncertainty attributed to self-similarity (orange), motion blur (yellow), occlusion (pink), or poor image quality (purple).

mesh [13, 25, 39, 41, 44, 46, 62, 85]. While these methods avoid the need for specialized equipment in constrained environments at inference time, they still rely upon it to obtain the 3D annotations required for *training* the deep models. The resulting limitations in the diversity and amount of data restrict the performance of these purely data-driven deep models. To address this challenge, some methods [17, 37, 47, 48] leverage synthetically generated training images. The synthetic data are rich in quantity, but limited in the realism of the images and hand poses. As illustrated in Fig. 1(a), many synthetically generated poses in DARTset appear unnatural due to a lack of systematic consideration of well-established principles of hand structure, functionality and movement during the data generation process. Additionally, approaches based on generating synthetic data still require some real 3D data for further model fine-tuning.

Other authors [5,12,35,63] have exploited weakly-supervised learning, whereby the models are trained on real images with 2D hand landmark annotations. The advantage of such approaches is that 2D hand landmark labels are much more readily acquired in practice than 3D annotations. Weakly-supervised 3D hand models are typically trained by minimizing two loss terms: (1) a prior term imposed on 3D hand prediction to encourage its realism under weak supervision, and (2) a data term measuring the consistency between the projection of 3D prediction and 2D image observations. When constructing the prior term, some methods [2, 5] learn the prior from data. However, there is no sufficiently homogeneous and dense data set that precisely captures realistic hand movement patterns, and moreover it is a significant challenge to acquire such data [28]. Other works [12, 63] attempt to derive the prior from hand literature, but they are often limited to a certain type of knowledge. Another issue exhibited in existing weakly-supervised approaches lies in their formulation of the data term.

They overlook the uncertainty in image observations and employ standard regression losses, such as Mean Square Error (MSE). As shown in Fig. 1(b), various types of image ambiguities may be present, posing a significant challenge to the reconstruction process. Failing to address such inherent uncertainty may lead to degraded model performance [32].

In this paper, we address the two issues prevalent in current weakly-supervised 3D hand reconstruction models by (1) systematically and effectively leveraging well-established knowledge about the human hand, and (2) explicitly modeling the uncertainty inherent in input images. Our method draws inspiration from KNOWN [84], which leverages body-specific knowledge and uncertainty for human body reconstruction. Here we adapt that approach to the hand. Specifically, we extract from a comprehensive study of literature on hand biomechanics, functional anatomy, and physics a useful body of hand knowledge. We encode it as a set of differentiable losses to enable training on images solely with 2D weak supervision. Moreover, we consider that the observation uncertainty varies at different hand joints for different input images. We model such heteroscedastic uncertainty by capturing the distribution of 2D hand landmark positions. We improve the training by exploiting a simple yet effective Negative Log-Likelihood (NLL) loss that automatically assigns weights to different 2D labels based on their captured uncertainty. Through extensive experiments, we demonstrate the effectiveness of the proposed method and its significant improvements over the existing weakly-supervised 3D hand reconstruction models.

In summary, our main contributions lie in:

- identifying valuable generic knowledge from a comprehensive study of hand literature, including hand biomechanics, functional anatomy, and physics;
- introducing a set of differentiable training losses to effectively integrate the identified knowledge into 3D hand reconstruction models;
- exploiting a simple yet effective NLL loss that incorporates the uncertainty in image observations to improve the training; and
- showing through extensive experiments that our method significantly outperforms existing methods under the challenging weakly-supervised setting.

2 Related Work

In this section, we discuss recent advancements in monocular 3D hand reconstruction, considering fully-supervised and weakly-supervised settings.

2.1 Fully-Supervised Approaches

Fully-supervised 3D hand reconstruction requires that 3D labels, such as ground truth 3D hand meshes, are sufficiently available. They focus on designing different model architectures for improved performance. One line of work follows a model-based reconstruction pipeline, wherein a 3D hand is represented by a deformable 3D hand model and reconstructed by estimating low-dimensional

pose and shape parameters of the hand model [1, 81]. These model-based approaches can struggle to capture fine reconstruction details. Another line of work exploits a model-free reconstruction pipeline that directly predicts 3D hand mesh vertex positions [10, 11, 13, 20, 39, 40, 46, 55]. Such model-free approaches are typically data-hungry and less robust to occlusions and truncations. To address the issues inherent in both approaches, recent works [30, 75] propose unifying the two pipelines into a single framework to enhance overall performance. Additionally, some models are specifically designed for handling cases like occlusion [54], hand-object interaction [26, 67, 73] or two-hand reconstruction [38, 58, 70, 74, 78, 88]. While such innovations improve estimation accuracy, none of them address the significant challenge of acquiring a sufficient amount of 3D data for fully-supervised learning.

2.2 Weakly-Supervised Approaches

Weak supervision approaches have made significant progress in enhancing the generalization and data efficiency of 3D reconstruction models [31, 84]. In the context of 3D hand reconstruction, 2D hand landmark annotation proves to be a valuable form of weak supervision given its wide accessibility and the structural information it captures. Early works [5, 9, 81] relied on Principle Component Analysis (PCA) pose bases of the MANO hand model [59] and encouraged plausible 3D prediction by regularizing the prediction to be closer to the mean pose. Some works [18,81] impose geometry constraints that assumed finger joints were located in the same plane during movement. Baek *et al.* [2] propose capturing the complex 3D hand pose data distribution via Generative Adversarial Networks [21] and utilize the trained generative model as guidance for predicting realistic outputs. Instead of relying on data-driven priors or heuristic constraints, other works [12, 35, 48, 57, 63, 68] impose joint rotation constraints with ranges retrieved from hand biomechanics literature and achieve improved performance. However, they overlook other sources of useful hand knowledge. Moreover, Tzionas *et al.* [69] propose preventing invalid penetration in reconstructions by utilizing a non-penetration loss formulated over colliding mesh triangles. The proposed non-penetration loss only handles shallow penetration and cannot accommodate soft deformations that often occur in hand contact.

The contributions that differentiate our method from existing works are as follows. First, our study and utilization of generic hand knowledge is more comprehensive, and includes a novel inter-dependency derived from hand functional anatomy. Second, our encoding of knowledge is more effective. In particular, our formulation of the non-penetration loss effectively handles soft surface deformations by accurately pulling out deep inside vertices. Third, unlike existing works that neglect the heteroscedastic uncertainty in input images or limit their uncertainty modeling to hypothesis generation [9], our method explicitly models the uncertainty and incorporates it into the training loss through a simple yet effective NLL loss, directly improving the training process. While this strategy has been studied in other applications [14, 15, 36, 80], we are the first to apply it to monocular 3D hand reconstruction.

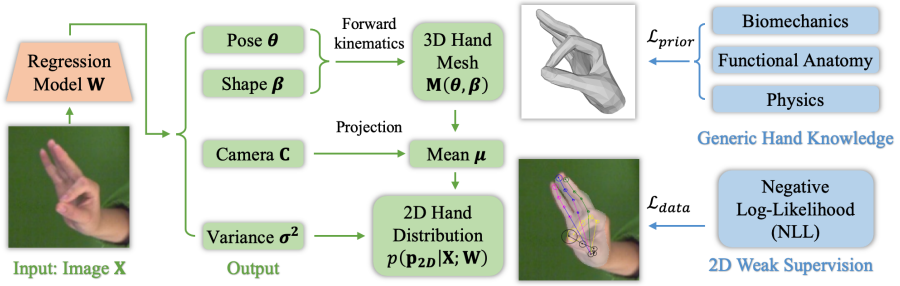

Fig. 2: Overview of the proposed method. Given a hand image, the regression model predicts the 3D hand pose and shape for recovering the 3D hand mesh through forward kinematics. The distribution of 2D hand landmark positions is specified via the projection of 3D hand and the predicted variance. The model is trained by incorporating generic hand knowledge and utilizing 2D hand landmark annotations.

3 Method

Fig. 2 overviews our proposed method. We begin by introducing our 3D hand representation and camera projection model in Sec. 3.1. Then, we systematically survey valuable hand knowledge and describe how we encode it as differentiable model training losses in Sec. 3.2. We discuss the modeled distribution of 2D hand landmark positions and our formulation of the NLL loss in Sec. 3.3. Finally, we summarize the overall training loss for our model in Sec. 3.4.

3.1 Preliminaries

3D Hand Representation. We employ MANO [59] to represent a 3D hand. MANO is a deformable 778-vertex 3D mesh model. It is parameterized by pose parameters $\boldsymbol{\theta} \in \mathbb{R}^{15\times3}$ that govern the rotation of 15 hand joints, and shape parameters $\boldsymbol{\beta} \in \mathbb{R}^{10}$ that represent the coefficients of PCA shape bases, capturing variations like hand length and width. Given $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$, 3D mesh vertices $\mathbf{M}(\boldsymbol{\theta}, \boldsymbol{\beta}) \in \mathbb{R}^{778\times3}$ are obtained through forward kinematics. 3D hand joints $\mathbf{P}(\boldsymbol{\theta}, \boldsymbol{\beta}) \in \mathbb{R}^{J\times3}$ are a linear combination of the vertices as $\mathbf{P}(\boldsymbol{\theta}, \boldsymbol{\beta}) = \mathbf{H}\mathbf{M}(\boldsymbol{\theta}, \boldsymbol{\beta})$, where $\mathbf{H} \in \mathbb{R}^{J\times778}$ is a joint regressor learned from data during the development of MANO, and $J = 21$ indicates the number of modeled hand joints.

Camera Projection Model. Similar to existing practices, we estimate camera parameters $\mathbf{C} = [s, \mathbf{R}, \mathbf{t}]$, where $s \in \mathbb{R}$, $\mathbf{R} \in \mathbb{R}^3$, and $\mathbf{t} \in \mathbb{R}^2$ denote the scale factor, camera rotation, and global translation, respectively. The projection of 3D hand joints is obtained as $\mathbf{p}_{2D} = Proj(\mathbf{P}; \mathbf{C})$, where $Proj(\cdot)$ denotes the full-perspective projection function with a constant focal length, as in [31].

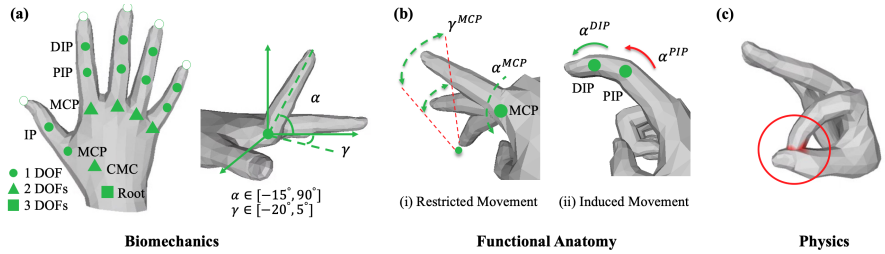

Fig. 3: Illustration of Generic Hand Knowledge. (a) Different hand joints have different degrees of freedom (DOFs) and ranges of motion [23]. (b) For the four fingers, (i) mutual restrictions exist between joint bending (α) and splaying (γ) of the MCPs (metacarpophalangeal joints); (ii) the bending of the DIP (distal interphalangeal joint) induces bending in the PIP (proximal interphalangeal joint) due to tighter ligaments [60]. (c) Different hand digits are prevented from penetrating into each other.

3.2 Study and Incorporation of Generic Hand Knowledge

Hand movement adheres to fundamental principles applicable across different subjects and gestures, serving as foundational insights for realistic 3D hands reconstruction. In this section, we systematically survey the generic hand knowledge from various sources, including hand biomechanics, functional anatomy, and physics. We introduce a set of differentiable losses over the 3D hand pose and shape parameters to integrate the knowledge into the reconstruction model.

Hand Biomechanics involves the quantitative study of hand movement mechanisms. There are 15 hand joints contributing to movement: a metacarpophalangeal joint (MCP), a proximal interphalangeal joint (PIP), and a distal interphalangeal joint (DIP) for each of the four fingers, and a carpometacarpal joint (CMC), an MCP, and an IP for the thumb. Each joint's movement can be described via three Euler angles corresponding to joint bending, splaying, and twisting, respectively. Hand biomechanics studies specify the DOFs and ranges of motion for each joint as illustrated in Fig. 3(a). To impose these constraints, we introduce the following pose loss:

$$\mathcal{L}_{pose} = \sum_{j=1}^{15} (\max\{\boldsymbol{\theta}_j - \bar{\boldsymbol{\theta}}_{j,max}, \bar{\boldsymbol{\theta}}_{j,min} - \boldsymbol{\theta}_j, \mathbf{0}\})^2, \tag{1}$$

where $\boldsymbol{\theta}_j$ represents the three Euler angles predicted for the j^{th} joint. Their range, denoted by $(\bar{\boldsymbol{\theta}}_{j,min}, \bar{\boldsymbol{\theta}}_{j,max})$, is obtained from literature [23], with the ranges set to zero for directions without degrees of freedom. Since the joint rotation coordinates used by MANO differ from those defined above, we adjust MANO's original coordinates by aligning its movement axes with the three Euler angles defined above. We also design the Euler angle rotation order for a joint based on their rotation range to avoid singularity, following [82–84].

Hand Functional Anatomy investigates how the hand's anatomical structure influences its movement. In contrast to hand biomechanics, which delineates the

range of motion for each individual joint, the study of functional anatomy stipulates essential inter-joint dependencies during hand movement. As illustrated in Fig. 3(b), there are two types of dependencies: (i) the bending of the MCP restricts its splaying, following a linear relationship that peaks at the maximum bending angle [61]; and (ii) the bending of the DIP induces bending of the PIP within the same finger [60]. These inter-dependencies highlight that the range of motion of the hand joints can be dynamic and dependent on each other. Specifically, denote the predicted bending and splaying angles of an MCP joint j as α_j^{MCP} and γ_j^{MCP}, respectively. Based on the Type-(i) dependency, their range of motion should be updated as:

$$\begin{aligned}\hat{\gamma}_{j,min}^{MCP} &= \bar{\gamma}_{j,min}^{MCP}(1 - \frac{\alpha_j^{MCP}}{\bar{\alpha}_{j,min}^{MCP}}),\ \text{if } \bar{\alpha}_{j,min}^{MCP} < \alpha_j^{MCP} < 0, \\ \hat{\gamma}_{j,max}^{MCP} &= \bar{\gamma}_{j,max}^{MCP}(1 - \frac{\alpha_j^{MCP}}{\bar{\alpha}_{j,max}^{MCP}}),\ \text{if } 0 < \alpha_j^{MCP} < \bar{\alpha}_{j,max}^{MCP},\end{aligned} \tag{2}$$

where $(\bar{\gamma}_{j,min}^{MCP}, \bar{\gamma}_{j,max}^{MCP})$ and $(\bar{\alpha}_{j,min}^{MCP}, \bar{\alpha}_{j,max}^{MCP})$ are the ranges based on hand biomechanics, while $(\hat{\gamma}_{j,min}^{MCP}, \hat{\gamma}_{j,max}^{MCP})$ denotes the refined range. As shown, the range of γ_j^{MCP} becomes very limited as α_j^{MCP} approaches extreme angles. Similarly, the range of motion for α_j^{MCP} needs to be further constrained based on the value of γ_j^{MCP}. Moreover, denote the predicted bending of the PIP and DIP of finger k as α_k^{PIP} and α_k^{DIP}, respectively. According to the Type-(ii) dependency, the lower bound of α_k^{PIP} should be refined as:

$$\hat{\alpha}_{k,min}^{PIP} = 0,\ \text{if } \alpha_k^{DIP} > 0, \tag{3}$$

where α_k^{PIP} is encouraged to be greater than zero given a flexion DIP. In summary, the two types of dependencies refine the ranges ($\bar{\boldsymbol{\theta}}_{min}$, $\bar{\boldsymbol{\theta}}_{max}$) provided by the hand biomechanics to ($\hat{\boldsymbol{\theta}}_{min}$, $\hat{\boldsymbol{\theta}}_{max}$) based on the current hand pose prediction $\boldsymbol{\theta}$. To integrate this valuable anatomical knowledge into the 3D reconstruction model, we dynamically update the joint rotation ranges following Eq. 2 and Eq. 3, and utilize the refined ranges to calculate the pose loss in Eq. 1.

Hand Physics studies assert various principles governing the physical interactions of the human hand. As our model reconstructs a single 3D hand from a single image, we mainly consider static physics, particularly the principle of non-penetration, according to which different hand parts cannot penetrate into each other. Fig. 3(c) illustrates a failure case. To integrate the non-penetration principle into the 3D reconstruction model, we first identify a set $\mathbf{M}$ comprising vertices located inside the mesh through the generalized winding number [29,50]. For each vertex $\mathbf{v} \in \mathbf{M}$, we then apply the following non-penetration loss:

$$\mathcal{L}_{non-penetration} = \sum_{\mathbf{v} \in \mathbf{M}} \max\{d(\mathbf{v}) - d_{tol}, 0\}, \tag{4}$$

where $d(\mathbf{v})$ denotes the minimum distance from vertex $\mathbf{v}$ to another vertex that is not a neighbor of $\mathbf{v}$ (where the geodesic distance exceeds the average length

of phalanges, e.g., 2cm). In other words, $d(\mathbf{v})$ represents the minimum distance from vertex $\mathbf{v}$ to 3D hand surface. Meanwhile, recognizing MANO's limitation in modeling soft surface deformations during contact, we introduce a tolerance distance d_{tol} to accommodate shallow penetrations. Unlike existing methods [69] that formulate the loss based on collision triangles and only deter shallow penetrations, our proposed loss is applied to vertices with distances to the surface exceeding d_{tol}, effectively pulling out the those deeply embedded vertices.

Overall Knowledge-Encoded Prior. Incorporating the knowledge discussed above ensures natural 3D hand pose predictions. Similar to existing methods [12, 35], we apply a shape regularization $\mathcal{L}_{shape} = \|\boldsymbol{\beta}\|_2$ to promote plausible hand shape predictions. Assembling all the losses together, we obtain the overall prior:

$$\mathcal{L}_{prior} = \lambda_1 \mathcal{L}_{pose} + \lambda_2 \mathcal{L}_{non-penetration} + \lambda_3 \mathcal{L}_{shape} \tag{5}$$

It is worth noting that the prior term in Eq. 5 is derived from generic hand knowledge, which is applicable to all subjects and gestures. Notably, its formulation does not require any 3D data and is independent of any specific dataset.

3.3 Training with Negative Log-Likelihood

To further ensure that predictions are consistent with the image observations, we utilize 2D hand landmark annotations. The input images can often exhibit challenges, such as occlusion or low image quality, that result in inherently ambiguous 2D hand positions or high uncertainty in the 3D reconstruction. Unlike existing methods that overlook this inherent uncertainty and train on 2D hand labels using standard regression loss, we explicitly model the uncertainty and incorporate it into the loss function to enhance model performance. Specifically, we model the uncertainty by capturing the distribution of 2D hand landmark positions. As different 2D hand landmarks exhibit different appearance features that vary across input images, we model each joint independently and capture input-dependent uncertainty. We assume the distribution of 2D hand landmark positions $\mathbf{p}_{2D}$ of an image $\mathbf{X}$ as:

$$p(\mathbf{p}_{2D}|\mathbf{X};\mathbf{W}) = \prod_i \frac{1}{\sqrt{2\pi}\boldsymbol{\sigma}_i} \exp\left(-\frac{(\mathbf{p}_{2D,i} - \boldsymbol{\mu}_i)^2}{2\boldsymbol{\sigma}_i^2}\right), \tag{6}$$

where i is the image location index of hand joints. The adoption of Gaussian distributions is based on their wide utility in modeling observation noise [32]. $\boldsymbol{\mu}$ represents the mean of the Gaussian distributions computed through the projection of 3D hand joint positions $\mathbf{P}$ using the camera parameters $\mathbf{C}$, while the variance $\boldsymbol{\sigma}^2$ are directly predicted by the regression model with parameters $\mathbf{W}$.

The modeled distribution $p(\mathbf{p}_{2D}|\mathbf{X};\mathbf{W})$ specifies the probability of the ground truth appearing at position $\mathbf{p}_{2D}$. The labeled position $\bar{\mathbf{p}}_{2D}$ can be viewed as an observed data sample. We can thus train the model through Maximum Likelihood Estimation. It minimizes the Negative Log-Likelihood (NLL) to construct

the data term to ensure 3D-2D consistency as:

$$\begin{aligned}\mathcal{L}_{data} &= -\log p(\mathbf{p}_{2D} = \bar{\mathbf{p}}_{2D}|\mathbf{X};\mathbf{W}) \\ &\propto \sum_i \left(\log \boldsymbol{\sigma}_i + \frac{(\bar{\mathbf{p}}_{2D,i} - \boldsymbol{\mu}_i)^2}{2\boldsymbol{\sigma}_i^2}\right).\end{aligned} \tag{7}$$

Note that the variance $\boldsymbol{\sigma}^2$ in Eq. 7 depends on the individual hand joint i. Omitting the variance estimation or treating it as a constant would be equivalent to using the standard MSE loss, which is agnostic to uncertainty and assigns weights to all samples uniformly. In contrast, our method assigns reduced weights to images and joints with high uncertainty in a principled fashion, thereby producing a more robust model with improved performance.

3.4 Total Training Loss

By combining the prior term in Eq. 5 and the data term in Eq. 7, we obtain the total loss for training the regression model as:

$$\mathcal{L} = \mathcal{L}_{prior} + \mathcal{L}_{data}. \tag{8}$$

During testing, 3D hands can be directly reconstructed through the hand pose and shape parameters estimated by the regression model.

4 Experiment

We briefly introduce our data sets, evaluation metrics, and implementation details in Sec. 4.1. Then, in Sec. 4.2, we discuss an ablation study that demonstrates the effectiveness of incorporating various sources of generic hand knowledge and training with the Negative Log-Likelihood (NLL). Finally, in Sec. 4.3, we assess the improved performance of our method in comparison to existing weakly-supervised State-of-the-Art (SOTA) approaches.

4.1 Datasets, Metrics, and Implementation Details

Datasets. We employ three widely adopted datasets: FreiHAND [87], DexYCB [8], and HO3Dv3 [24], all of which have been captured by multi-view data collection systems. FreiHAND features a diverse range of daily hand poses. DexYCB and HO3Dv3 contain hand-object interaction images, some of which are significantly occluded. We follow the established training and testing splits to facilitate comparison with other methods.

Evaluation Metrics. Like existing methods [12, 30], we compute the average Euclidean distance between the predicted and the ground truth 3D hand joint and vertex positions after procrustes alignment ($\mathrm{E}_J/\mathrm{E}_V$). The evaluation on HO3Dv3 is obtained through the online submission system. It further includes $\mathrm{AUC}_J/\mathrm{AUC}_V$, the area under the percentage of correct keypoint (PCK) curves

with thresholds between 0mm and 50mm. Additionally, we compute penetration rate (PR), the percentage of reconstructions exhibiting penetration with a depth greater than d_{tol}, to assess the physical plausibility of 3D reconstructions.

Implementation. We implemented our framework using PyTorch. The regression model consists of a ResNet-50 model [27] to extract image features and an iterative error feedback regression model [7] to predict the unknown parameters from the extracted features. The hand images are scaled to 224×224 while preserving the aspect ratio. The training images are augmented with random scaling and flipping. The training batch size and epochs is 64 and 200, respectively. Following the training strategy in [84], we initially employ the MSE for the data term and then utilize the NLL for faster convergence. We use the Adam optimizer [34] with a learning rate of 10^{-5} and weight decay of 10^{-4}. The hyperparameters are set to $d_{tol} = 6mm$, $\lambda_1 = 20000$, $\lambda_2 = 20000$, and $\lambda_3 = 10$.

4.2 Ablation Study

Table 1 summarizes the impact of a) incorporating hand knowledge and b) training with the NLL loss. We provide a detailed analysis of these results below.

Incorporating Generic Hand Knowledge. To provide valuable insights about the effectiveness of leveraging different sources of hand knowledge, we supplement Table 1 with a qualitative evaluation in Fig. 4. When not integrating any hand knowledge, the model is trained using 2D hand landmark annotations with a prior term that only includes the shape regularization. This model produces large reconstruction errors (Table 1, row1). As illustrated in Fig. 4 ("No Knowledge"), the reconstructions can align with the image observations, but the predicted 3D hand poses are fairly unrealistic. The infeasible twisted fingers significantly violate the joint range of motion specified by hand biomechanics. This issue is addressed by introducing hand biomechanics into the training, resulting in a significant model performance boost (Table 1, row2 over row1). For example, E_J is improved from 22.4mm to 10.9mm. Meanwhile, the estimated 3D hand poses become more plausible as shown in Fig. 4 ("+Biomechanics"). Nonetheless, poor reconstructions can still occur due to the inherent depth ambiguity. Specifically, the relative depth of hand joints can be incorrect. Mitigating this issue requires the further incorporation of the functional anatomy knowledge (Fig. 4, "++F-Anatomy"), leading to a reduction of the reconstruction errors from 10.9mm to 9.6mm for E_J and from 11.4mm to 10.0mm for E_V (Table 1, row3 over row2). The functional anatomy captures inter-joint dependency, alleviating the depth ambiguity by introducing additional constraints on the 3D reconstruction space. Furthermore. avoiding invalid penetrations in the reconstructions requires adding the proposed non-penetration loss (Fig. 4, bottom example). Incorporating this physics knowledge effectively reduces the percentage of reconstructions with invalid penetration from 11.4% to 1.9%. To a lesser degree, but still significantly, it also improves the other reconstruction accuracy metrics. In summary, by incorporating hand knowledge gleaned from literature into our proposed training loss functions, we are able to generate accurate 3D hand reconstruction models based solely on 2D weak supervision.

Table 1: Quantitative Evaluation of Incorporating Hand Knowledge and Training into the NLL Loss. The evaluation is on FreiHAND. Without incorporating any hand knowledge, the model is trained by utilizing 2D hand landmark annotations and the shape regularization. "F-Anatomy" denotes "Functional Anatomy". The units of E_J and E_V are in mm, while PR is in percentage.

Hand Knowledge			Weak Supervision	Reconstruction		
Biomechanics	F-Anatomy	Physics	NLL	$E_J\downarrow$	$E_V\downarrow$	PR$\downarrow$
				22.4	24.8	38.8
✓				10.9	11.4	11.4
✓	✓			9.6	10.0	11.4
✓	✓	✓		9.4	9.8	1.9
✓	✓	✓	✓	**8.5**	**8.9**	**1.3**

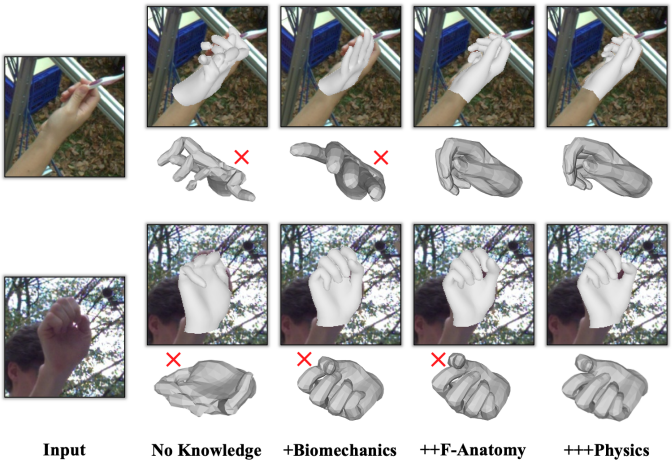

Fig. 4: Qualitative Evaluation of Incorporating Hand Knowledge. The images are from FreiHAND's test set. For each example, we present the rendered 3D hand overlaid on the input image, along with the reconstructed 3D hand viewed from a different angle. The results from left to right are obtained by incorporating the additional knowledge specified at the bottom. "F-Anatomy" denotes "Functional Anatomy". Reconstructions with notable errors are marked by red crosses.

Training with Negative Log-likelihood. As discussed in Sec. 3.3, the NLL loss takes into account the increased reconstruction uncertainty of images containing occlusions or other degradations at the granularity of individual hand joints. Comparing rows 5 and 4 of Table 1 we see that the 3D joint position error E_J is reduced from 9.4mm to 8.5mm and the 3D mesh reconstruction error E_V is decreased from 9.8mm to 8.9mm. In Fig. 5(a), we provide qualitative comparison

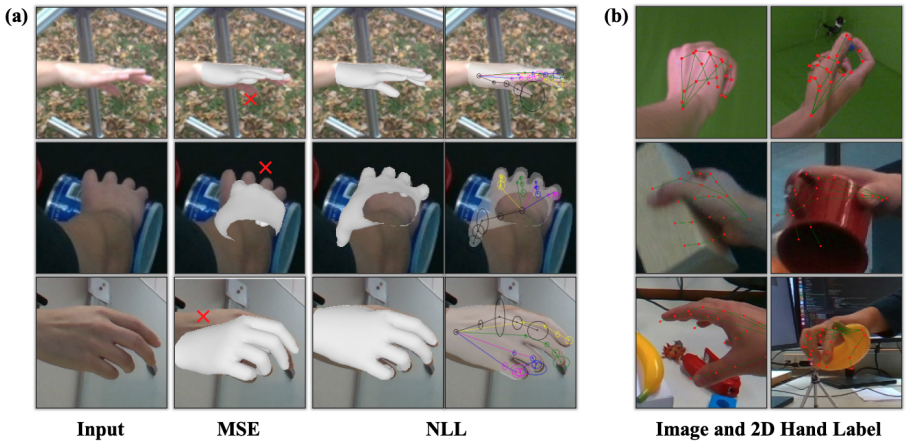

Fig. 5: Qualitative Evaluation of Training with the NLL. (a) Evaluation of the models trained without and with the NLL. Notable errors are marked by red crosses. Colors indicate finger identity: thumb (black), index (yellow), middle (green), ring (blue), and pinky (magenta). The width and height of the ellipse at each joint represent the magnitude of the estimated variance along the horizontal and vertical directions, respectively. (b) Training images with high uncertainty, captured by large estimated variances (averaged over all joints). The images in (a) and (b) are from FreiHAND (top), DexYCB (middle), and HO3D (bottom).

between the models trained without and with the NLL loss. When not utilizing the NLL, the model is trained using the MSE loss. As shown, such models can be adversely affected by low image quality, image occlusion, and image truncation occurring at various hand joints, resulting in poor 3D hand reconstructions. In contrast, the model trained with the NLL is more robust to these situations. For example, the alignment to the input image is significantly improved compared to training with the MSE even when the hand is heavily occluded (Fig. 5(a), middle example). Furthermore, the model trained with the NLL captures the distribution of 2D hand positions. As shown in Fig. 5(a) (column4), the hand joints in low-quality or occluded regions are captured by high variance estimates (visualized by large ellipses). In Fig. 5(b), we present training images with large estimated variances. As shown, the images with excessive occlusion, truncation, and ambiguity appearance exhibit large variance estimations. During training, the utilization of the NLL effectively enhances the final model performance by incorporating the uncertainty into the training loss function.

4.3 Comparison with State-of-the-Art

In this section, we showcase the enhanced performance of our approach compared to state-of-the-art (SOTA) methods in the challenging weakly-supervised setting. We summarize the quantitative evaluation on three different datasets:

Table 2: Comparison with Weakly-Supervised SOTA Methods. The evaluation of other methods are obtained from published papers. * denotes the methods using 3D annotations in synthetic or real 3D hand datasets during training. For the evaluation on DexYCB, we report the 3D mesh reconstruction error with the root translation alignment E_{RV} to compare with others. The units of E_J, E_V, and E_{RV} are in mm.

Method	FreiHAND		DexYCB		HO3D			
	$E_J\downarrow$	$E_V\downarrow$	$E_J\downarrow$	$E_{RV}\downarrow$	$E_J\downarrow$	$AUC_J\uparrow$	$E_V\downarrow$	$AUC_V\uparrow$
*Boukhayma *et al.* [5]	11.0	10.9	-	27.3	-	-	-	-
*Spurr *et al.* [63]	11.3	-	7.1	-	-	-	-	-
Chen *et al.* [12]	11.8	11.9	-	-	11.5	0.769	11.1	0.778
Ren *et al.* [57]	10.7	11.0	-	-	-	-	-	-
Jiang *et al.* [30]	10.8	10.9	-	-	10.5	0.789	10.7	0.785
Ours	**8.5**	**8.9**	**6.7**	**22.0**	**10.0**	**0.800**	**9.8**	**0.804**

FreiHAND, DexYCB, and HO3D in Table 2. Our method consistently outperforms existing approaches across all three datasets, which include diverse images depicting daily hand poses and hand-object interactions. Specifically, early methods using 2D weak supervision often require training with 3D annotations due to limited constraints on the 3D predictions [5,63]. Chen *et al.* [12] avoid the dependency on 3D data by employing different statistical regularizations during training, achieving performance comparable to that of methods utilizing 3D data. Ren *et al.* [57] further enhance the performance by leveraging feature consistency constraints. However, these methods are confined to heuristic constraints, such as enforcing a mean pose prediction, or partial types of hand knowledge, like hand biomechanics alone. In contrast, we systematically study and exploit generic hand knowledge, resulting in significant performance improvements. Notably, our improvements over these methods are achieved even without utilizing the NLL loss. For instance, on the FreiHAND dataset, our method achieves E_J of 9.4mm, reducing the second-best's 10.7mm by 12%. Further utilization of the NLL leads to a more significant error reduction of 21%. Additionally, Jiang *et al.* [30] propose a probabilistic framework to combine model-based and model-free reconstruction models. Despite their incorporation of additional models, our method outperforms them by a large margin. For example, E_V is decreased from 10.9mm to 8.9mm on FreiHAND, and from 10.7mm to 9.8mm on HO3D. Particularly, our method achieves the improved performance by effectively utilizing the generic hand knowledge and modeling the input uncertainty.

5 Discussion

In Sec. 4, we validated our method under the challenging weakly-supervised setting. Here, we demonstrate the advantages of leveraging generic hand knowledge even when 3D annotations are available. Table 3(a) shows that our method's

Table 3: Benefits of Utilizing Generic Hand Knowledge When 3D Annotation is Available. The units of E_J, E_V, and E_{RV} are in mm.

(a) Generalization. The evaluation is on DexYCB. The models are trained using 2D hand landmark annotation. "FreiHAND-3D" exploits 3D annotated images from FreiHAND to regularize the training on DexYCB, while "Ours" uses generic hand knowledge as the prior.

Prior	$E_J\downarrow$	$E_V\downarrow$
FreiHAND-3D	8.3	8.5
Ours	6.7	7.1

(b) Data Efficiency. The evaluation is on FreiHAND. The models are trained using 2D hand landmark annotation and incorporating different percentages of 3D annotation during training.

3D Annotation	$E_J\downarrow$	$E_V\downarrow$	$E_{RV}\downarrow$
100%	8.30	8.4	21.5
Ours (0%)	8.52	8.9	18.4
Ours+10%	8.26	8.6	16.9

performance compares favorably with that attained by using a data-driven prior extracted from FreiHAND (following [31]) and then evaluated on DexYCB. Thus our method's use of generic hand knowledge gives it a significant advantage over data-driven, domain-specific approaches. Furthermore, our method can take advantage of 3D annotations when they are available. As illustrated in Table 3(b), when not leveraging any 3D annotation, our method performs just slightly worse than the fully-supervised model ("100%") on 2 of the 3 metrics, and it performs comparably to the fully-supervised model using only 10% of the 3D annotations ("Ours+10%"). The advantages of generic hand knowledge, including its generalizability and its role in improving data efficiency of monocular 3D hand reconstruction models, further demonstrate its significance.

6 Conclusion

We comprehensively study generic hand knowledge, including hand biomechanics, functional anatomy, and physics. We effectively encode these foundational insights as differentiable prior losses, enabling the training of 3D hand reconstruction models solely using 2D annotation. Moreover, we explicitly model image uncertainty with a simple yet effective Negative Log-Likelihood (NLL) loss that incorporates the well-captured uncertainty into the training loss function. Our method significantly outperforms existing weakly-supervised methods. On the widely adopted FreiHAND dataset, the improvement is nearly 21%.

Society Impact. Our work highlights the importance of integrating hand knowledge and modeling uncertainty to produce reliable predictions, grounded in hand mechanics and with confidence estimates. It can potentially benefit many downstream tasks like synthetic data generation, biomechanics, and robotics.

Limitations & Future Work. Our method focuses on static generic hand knowledge for image-based reconstruction. A natural extension to our work would be to estimate hand dynamics from monocular videos.

Acknowledgement

This work is supported in part by IBM through the IBM-Rensselaer Future of Computing Research Collaboration.

References

1. Baek, S., Kim, K.I., Kim, T.K.: Pushing the envelope for rgb-based dense 3d hand pose estimation via neural rendering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1067–1076 (2019)
2. Baek, S., Kim, K.I., Kim, T.K.: Weakly-supervised domain adaptation via gan and mesh model for estimating 3d hand poses interacting objects. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6121–6131 (2020)
3. Bai, H., Sasikumar, P., Yang, J., Billinghurst, M.: A user study on mixed reality remote collaboration with eye gaze and hand gesture sharing. In: Proceedings of the 2020 CHI conference on human factors in computing systems. pp. 1–13 (2020)
4. Ballan, L., Taneja, A., Gall, J., Van Gool, L., Pollefeys, M.: Motion capture of hands in action using discriminative salient points. In: European Conference on Computer Vision. pp. 640–653. Springer (2012)
5. Boukhayma, A., Bem, R.d., Torr, P.H.: 3d hand shape and pose from images in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10843–10852 (2019)
6. de Campos, T.E., Murray, D.W.: Regression-based hand pose estimation from multiple cameras. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06). vol. 1, pp. 782–789. IEEE (2006)
7. Carreira, J., Agrawal, P., Fragkiadaki, K., Malik, J.: Human pose estimation with iterative error feedback. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4733–4742 (2016)
8. Chao, Y.W., Yang, W., Xiang, Y., Molchanov, P., Handa, A., Tremblay, J., Narang, Y.S., Van Wyk, K., Iqbal, U., Birchfield, S., et al.: Dexycb: A benchmark for capturing hand grasping of objects. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9044–9053 (2021)
9. Chen, R., Yang, L., Yao, A.: Mhentropy: Entropy meets multiple hypotheses for pose and shape recovery. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14840–14849 (2023)
10. Chen, X., Liu, Y., Dong, Y., Zhang, X., Ma, C., Xiong, Y., Zhang, Y., Guo, X.: Mobrecon: Mobile-friendly hand mesh reconstruction from monocular image. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20544–20554 (2022)
11. Chen, X., Liu, Y., Ma, C., Chang, J., Wang, H., Chen, T., Guo, X., Wan, P., Zheng, W.: Camera-space hand mesh recovery via semantic aggregation and adaptive 2d-1d registration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13274–13283 (2021)
12. Chen, Y., Tu, Z., Kang, D., Bao, L., Zhang, Y., Zhe, X., Chen, R., Yuan, J.: Model-based 3d hand reconstruction via self-supervised learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10451–10460 (2021)

13. Choi, H., Moon, G., Lee, K.M.: Pose2mesh: Graph convolutional network for 3d human pose and mesh recovery from a 2d human pose. In: European Conference on Computer Vision. pp. 769–787. Springer (2020)
14. Duan, R., Caffo, B., Bai, H.X., Sair, H.I., Jones, C.: Evidential uncertainty quantification: A variance-based perspective. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 2132–2141 (2024)
15. Dwivedi, S.K., Schmid, C., Yi, H., Black, M.J., Tzionas, D.: Poco: 3d pose and shape estimation with confidence. arXiv preprint arXiv:2308.12965 (2023)
16. Fan, Z., Spurr, A., Kocabas, M., Tang, S., Black, M.J., Hilliges, O.: Learning to disambiguate strongly interacting hands via probabilistic per-pixel part segmentation. In: 2021 International Conference on 3D Vision (3DV). pp. 1–10. IEEE (2021)
17. Gao, D., Xiu, Y., Li, K., Yang, L., Wang, F., Zhang, P., Zhang, B., Lu, C., Tan, P.: Dart: Articulated hand model with diverse accessories and rich textures. Advances in Neural Information Processing Systems **35**, 37055–37067 (2022)
18. Gao, D., Zhang, X., Chen, X., Tan, A., Zhang, B., Pan, P., Tan, P.: Cyclehand: Increasing 3d pose estimation ability on in-the-wild monocular image through cyclic flow. In: Proceedings of the 30th ACM International Conference on Multimedia. pp. 2452–2463 (2022)
19. Garcia-Hernando, G., Yuan, S., Baek, S., Kim, T.K.: First-person hand action benchmark with rgb-d videos and 3d hand pose annotations. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 409–419 (2018)
20. Ge, L., Ren, Z., Li, Y., Xue, Z., Wang, Y., Cai, J., Yuan, J.: 3d hand shape and pose estimation from a single rgb image. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10833–10842 (2019)
21. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. Communications of the ACM **63**(11), 139–144 (2020)
22. Grubert, J., Witzani, L., Ofek, E., Pahud, M., Kranz, M., Kristensson, P.O.: Effects of hand representations for typing in virtual reality. In: 2018 IEEE Conference on Virtual Reality and 3D User Interfaces (VR). pp. 151–158. IEEE (2018)
23. Hamill, J., Knutzen, K.M.: Biomechanical basis of human movement. Lippincott Williams & Wilkins (2006)
24. Hampali, S., Sarkar, S.D., Lepetit, V.: Ho-3d_v3: Improving the accuracy of hand-object annotations of the ho-3d dataset. arXiv preprint arXiv:2107.00887 (2021)
25. Hasson, Y., Tekin, B., Bogo, F., Laptev, I., Pollefeys, M., Schmid, C.: Leveraging photometric consistency over time for sparsely supervised hand-object reconstruction. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 571–580 (2020)
26. Hasson, Y., Varol, G., Tzionas, D., Kalevatykh, I., Black, M.J., Laptev, I., Schmid, C.: Learning joint reconstruction of hands and manipulated objects. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11807–11816 (2019)
27. He, K., Zhang, X., Ren, S., Sun, J.: Identity mappings in deep residual networks. In: European conference on computer vision. pp. 630–645. Springer (2016)
28. Herda, L., Urtasun, R., Fua, P.: Hierarchical implicit surface joint limits for human body tracking. Computer Vision and Image Understanding **99**(2), 189–209 (2005)
29. Jacobson, A., Kavan, L., Sorkine-Hornung, O.: Robust inside-outside segmentation using generalized winding numbers. ACM Transactions on Graphics (TOG) **32**(4), 1–12 (2013)

30. Jiang, Z., Rahmani, H., Black, S., Williams, B.M.: A probabilistic attention model with occlusion-aware texture regression for 3d hand reconstruction from a single rgb image. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 758–767 (2023)
31. Kanazawa, A., Black, M.J., Jacobs, D.W., Malik, J.: End-to-end recovery of human shape and pose. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7122–7131 (2018)
32. Kendall, A., Gal, Y.: What uncertainties do we need in bayesian deep learning for computer vision? Advances in neural information processing systems **30** (2017)
33. Kim, D.U., Kim, K.I., Baek, S.: End-to-end detection and pose estimation of two interacting hands. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11189–11198 (2021)
34. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
35. Kulon, D., Guler, R.A., Kokkinos, I., Bronstein, M.M., Zafeiriou, S.: Weakly-supervised mesh-convolutional hand reconstruction in the wild. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4990–5000 (2020)
36. Li, J., Bian, S., Zeng, A., Wang, C., Pang, B., Liu, W., Lu, C.: Human pose regression with residual log-likelihood estimation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 11025–11034 (2021)
37. Li, L., Tian, L., Zhang, X., Wang, Q., Zhang, B., Bo, L., Liu, M., Chen, C.: Renderih: A large-scale synthetic dataset for 3d interacting hand pose estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 20395–20405 (2023)
38. Li, M., An, L., Zhang, H., Wu, L., Chen, F., Yu, T., Liu, Y.: Interacting attention graph for single image two-hand reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2761–2770 (2022)
39. Lin, K., Wang, L., Liu, Z.: End-to-end human pose and mesh reconstruction with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1954–1963 (2021)
40. Lin, K., Wang, L., Liu, Z.: Mesh graphormer. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 12939–12948 (2021)
41. Liu, S., Jiang, H., Xu, J., Liu, S., Wang, X.: Semi-supervised 3d hand-object poses estimation with interactions in time. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14687–14697 (2021)
42. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9**(11) (2008)
43. Meng, H., Jin, S., Liu, W., Qian, C., Lin, M., Ouyang, W., Luo, P.: 3d interacting hand pose estimation by hand de-occlusion and removal. In: European Conference on Computer Vision. pp. 380–397. Springer (2022)
44. Moon, G.: Bringing inputs to shared domains for 3d interacting hands recovery in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17028–17037 (2023)
45. Moon, G., Chang, J.Y., Lee, K.M.: V2v-posenet: Voxel-to-voxel prediction network for accurate 3d hand and human pose estimation from a single depth map. In: Proceedings of the IEEE conference on computer vision and pattern Recognition. pp. 5079–5088 (2018)
46. Moon, G., Lee, K.M.: I2l-meshnet: Image-to-lixel prediction network for accurate 3d human pose and mesh estimation from a single rgb image. In: Computer Vision–

ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16. pp. 752–768. Springer (2020)
47. Moon, G., Saito, S., Xu, W., Joshi, R., Buffalini, J., Bellan, H., Rosen, N., Richardson, J., Mize, M., De Bree, P., et al.: A dataset of relighted 3d interacting hands. arXiv preprint arXiv:2310.17768 (2023)
48. 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: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 49–59 (2018)
49. Mueller, F., Mehta, D., Sotnychenko, O., Sridhar, S., Casas, D., Theobalt, C.: Real-time hand tracking under occlusion from an egocentric rgb-d sensor. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1154–1163 (2017)
50. Muller, L., Osman, A.A., Tang, S., Huang, C.H.P., Black, M.J.: On self-contact and human pose. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9990–9999 (2021)
51. Oberweger, M., Riegler, G., Wohlhart, P., Lepetit, V.: Efficiently creating 3d training data for fine hand pose estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4957–4965 (2016)
52. Oberweger, M., Wohlhart, P., Lepetit, V.: Generalized feedback loop for joint hand-object pose estimation. IEEE transactions on pattern analysis and machine intelligence **42**(8), 1898–1912 (2019)
53. Parelli, M., Papadimitriou, K., Potamianos, G., Pavlakos, G., Maragos, P.: Exploiting 3d hand pose estimation in deep learning-based sign language recognition from rgb videos. In: European Conference on Computer Vision. pp. 249–263. Springer (2020)
54. Park, J., Oh, Y., Moon, G., Choi, H., Lee, K.M.: Handoccnet: Occlusion-robust 3d hand mesh estimation network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1496–1505 (2022)
55. Pavlakos, G., Shan, D., Radosavovic, I., Kanazawa, A., Fouhey, D., Malik, J.: Reconstructing hands in 3d with transformers. arXiv preprint arXiv:2312.05251 (2023)
56. Qian, C., Sun, X., Wei, Y., Tang, X., Sun, J.: Realtime and robust hand tracking from depth. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1106–1113 (2014)
57. Ren, J., Zhu, J., Zhang, J.: End-to-end weakly-supervised single-stage multiple 3d hand mesh reconstruction from a single rgb image. arXiv preprint arXiv:2204.08154 (2022)
58. Ren, P., Wen, C., Zheng, X., Xue, Z., Sun, H., Qi, Q., Wang, J., Liao, J.: Decoupled iterative refinement framework for interacting hands reconstruction from a single rgb image. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8014–8025 (2023)
59. Romero, J., Tzionas, D., Black, M.J.: Embodied hands: Modeling and capturing hands and bodies together. ACM Transactions on Graphics (ToG) **36**(6), 1–17 (2017)
60. Schreuders, T.A., Brandsma, J.W., Stam, H.J.: Functional anatomy and biomechanics of the hand. In: Hand Function, pp. 3–22. Springer (2014)
61. Schultz, R., Storace, A., Krishnamurthy, S.: Metacarpophalangeal joint motion and the role of the collateral ligaments. International orthopaedics **11**(2), 149–155 (1987)

62. Spurr, A., Dahiya, A., Wang, X., Zhang, X., Hilliges, O.: Self-supervised 3d hand pose estimation from monocular rgb via contrastive learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11230–11239 (2021)
63. Spurr, A., Iqbal, U., Molchanov, P., Hilliges, O., Kautz, J.: Weakly supervised 3d hand pose estimation via biomechanical constraints. arXiv preprint arXiv:2003.09282 (2020)
64. Spurr, A., Song, J., Park, S., Hilliges, O.: Cross-modal deep variational hand pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 89–98 (2018)
65. Sridhar, S., Feit, A.M., Theobalt, C., Oulasvirta, A.: Investigating the dexterity of multi-finger input for mid-air text entry. In: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems. pp. 3643–3652 (2015)
66. Sridhar, S., Rhodin, H., Seidel, H.P., Oulasvirta, A., Theobalt, C.: Real-time hand tracking using a sum of anisotropic gaussians model. In: 2014 2nd International Conference on 3D Vision. vol. 1, pp. 319–326. IEEE (2014)
67. Tse, T.H.E., Kim, K.I., Leonardis, A., Chang, H.J.: Collaborative learning for hand and object reconstruction with attention-guided graph convolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1664–1674 (2022)
68. Tu, Z., Huang, Z., Chen, Y., Kang, D., Bao, L., Yang, B., Yuan, J.: Consistent 3d hand reconstruction in video via self-supervised learning. IEEE Transactions on Pattern Analysis and Machine Intelligence (2023)
69. Tzionas, D., Ballan, L., Srikantha, A., Aponte, P., Pollefeys, M., Gall, J.: Capturing hands in action using discriminative salient points and physics simulation. International Journal of Computer Vision (IJCV) **118**(2), 172–193 (Jun 2016), https://doi.org/10.1007/s11263-016-0895-4
70. Wang, C., Zhu, F., Wen, S.: Memahand: Exploiting mesh-mano interaction for single image two-hand reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 564–573 (2023)
71. Yang, L., Li, S., Lee, D., Yao, A.: Aligning latent spaces for 3d hand pose estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2335–2343 (2019)
72. Yang, L., Yao, A.: Disentangling latent hands for image synthesis and pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9877–9886 (2019)
73. Yang, L., Li, K., Zhan, X., Lv, J., Xu, W., Li, J., Lu, C.: Artiboost: Boosting articulated 3d hand-object pose estimation via online exploration and synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2750–2760 (2022)
74. Yu, Z., Huang, S., Fang, C., Breckon, T.P., Wang, J.: Acr: Attention collaboration-based regressor for arbitrary two-hand reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12955–12964 (2023)
75. Yu, Z., Li, C., Yang, L., Zheng, X., Mi, M.B., Lee, G.H., Yao, A.: Overcoming the trade-off between accuracy and plausibility in 3d hand shape reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 544–553 (2023)
76. Yuan, S., Garcia-Hernando, G., Stenger, B., Moon, G., Chang, J.Y., Lee, K.M., Molchanov, P., Kautz, J., Honari, S., Ge, L., et al.: Depth-based 3d hand pose

estimation: From current achievements to future goals. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2636–2645 (2018)
77. Yuan, S., Ye, Q., Stenger, B., Jain, S., Kim, T.K.: Bighand2. 2m benchmark: Hand pose dataset and state of the art analysis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4866–4874 (2017)
78. 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: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11354–11363 (2021)
79. Zhang, J., Jiao, J., Chen, M., Qu, L., Xu, X., Yang, Q.: 3d hand pose tracking and estimation using stereo matching. arXiv preprint arXiv:1610.07214 (2016)
80. Zhang, X., Pak, D.H., Ahn, S.S., Li, X., You, C., Staib, L., Sinusas, A.J., Wong, A., Duncan, J.S.: Heteroscedastic uncertainty estimation for probabilistic unsupervised registration of noisy medical images. arXiv preprint arXiv:2312.00836 (2023)
81. Zhang, X., Li, Q., Mo, H., Zhang, W., Zheng, W.: End-to-end hand mesh recovery from a monocular rgb image. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2354–2364 (2019)
82. Zhang, Y., Kephart, J.O., Cui, Z., Ji, Q.: Physpt: Physics-aware pretrained transformer for estimating human dynamics from monocular videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2305–2317 (June 2024)
83. Zhang, Y., Kephart, J.O., Ji, Q.: Incorporating physics principles for precise human motion prediction. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 6164–6174 (2024)
84. Zhang, Y., Wang, H., Kephart, J.O., Ji, Q.: Body knowledge and uncertainty modeling for monocular 3d human body reconstruction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9020–9032 (2023)
85. Zhou, Y., Habermann, M., Xu, W., Habibie, I., Theobalt, C., Xu, F.: Monocular real-time hand shape and motion capture using multi-modal data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5346–5355 (2020)
86. Zimmermann, C., Brox, T.: Learning to estimate 3d hand pose from single rgb images. In: Proceedings of the IEEE international conference on computer vision. pp. 4903–4911 (2017)
87. Zimmermann, C., Ceylan, D., Yang, J., Russell, B., Argus, M., Brox, T.: Freihand: A dataset for markerless capture of hand pose and shape from single rgb images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 813–822 (2019)
88. Zuo, B., Zhao, Z., Sun, W., Xie, W., Xue, Z., Wang, Y.: Reconstructing interacting hands with interaction prior from monocular images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9054–9064 (2023)