This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

JUMP-Hand: Learning Joint-wise Uncertainty to Gate Mixture of View Experts for Multi-View 3D Hand Reconstruction

Haohong Kuang[1], Yang Xiao[1]*, Changlong Jiang[1], Jinghong Zheng[1], Hang Xu[1], Ran Wang[2], Zhiguo Cao[3], and Joey Tianyi Zhou[4,5]

[1]National Key Laboratory of Multispectral Information Intelligent Processing Technology, School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan 430074, China
[2]School of Journalism and Information Communication, Huazhong University of Science and Technology
[3]School of Artificial Intelligence and Automation, Huazhong University of Science and Technology
[4]Centre for Frontier AI Research (CFAR), Agency for Science, Technology and Research (A*STAR), Singapore
[5]Institute of High Performance Computing (IHPC), Agency for Science, Technology and Research (A*STAR), Singapore
{haohong_kuang, Yang_Xiao, cljiang, deepzheng, hang_xu, rex_wang, zgcao}@hust.edu.cn
joey_zhou@a-star.edu.sg

Abstract

We propose JUMP-Hand, a novel multi-view 3D hand reconstruction method that explicitly models probabilistic joint-wise uncertainty as a gating mechanism for multi-view fusion. Existing approaches usually rely on naive pooling or implicit attention, overlooking that each hand joint exhibits varying visibility and reliability across views. Such indiscriminate aggregation of noisy or unreliable information inevitably degrades overall performance. For instance, a joint severely occluded in one view might be clearly visible in another. To address this, JUMP-Hand draws inspiration from the Mixture of Experts (MoE) paradigm, treating each 2D view as a specialized expert. The key idea is that the reliability of each view expert is quantified through joint-wise uncertainty modeling, serving as an explicit gating signal to route experts' partial yet complementary clues for each joint in a coarse-to-fine reconstruction paradigm. In this design, uncertainty not only guides the uncertainty-aware triangulation for reliable 3D hand initialization during the coarse stage, but also acts as a gating signal during the refinement stage to adaptively aggregate multi-scale features from different view experts on a joint-wise basis, enabling robust 3D hand reconstruction. Extensive experiments on DexYCB-MV, HO3D-MV, and OakInk-MV demonstrate that our method consistently achieves state-of-the-art results, validating the effectiveness of the proposed method with joint-wise uncertainty gating for reliable 3D hand reconstruction. Code is available at https://github.com/HaohongKuang/JUMP-Hand.

*Yang Xiao is corresponding author (Yang_Xiao@hust.edu.cn).

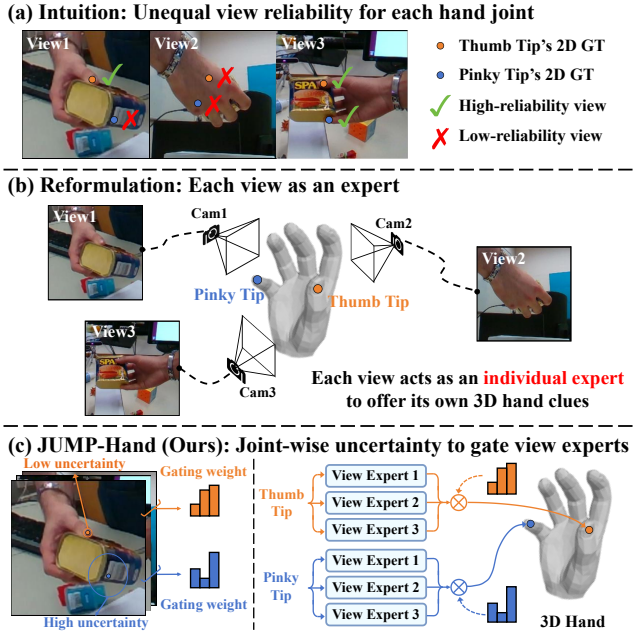

Figure 1. **Our main research intuition and idea.** (a) Since hand joints have varying reliability across different views, (b) we formulate the multi-view 3D hand reconstruction task into a MoE task, treating different views as different experts, and (c) using the predicted uncertainty to gate the fusion among view experts.

1. Introduction

Accurate reconstruction of 3D hand pose and shape is a fundamental task in computer vision, crucial for downstream tasks such as AR/VR, human-object interaction, and robotic grasping [2, 24, 28]. While monocular 3D hand reconstruction and hand pose estimation have pro-

gressed [1, 15, 16, 23, 31, 38, 39], it remains an ill-posed problem due to depth ambiguity and occlusions [12, 27].

Relative to monocular setting, multi-view paradigms benefit from richer cross-view geometry and lead to more stable reconstruction. Nonetheless, despite the promising potential, current multi-view solutions are hindered by the difficulty of effectively integrating features from multiple viewpoints. Recent state-of-the-art (SOTA) methods, such as POEM [35] and MLPHand [33], have adopted advanced fusion strategies including transformer-based architectures [29] and GCN-based methods [4]. However, a key limitation of them is their tendency to aggregate information indiscriminately. They rely on implicit, black-box attention [30] or naive pooling, treating all views equally. This overlooks a core problem: observation reliability is inherently joint-wise and view-dependent. As in Fig. 1(a), the thumb tip is visible in one view but occluded in another. Indiscriminately aggregating low-reliability information can contaminate the fusion process and degrade performance.

To address this, we draw inspiration from the Mixture-of-Experts (MoE) paradigm [5, 8, 26] and propose a novel perspective for multi-view 3D hand reconstruction. As in Fig. 1(b), each camera captures a 2D projection of the 3D hand, providing partial yet complementary geometric cues from its unique perspective. These views serve as independent yet specialized **view experts** for reconstructing the 3D hand. Although an individual expert may suffer from occlusions or ambiguities, each expert contributes unique observations, which enable complete and accurate reconstruction when properly combined. The key challenge, therefore, lies in *how to effectively aggregate these expert opinions on a per-joint basis?* We argue that prediction uncertainty provides a natural and learnable confidence indicator [9, 18]: low uncertainty corresponds to well-observed joints with high-quality detections, while high uncertainty arises from low-quality conditions such as occlusions or motion blur. Therefore, *the joint-wise and view-dependent uncertainty can serve as an explicit, physically interpretable gating signal to route expert contributions adaptively* (Fig. 1(c)), allowing reliable views to dominate reconstruction while suppressing noisy or unreliable ones.

Building on this insight, we propose JUMP-Hand, a novel method that introduces joint-wise uncertainty modeling as a principled and learnable gating mechanism for multi-view 3D hand reconstruction. Our approach consists of three key components: (1) a joint-wise uncertainty modeling module where each view independently predicts 2D joint locations along with their associated uncertainties as Gaussian distributions, supervised via negative log-likelihood loss to provide explicit, physically interpretable confidence indicators; (2) uncertainty-aware triangulation for coarse 3D initialization, where each view's contribution is weighted according to its predicted uncertainty, providing geometrically consistent initialization robust to low-quality observations; and (3) an uncertainty-gated iterative decoder that employs uncertainty-gated cross-attention (UGCA) to adaptively fuse multi-view features on a per-joint basis, progressively refining the 3D hand mesh. Throughout the entire pipeline, joint-wise uncertainty serves as an explicit gating signal, enabling robust reconstruction with clear interpretability and natural adaptivity to occlusion patterns.

Extensive experiments on three large-scale multi-view benchmarks [3, 10, 34] validate that JUMP-Hand achieves SOTA performance and significantly outperforms other methods. This superiority is especially evident in challenging scenarios: on a difficult subset of HO3D, JUMP-Hand achieves a substantial 29.9% MPVPE improvement over POEM. These results validate the effectiveness of our uncertainty-guided paradigm for reliable 3D hand reconstruction, especially in scenarios with heavy occlusion.

Our contributions are summarized as follows:

- We are the first to introduce probabilistic, joint-wise uncertainty as a principled and explicit gating signal for multi-view 3D hand reconstruction, which quantifies the reliability of each view on a per-joint basis.
- We propose a novel two-stage reconstruction paradigm where uncertainty-gating is integrated throughout: Uncertainty-Guided DLT (UG-DLT) for robust geometric initialization, and Uncertainty-Gated Cross-Attention (UGCA) for adaptive refinement.
- Extensive experiments on three benchmarks demonstrate that our method achieves SOTA results, especially under challenging scenarios, validating exceptional robustness under severe occlusions and degraded viewing conditions.

2. Related Works

Multi-view 3D Hand Reconstruction. Multi-view 3D hand reconstruction aims to overcome the inherent occlusion and depth ambiguity issues of monocular methods, with its core challenge lying in how to effectively fuse features from different viewpoints. Recent SOTA methods, such as POEM [35], MLPHand [33], and other Transformer or GCN-based architectures [4, 29, 36], have adopted advanced fusion strategies, including cross-set point attention, MLP-based geometric fusion, and cross-attention. Another important line of work focuses on improving the triangulation process itself. [14] proposes two approaches: one introduces learnable confidence weights into algebraic triangulation, and the other constructs a 3D feature voxel grid for aggregation. However, they typically rely on implicit, black-box attention mechanisms or simple pooling to treat all views blindly, without explicitly modeling observation reliability, which is joint-wise and view-dependent. To fill this gap, we propose JUMP-Hand, which is the first method to introduce probabilistic, joint-wise uncertainty as an explicit, physically-interpretable gating signal. This design

allows us to dynamically and adaptively aggregate information from the most reliable views while suppressing noise.

Mixture-of-Experts. Mixture-of-Experts (MoE) model is a powerful paradigm that aggregates contributions from multiple specialized experts using a gating mechanism [5, 8, 21, 26, 32]. Our work innovatively introduces the core concept of MoE to multi-view 3D hand reconstruction task. Instead of using different networks as experts in the naive MoE architecture, such as FFNs in transformers, we innovatively treat each camera view as an independent view expert. Furthermore, rather than using a black-box learnable gating network, we use probabilistic joint-wise uncertainty as the gating signal. This design provides clear physical interpretability: for each hand joint, view experts with low uncertainty are assigned a larger weight during aggregation.

Uncertainty Modeling in 3D Vision. Uncertainty estimation, which distinguishes between aleatoric and epistemic uncertainty [18], has become critical in 3D vision tasks. Techniques like MC-Dropout [9] and Bayesian fusion [20] have been widely explored for quantifying prediction confidence. In hand reconstruction, recent works employ NLL loss [17, 37] or diffusion models [22] for single-view settings. For multi-view scenarios, UPose3D [6] uses uncertainty in an MLE framework to improve robustness, but treats it as optimization input rather than explicit routing. However, they typically focus on single-view scenarios or treat uncertainty passively as a post-hoc confidence metric, rather than as an active, integral component for multi-view fusion. In contrast, JUMP-Hand is the first to employ joint-wise uncertainty as an explicit, physically interpretable gating signal, enabling adaptive per-joint aggregation that suppresses unreliable views while emphasizing confident ones.

3. Method

3.1. Overview

Preliminaries. Given a set of N synchronized multi-view images $\mathcal{I} = \{\mathbf{I}_i\}_{i=1}^{N}$ as input, our model reconstructs the 3D hand pose and shape, directly represented by the spatial coordinates of hand joints $\mathbf{J}_{3D} \in \mathbb{R}^{21\times 3}$ and mesh vertices $\mathbf{V}_{3D} \in \mathbb{R}^{778\times 3}$ in MANO format.

Pipeline. The overall pipeline of JUMP-Hand is illustrated in Fig. 2. Each input view $\mathrm{I_n}$ is first processed by a shared CNN backbone to extract multi-scale visual features $\mathrm{F_n^s}$. Guided by our proposed gating mechanism, the pipeline operates as a coarse-to-fine Mixture-of-Experts (MoE) architecture. Specifically, we first formulate a joint-wise gating signal to explicitly quantify the confidence of each view expert (Sec. 3.2). This principled signal then dictates the uncertainty-guided triangulation (UG-DLT) to establish a robust coarse 3D initialization (Sec. 3.3). Finally, it also guides deformable multi-view feature fusion across views to recover fine-grained geometric details (Sec. 3.4).

3.2. Joint-wise Uncertainty Modeling

To construct the gating signal, we model each 2D joint observation as a Gaussian variable to quantify view-dependent observation uncertainty [18]. Specifically, for the j-th hand joint observed from the n-th view, we model the observed 2D joint location $\mathbf{p}_{j,n}^{2D} \in \mathbb{R}^2$ as

$$\mathbf{p}_{j,n}^{2D} \sim \mathcal{N}(\boldsymbol{\mu}_{j,n}, \boldsymbol{\Sigma}_{j,n}), \tag{1}$$

where $\boldsymbol{\mu}_{j,n} = [\mu_{j,n}^x, \mu_{j,n}^y]^\top$ denotes the predicted mean 2D location, and $\boldsymbol{\Sigma}_{j,n} = \mathrm{diag}\left((\sigma_{j,n}^x)^2, (\sigma_{j,n}^y)^2\right)$ is a diagonal covariance matrix that captures uncertainty along the x and y directions. This diagonal form assumes independent uncertainty along the two image axes, which is sufficient for our joint-wise gating design. Intuitively, a low variance indicates that the joint is clearly observed in that view, while a high variance reflects ambiguity caused by occlusion, motion blur, or poor imaging conditions.

To estimate these parameters in a learnable manner, we design a specialized uncertainty estimation branch built upon the backbone features $\mathbf{F}_n$. The branch consists of three components: (1) a Feature Pyramid Network (FPN) that captures multi-scale spatial information, (2) a heatmap head $\mathcal{H}$ that predicts the mean location, and (3) a variance head $\mathcal{V}$ that directly predicts joint-wise variance:

$$\boldsymbol{\mu}_{j,n} = \mathcal{H}(\mathrm{FPN}(\mathbf{F}_n)), \quad \boldsymbol{\Sigma}_{j,n} = \mathcal{V}(\mathrm{FPN}(\mathbf{F}_n)). \tag{2}$$

To ensure valid uncertainty estimates, the output of variance head $\mathcal{V}$ is passed through a Softplus activation function [7] to ensure the predicted variance $\sigma_{j,n}^2$ is always positive. Here, $\sigma_{j,n}^2$ denotes the scalar joint-wise uncertainty obtained by averaging the predicted variances along the x and y directions.

To train this probabilistic model, we employ a negative log-likelihood (NLL) loss [18]. Specifically, using the ground-truth 2D joint locations as observations, we minimize their negative log-likelihood under the predicted Gaussian distribution, thereby encouraging the network to jointly learn both the mean and the variance.

The learned variance $\boldsymbol{\sigma}_{j,n}^2$ serves two roles in our pipeline. First, it provides interpretability by explicitly quantifying the model's confidence for each joint prediction, as shown in Fig. 3(b). Second, it serves as the soft gating signal used in both the coarse triangulation and refinement stages. This joint-wise uncertainty acts as a probabilistic routing signal, connecting the 2D observation space and 3D reconstruction space.

3.3. Uncertainty-Guided Coarse Initialization

After obtaining multi-view 2D joint predictions $\boldsymbol{\mu}_{j,n}$ and their associated uncertainty $\sigma_{j,n}^2$, the coarse reconstruction

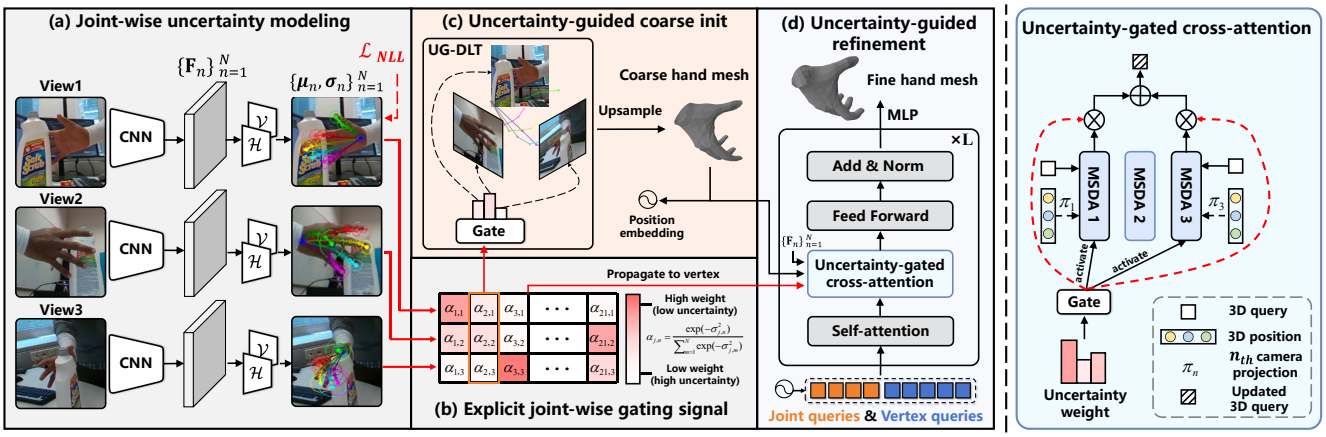

Figure 2. The pipeline of JUMP-Hand. (a) Joint-wise uncertainty modeling: each view independently predicts 2D joint locations and uncertainties (μ_n, σ_n) supervised by NLL loss. $\mathcal{H}$ and $\mathcal{V}$ represents heatmap head and variance head respectively. (b) Uncertainty is converted to explicit joint-wise gating signal to guide coarse-to-fine reconstruction. (c) Coarse stage: uncertainty-guided DLT (UG-DLT) performs weighted triangulation to obtain robust 3D initialization. (d) Refinement stage: refined iterative regression through query modeling. Uncertainty-gated cross attention (UGCA) is the core to fuse each view's features through gating mechanism.

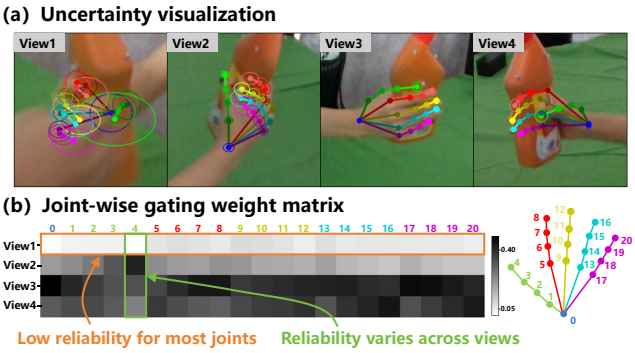

Figure 3. (a) Visualization of uncertainty. The joint-wise ellipses display the standard deviation in two directions. (b) Corresponding joint-wise gating weights.

stage aims to produce a robust initial 3D hand via geometric triangulation. A fundamental challenge is that the classical triangulation method DLT [11] assigns uniform confidence across all views, making it vulnerable to noisy observations from occluded or low-reliability views.

To address this limitation, we reformulate the triangulation process from a probabilistic perspective and propose Uncertainty-Guided DLT (UG-DLT). The core idea is to leverage the joint-wise uncertainty $\sigma_{j,n}^2$ as an explicit gating signal to control each view's contribution to the reconstructed 3D point. Specifically, we first convert the predicted variances into normalized weights $\alpha_{j,n}$, where views with high uncertainty receive low weights:

$$\alpha_{j,n} = \frac{\exp(-\sigma_{j,n}^2)}{\sum_{m=1}^{N} \exp(-\sigma_{j,m}^2)}, \quad j = 1, \ldots, 21. \tag{3}$$

Standard DLT solves $A_j J_j^{3D} = 0$. To favor more reliable views in triangulation, we construct a weighted constraint equation by expanding $\alpha_{j,n}$ into a weight vector $\mathbf{w}_j$ and applying an element-wise multiplication (Hadamard product $\circ$) to the constraint matrix $\mathbf{A}_j$:

$$(\mathbf{w}_j \circ \mathbf{A}_j)\mathbf{J}_j^{3D,(0)} = \mathbf{0}, \tag{4}$$

which is solved with SVD in a differentiable manner. In this way, UG-DLT suppresses the influence of unreliable views on a per-joint basis, yielding coarse 3D joints $\hat{\mathbf{J}}^{3D,(0)}$ that are more robust to occlusion and noise. This initialization, which already respects multi-view geometric consistency, provides a solid foundation for the subsequent refinement stage. Finally, we upsample the coarse joints $\hat{\mathbf{J}}^{3D,(0)}$ to obtain a coarse mesh $\hat{\mathbf{V}}^{3D,(0)}$ using a learnable linear layer.

3.4. Uncertainty-Guided Refinement

While UG-DLT provides a geometrically sound coarse initialization, it operates only at the joint level and cannot recover fine-grained surface details. The refinement stage therefore fuses multi-scale visual features from multiple viewpoints to fully reconstruct the detailed hand geometry. The key challenge remains consistent: how to aggregate contributions from multiple views without being corrupted by unreliable observations?

Following the expert-routing perspective established in the coarse stage, we design the refinement process such that each view continues to serve as an independent expert, specializing in extracting visual features from its own perspective. We implement an iterative Transformer decoder with L stacked layers, using the coarse 3D joints $\mathbf{J}^{3D,(0)}$ and vertices $\mathbf{V}^{3D,(0)}$ as the initial 3D queries. Each decoder layer consists of three modules: self-attention to model structural dependencies among joints and vertices, uncertainty-gated cross-attention (UGCA) to aggregate multi-view image features, and a feed-forward network.

UGCA is the core module for view-aware feature aggregation in the refinement stage. For each query, we project its 3D coordinate onto all N views using the camera parameters, obtaining N reference points $\{\mathbf{p}_{j,n}^{2D}\}_{n=1}^{N}$. Around each reference point, the corresponding image feature is extracted independently from each view. For the n-th view, we use multi-scale deformable attention [40] to sample informative features around the projected location:

$$y_{j,n} = \sum_{s=1}^{S}\sum_{k=1}^{K} A_{nsk} \cdot \phi(\mathbf{F}_n^s, \mathbf{p}_{j,n}^{2D} + \Delta p_{nsk}), \quad (5)$$

where the learnable offsets Δp_{nsk} allow the model to adaptively sample informative positions around the projected reference point across multiple feature scales. In this way, deformable attention determines where to extract features within each view, while the uncertainty-guided fusion described below determines how much each view should contribute across views. Different views provide complementary evidence: views with clear visibility contribute precise local details, while partially occluded views may still provide useful contextual cues.

After obtaining N expert responses $\{y_{j,n}\}_{n=1}^{N}$, a robust gating mechanism is essential to suppress noise from occluded or low-quality views. Rather than uniform averaging, we reuse the uncertainty-derived weights $\alpha_{j,n}$ from Eq. 3 as our gating signal to route expert contributions:

$$y_j = \sum_{n=1}^{N} \alpha_{j,n} \cdot y_{j,n}. \quad (6)$$

This design ensures architectural consistency with our coarse stage, where uncertainty serves as an explicit, physically interpretable gating signal throughout the entire pipeline. For vertex queries, we extend this mechanism by propagating joint-wise uncertainty via the MANO [25] skinning weight matrix $\mathbf{W} \in \mathbb{R}^{778\times 21}$:

$$\sigma_{v,n}^2 = \sum_{j=1}^{21} W_{v,j} \cdot \sigma_{j,n}^2, \quad (7)$$

where $W_{v,j}$ encodes the kinematic influence of joint j on vertex v. Consequently, vertex queries inherit uncertainty from their driving joints, enabling a unified gating strategy for both joints and the full mesh.

This uncertainty-guided gating design offers three key advantages. First, unlike implicit attention mechanisms or black-box routers, it provides an explicit and physically interpretable confidence measure for each view. Second, the joint-wise (and vertex-wise) design naturally captures that different parts of the hand may rely on different combinations of views. Third, it maintains architectural consistency across the entire pipeline: the same uncertainty-guided gating mechanism governs both the coarse (UG-DLT) and refinement stages, creating a unified framework aligned with the MoE philosophy where each expert's contribution is strictly proportional to its estimated reliability.

The fused feature y_j is then processed through an FFN with residual connections, followed by an MLP regressor to predict coordinate updates. In each decoder layer, the updated 3D queries are re-projected to generate refined reference points for the subsequent UGCA module. This iterative process progressively improves the query locations and enables more accurate feature extraction in subsequent layers. Combined with uncertainty-guided gating, our method recovers fine-grained geometric details such as finger articulation and local surface variations while maintaining robustness to occlusions and ambiguous observations.

3.5. Loss

Our training objective consists of multiple loss terms that supervise different aspects of the reconstruction pipeline.

Probabilistic 2D Joint Supervision. To train the uncertainty modeling module, we employ the Gaussian Negative Log-Likelihood (NLL) loss [18]. Given the ground-truth 2D joint position $\bar{\mathbf{J}}_{j,n}^{2D}$ and the predicted Gaussian parameters $\boldsymbol{\mu}_{j,n}, \boldsymbol{\sigma}_{j,n}^2$ for the j-th joint in the n-th view, the NLL loss encourages both accurate mean predictions and well-calibrated uncertainty estimates:

$$\mathcal{L}_{NLL} = \frac{1}{N}\sum_{n=1}^{N}\sum_{j=1}^{J}\left(\log \boldsymbol{\sigma}_{j,n} + \frac{(\bar{\mathbf{J}}_{j,n}^{2D} - \boldsymbol{\mu}_{j,n})^2}{2\boldsymbol{\sigma}_{j,n}^2}\right). \quad (8)$$

This probabilistic formulation naturally balances two objectives: minimizing prediction error while preventing overconfident predictions.

3D Reconstruction Supervision. For end-to-end training, we combine several geometric consistency losses:

$$\begin{aligned}
\mathcal{L}_{3D} &= \|\hat{\mathbf{J}}^{3D} - \mathbf{J}^{3D}\|_1 + \|\hat{\mathbf{V}}^{3D} - \mathbf{V}^{3D}\|_1, \\
\mathcal{L}_{2D} &= \|\Pi_n(\hat{\mathbf{J}}^{3D}) - \mathbf{J}_n^{2D}\|_1 + \|\Pi_n(\hat{\mathbf{V}}^{3D}) - \mathbf{V}_n^{2D}\|_1, \\
\mathcal{L}_{DLT} &= \|\hat{\mathbf{J}}^{3D,(0)} - \mathbf{J}^{3D}\|_1, \\
\mathcal{L} &= \lambda_1\mathcal{L}_{NLL} + \lambda_2\mathcal{L}_{3D} + \lambda_3\mathcal{L}_{2D} + \lambda_4\mathcal{L}_{DLT},
\end{aligned} \quad (9)$$

where $\mathcal{L}_{3D}$ enforces direct 3D supervision on joints and vertices, $\mathcal{L}_{2D}$ ensures multi-view reprojection consistency by projecting 3D predictions back to each view via camera projection Π_i, and $\mathcal{L}_{DLT}$ provides intermediate supervision at the coarse reconstruction stage. The balancing coefficients $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ control the relative importance of each term.

4. Experiment

4.1. Datasets and Evaluation Metrics

HO3D-MV builds on HO3D v3 [10], which provides RGB-D sequences of hand-object interactions from up to 5 calibrated cameras. Following POEM [35], we select only the

Dataset	Methods	MPVPE ↓	RR-V ↓	PA-V ↓	AUC-V ↑	MPJPE ↓	RR-J ↓	PA-J ↓	AUC-J ↑
HO3D-MV	MVP [36]	20.95	27.08	10.04	0.59	18.72	24.90	10.44	0.60
	POEM [35]	17.20	21.45	9.97	0.66	17.28	21.94	9.60	0.63
	MLPHand [33]	18.69	23.28	10.54	-	18.70	23.76	10.11	-
	Ours	**13.39**↓3.81	**16.96**↓4.49	**8.78**↓1.19	**0.73**↑0.07	**13.10**↓4.18	**17.07**↓4.87	**8.30**↓1.30	**0.72**↑0.09
DexYCB-MV	MVP [36]	9.77	12.18	8.14	0.53	6.23	9.47	4.26	0.69
	POEM [35]	6.13	7.21	4.00	0.70	6.06	7.30	3.93	0.68
	MLPHand [33]	6.20	7.30	4.15	-	6.23	7.41	4.11	-
	Ours	**5.45**↓0.68	**6.41**↓0.80	**3.77**↓0.23	**0.73**↑0.03	**5.31**↓0.75	**6.41**↓0.89	**3.61**↓0.32	**0.72**↑0.03
OakInk-MV	MVP [36]	9.69	11.75	7.74	0.53	7.32	9.99	4.97	0.64
	POEM [35]	6.20	7.63	4.21	0.70	6.01	7.46	4.00	0.69
	MLPHand [33]	6.17	7.49	4.26	-	6.00	7.52	4.09	-
	Ours	**5.94**↓0.23	**6.99**↓0.50	**4.13**↓0.08	**0.72**↑0.02	**5.72**↓0.28	**6.94**↓0.52	**3.86**↓0.14	**0.71**↑0.02

Table 1. Comparison with the state-of-the-art on **OakInk-MV**, **DexYCB-MV**, and **HO3D-MV** datasets. ↓ indicates lower is better, ↑ indicates higher is better. All distance metrics are in millimeters (mm). AUC-V/J represent the area under the PCK curve for vertex and joint accuracy, respectively. **Best** in bold.

7 sequences containing complete 5-camera observations. This split yields 9,087 training and 2,706 test frames.

DexYCB-MV is derived from the DexYCB dataset [3], which captures hand-object interactions using 8 synchronized RGB-D cameras across 10 subjects. Following the official "S0" split and consistent with POEM [35], we exclude all left-hand samples and construct a multi-view version, referred to as DexYCB-MV. This results in 25,387 training, 1,412 validation, and 4,951 test multi-view frames.

OakInk-MV originates from the OakInk dataset [34], a large-scale multi-view benchmark for diverse hand-object interactions. Following POEM [35], we use its multi-view split with synchronized RGB images from calibrated views and accurate 3D hand mesh annotations.

Evaluation Metrics. We report MPJPE and MPVPE (in mm), measuring average Euclidean errors for hand joints and vertices, respectively. Following MANO's vertex-to-keypoint mapping, reported joints are derived from predicted mesh vertices in the refinement stage. Both metrics are evaluated under two settings: root-relative (RR), which normalizes coordinates to the wrist joint to assess relative hand structure, and Procrustes-aligned (PA), which applies optimal rigid alignment to isolate shape accuracy from global pose. Additionally, we use AUC (Area Under the Curve) computed from the Percentage of Correct Keypoints across distance thresholds to evaluate localization precision.

4.2. Implementation Details

The input RGB images from all views are resized and center-cropped to 256×256. We employ a shared CNN backbone based on ResNet-34 [13], initialized with pre-trained weights. All models are trained on two NVIDIA RTX 3090 GPUs for 100 epochs, using the Adam optimizer [19], with an initial learning rate of 1×10^{-4}, which is decayed by a factor of 10 at the 60th epoch.

Subset	Methods	MPVPE↓	PA-V↓	AUC-V↑
HO3D	POEM	35.52	18.45	0.04
	Ours	**24.91**↓10.61	**14.12**↓4.33	**0.13**↑0.09
DexYCB	POEM	13.88	7.96	0.38
	Ours	**10.37**↓3.27	**6.67**↓1.29	**0.52**↑0.14
OakInk	POEM	13.85	9.39	0.42
	Ours	**11.75**↓2.10	**8.03**↓1.36	**0.50**↑0.08

Table 2. Quantitative comparison on 3 challenging subsets.

4.3. Comparison with State-of-the-Art Methods

We conduct a comprehensive comparison between our proposed JUMP-Hand method and state-of-the-art approaches on three multi-view hand reconstruction datasets: HO3D-MV, DexYCB-MV, and OakInk-MV.

As shown in Table 1, JUMP-Hand achieves the best performance across all datasets and metrics, significantly outperforming existing methods. On HO3D-MV, which features challenging hand-object occlusions, JUMP-Hand achieves 13.39mm MPVPE, improving over POEM (17.20mm) by 22.2% and over MLPHand (18.69mm) by 28.4%. The PA-V metric further validates our superior geometric accuracy (8.78mm vs. 9.97mm for POEM), indicating better pose structure recovery. The substantial gains on this dataset highlight the robustness of joint-wise uncertainty modeling when dealing with severe occlusions, precisely the scenario where different hand joints exhibit highly variable visibility across views.

On DexYCB-MV and OakInk-MV, our method maintains consistent advantages with 5.45mm and 5.94mm MPVPE respectively, outperforming POEM by 11.1% and 4.2%. Notably, our AUC scores (0.72-0.73) surpass all baselines, indicating superior localization accuracy across a wide range of distance thresholds.

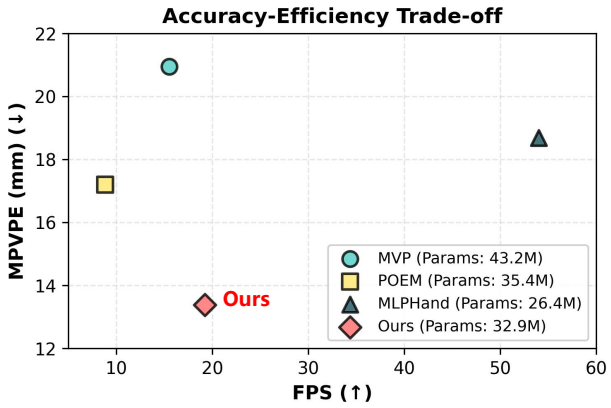

Figure 4. Accuracy-efficiency trade-off on HO3D-MV. Experiments are conducted on a single RTX 3090 GPU.

To further demonstrate the robustness of our method under degraded 2D image quality, we construct challenging subsets by selecting the 10% cases with the highest per-view 2D pose errors from POEM predictions across all three datasets. As shown in Table 2, our method demonstrates remarkable robustness in these challenging scenarios. On the HO3D challenging subset, JUMP-Hand achieves 24.91mm MPVPE compared to POEM's 35.52mm, a substantial 29.9% improvement. Crucially, the AUC-V score increases from 0.04 to 0.13, demonstrating significantly better keypoint localization under severe quality degradation. Similar improvements are observed on the challenging subsets of DexYCB (13.88mm to 10.37mm) and OakInk (13.85mm to 11.75mm). These results validate that our joint-wise uncertainty modeling provides explicit confidence indicators that naturally adapt to challenging viewing conditions, enabling the network to down-weight unreliable observations while leveraging high-quality cues from complementary views.

Efficiency analysis. Beyond accuracy, our method also demonstrates favorable computational efficiency. As illustrated in Fig. 4, we achieve an optimal balance between reconstruction quality and inference speed.

4.4. Ablation

We conduct ablation studies on HO3D-MV to validate the main design choices, as reported in Table 3 and Table 4.

Gating Signals. We compare three gating mechanisms in row (a)-(c) of Table 3. (a) Average Gating treats all views uniformly (16.31mm MPJPE), serving as a baseline but suffering when views exhibit vastly different reliability. (b) Learnable Gating employs a trainable router network, improving performance to 15.54mm (-0.77mm). However, it lacks explicit probabilistic grounding and struggles to generalize to novel occlusion patterns. (c) Uncertainty-based Gating (Ours) achieves the best result of 13.10mm, yielding substantial improvements of 3.21mm and 2.44mm over average and learnable gating, respectively. This validates that joint-wise uncertainty provides a principled, interpretable confidence measure that naturally adapts to view-dependent visibility variations. The progressive performance gain demonstrates that explicitly modeling observation reliability is crucial for robust multi-view fusion.

ID	Gating	C	R	MPJPE↓	PA-J↓	MPVPE↓	PA-V↓
(a)	Avg.	✔	✔	16.31	9.13	16.43	9.54
(b)	Learn.	✔	✔	15.54	8.66	15.68	9.19
(c)	Uncert.	✔	✔	**13.10**	**8.30**	**13.39**	**8.78**
(d)	Uncert.	✔	✗	16.61	9.80	16.67	10.32
(e)	Uncert.	✗	✔	21.55	10.98	22.19	11.83

Table 3. **Ablation on gating signals and reconstruction stages.** 'C' and 'R' denote the coarse and refinement stages. Experiment (c) is our full model with uncertainty-based gating in both stages. Metrics are reported on HO3D-MV (mm).

Datasets	Gating Mechanism	MPJPE↓	PA-J↓	MPVPE↓	PA-V↓
HO3D-MV	Hard (Top-2)	18.16	10.99	18.52	11.57
	Hard (Top-3)	15.47	9.09	15.70	9.62
	Soft (Ours)	**13.10**	**8.30**	**13.39**	**8.78**
DexYCB-MV	Hard (Top-2)	6.23	4.24	6.39	4.41
	Hard (Top-4)	5.70	3.79	5.90	4.01
	Soft (Ours)	**5.31**	**3.61**	**5.45**	**3.77**
OakInk-MV	Hard (Top-2)	6.21	4.15	6.45	4.46
	Hard (Top-3)	5.88	3.94	6.09	4.21
	Soft (Ours)	**5.72**	**3.86**	**5.94**	**4.13**

Table 4. **Comparison of gating mechanisms.** Soft gating uses all view experts, while Top-k selects the k most confident views for each joint and discards the rest. Metrics are reported in mm.

Reconstruction Stages. Table 3 (d-e) reveals critical insights into our two-stage design. (d) Removing the refinement stage (relying solely on UG-DLT) increases the error to 16.61mm (+3.51mm). This confirms that while geometric triangulation provides a sound initialization, iterative feature refinement is essential for recovering fine-grained details. Conversely, (e) Removing the coarse stage causes a severe degradation to 21.55mm (+8.45mm). This asymmetry highlights a crucial finding: a robust geometric initialization is fundamentally more critical than subsequent refinement. Without our uncertainty-weighted triangulation, corrupted initial estimates inherently propagate through the network, making accurate recovery nearly impossible. The 4.94mm larger performance drop in (e) vs. (d) quantitatively demonstrates that securing a reliable initialization is the primary bottleneck in multi-view reconstruction.

Hard vs. Soft Gating. Table 4 compares top-k hard gating (selecting only the k most confident views per joint) against our soft gating across three benchmarks. An interesting pattern emerges: hard gating steadily improves as k

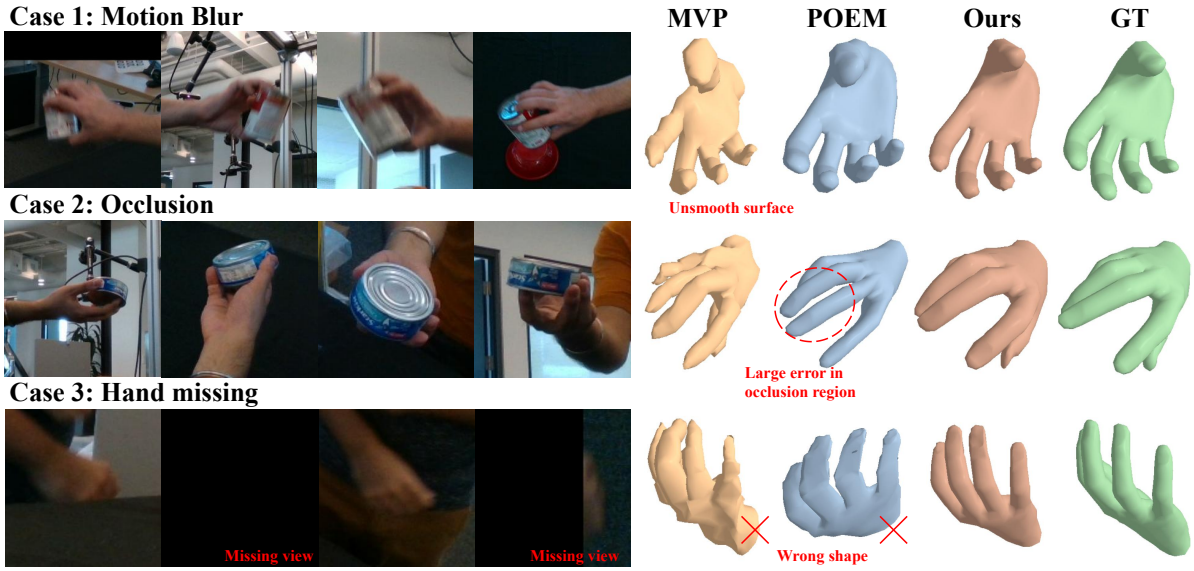

Figure 5. **Qualitative comparison.** We visualized 2D images with challenges from multiple perspectives in the dataset, including (1) motion blur, (2) occlusion, and (3) hand missing. When 3D hand reconstruction is performed based on these images, our method achieves better reconstruction results compared with SOTA methods. This indicates that our explicit modeling approach has superior robustness.

increases. On HO3D-MV, top-2 yields 18.16mm, while top-3 achieves 15.47mm (-2.69mm). However, even the best hard gating variants significantly underperform our soft gating: by 5.06mm on HO3D-MV, 0.92mm on DexYCB-MV, and 0.16mm on OakInk-MV. Notably, this performance gap is largest on HO3D-MV, the most challenging dataset with severe occlusions. This suggests that completely discarding less reliable views via hard thresholding irreversibly eliminates valuable contextual cues.

More revealingly, comparing across all three approaches unveils a clear hierarchy of fusion strategies. POEM's implicit attention achieves 17.28mm on HO3D-MV, falling between hard gating top-2 (18.16mm) and top-3 (15.47mm), yet still 4.18mm worse than our soft gating (13.10mm). This gap is most pronounced on HO3D-MV, where soft gating outperforms POEM by 24.2%, compared to more modest gains of 12.4% on DexYCB-MV and 4.8% on OakInk-MV. This pattern reveals a fundamental insight: the benefit of explicit, uncertainty-driven gating scales with scene complexity. When occlusions are severe, implicit attention mechanisms struggle to identify and suppress unreliable observations. Our soft gating strikes the optimal balance by proportionally down-weighting noisy contributions based on principled uncertainty while preserving all partial observations. This proves paramount in challenging scenarios where every view contains complementary geometric information necessary for accurate 3D hand reconstruction.

4.5. Qualitative Analysis

Fig. 5 compares reconstruction results under three challenging real-world conditions: motion blur, severe occlusion, and missing views. As shown, baseline methods such as MVP and POEM suffer from noticeable artifacts, including surface distortions, unnatural articulations, and large geometric errors in degraded regions. In contrast, JUMP-Hand consistently recovers accurate and complete hand geometry across all scenarios. This superior robustness stems from our explicit uncertainty modeling, which detects and down-weights unreliable observations while emphasizing high-quality views, enabling graceful adaptation to varying observation quality without requiring scenario-specific tuning.

5. Conclusion and Future Work

In this paper, we propose JUMP-Hand, the first method to leverage probabilistic joint-wise uncertainty as an explicit gating signal for multi-view 3D hand reconstruction. By conceptualizing each view as an independent expert and quantifying its per-joint reliability, our method effectively overcomes the limitations of prior works that indiscriminately aggregate multi-view information. This principled gating mechanism unifies both the coarse triangulation and refinement stages, establishing state-of-the-art performance across three benchmarks, with particularly substantial gains in highly challenging scenarios.

Limitation and future work. While our Gaussian-based uncertainty modeling is computationally efficient and generally effective, its uni-modal nature may not fully capture the complex, multi-modal ambiguities that arise from severe occlusions. Future work will explore more expressive probabilistic representations to better resolve uncertainty under these extreme conditions.

Acknowledgment

This work is jointly supported by the National Natural Science Foundation of China (Grant No. 62271221), the National Social Science Foundation of China (Grant No. 25BXW041), and the National Research Foundation, Singapore under its Digital Trust Centre Innovation Grant (DTC Award No: DTC-IGC-02). This research is supported by the National Research Foundation, Singapore and Infocomm Media Development Authority under its Trust Tech Funding Initiative. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of the National Research Foundation, Singapore and Infocomm Media Development Authority.

References

[1] Adnane Boukhayma, Rodrigo de Bem, and Philip HS Torr. 3d hand shape and pose from images in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10843–10852, 2019. 2

[2] Samarth Brahmbhatt, Cusuh Ham, Charles C Kemp, and James Hays. Contactdb: Analyzing and predicting grasp contact via thermal imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8709–8719, 2019. 1

[3] Yu-Wei Chao, Wei Yang, Yu Xiang, Pavlo Molchanov, Ankur Handa, Jonathan Tremblay, Yashraj S Narang, Karl Van Wyk, Umar Iqbal, Stan Birchfield, et al. Dexycb: A benchmark for capturing hand grasping of objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9044–9053, 2021. 2, 6

[4] Geng Chen, Xufeng Jian, Yuchen Chen, Pengfei Ren, Jingyu Wang, Haifeng Sun, Qi Qi, Jing Wang, and Jianxin Liao. A 3-net: Calibration-free multi-view 3d hand reconstruction for enhanced musical instrument learning. In *Proceedings of the International Joint Conference on Artificial Intelligence*, 2025. 2

[5] Damai Dai, Chengqi Deng, Chenggang Zhao, R. X. Xu, Huazuo Gao, Deli Chen, Jiashi Li, Wangding Zeng, Xingkai Yu, Y. Wu, Zhenda Xie, Y. K. Li, Panpan Huang, Fuli Luo, Chong Ruan, Zhifang Sui, and Wenfeng Liang. Deepseekmoe: Towards ultimate expert specialization in mixture-of-experts language models, 2024. 2, 3

[6] Vandad Davoodnia, Saeed Ghorbani, Marc-André Carbonneau, Alexandre Messier, and Ali Etemad. Upose3d: Uncertainty-aware 3d human pose estimation with cross-view and temporal cues. In *Proceedings of the European Conference on Computer Vision*, pages 19–38. Springer, 2024. 3

[7] Charles Dugas, Yoshua Bengio, François Bélisle, Claude Nadeau, and René Garcia. Incorporating second-order functional knowledge for better option pricing. In *Proceedings of Advances in Neural Information Processing Systems*, 2000. 3

[8] William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022. 2, 3

[9] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the International Conference on Machine Learning*, pages 1050–1059. PMLR, 2016. 2, 3

[10] Shreyas Hampali, Sayan Deb Sarkar, and Vincent Lepetit. Ho-3d_v3: Improving the accuracy of hand-object annotations of the ho-3d dataset. *arXiv preprint arXiv:2107.00887*, 2021. 2, 5

[11] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003. 4

[12] Yana Hasson, Gul Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11807–11816, 2019. 2

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 6

[14] Karim Iskakov, Egor Burkov, Victor Lempitsky, and Yury Malkov. Learnable triangulation of human pose. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7718–7727, 2019. 2

[15] Changlong Jiang, Yang Xiao, Cunlin Wu, Mingyang Zhang, Jinghong Zheng, Zhiguo Cao, and Joey Tianyi Zhou. A2j-transformer: Anchor-to-joint transformer network for 3d interacting hand pose estimation from a single rgb image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8846–8855, 2023. 2

[16] Changlong Jiang, Yang Xiao, Jinghong Zheng, Haohong Kuang, Cunlin Wu, Mingyang Zhang, Zhiguo Cao, Min Du, Joey Tianyi Zhou, and Junsong Yuan. 3d hand pose estimation via articulated anchor-to-joint 3d local regressors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 2

[17] Zheheng Jiang, Hossein Rahmani, Sue Black, and Bryan M Williams. 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*, pages 758–767, 2023. 3

[18] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *Proceedings of Advances in Neural Information Processing Systems*, 2017. 2, 3, 5

[19] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014. 6

[20] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Proceedings of Advances in Neural Information Processing Systems*, 2017. 3

[21] Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam

Shazeer, and Zhifeng Chen. Gshard: Scaling giant models with conditional computation and automatic sharding, 2020. 3

[22] Lijun Li, Li'an Zhuo, Bang Zhang, Liefeng Bo, and Chen Chen. Diffhand: end-to-end hand mesh reconstruction via diffusion models. *arXiv preprint arXiv:2305.13705*, 2023. 3

[23] Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-end human pose and mesh reconstruction with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1954–1963, 2021. 2

[24] Franziska Mueller, Dushyant Mehta, Oleksandr Sotnychenko, Srinath Sridhar, Dan Casas, and Christian Theobalt. Real-time hand tracking under occlusion from an egocentric rgb-d sensor. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1154–1163, 2017. 1

[25] Javier Romero, Dimitrios Tzionas, and Michael J Black. Embodied hands: Modeling and capturing hands and bodies together. *arXiv preprint arXiv:2201.02610*, 2022. 5

[26] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer, 2017. 2, 3

[27] Adrian Spurr, Jie Song, Seonwook Park, and Otmar Hilliges. Cross-modal deep variational hand pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 89–98, 2018. 2

[28] Omid Taheri, Nima Ghorbani, Michael J Black, and Dimitrios Tzionas. Grab: A dataset of whole-body human grasping of objects. In *Proceedings of the European Conference on Computer Vision*, pages 581–600. Springer, 2020. 1

[29] Tze Ho Elden Tse, Franziska Mueller, Zhengyang Shen, Danhang Tang, Thabo Beeler, Mingsong Dou, Yinda Zhang, Sasa Petrovic, Hyung Jin Chang, Jonathan Taylor, et al. Spectral graphormer: Spectral graph-based transformer for egocentric two-hand reconstruction using multi-view color images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14666–14677, 2023. 2

[30] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of Advances in Neural Information Processing Systems*, 2017. 2

[31] Fu Xiong, Boshen Zhang, Yang Xiao, Zhiguo Cao, Taidong Yu, Joey Tianyi Zhou, and Junsong Yuan. A2j: Anchor-to-joint regression network for 3d articulated pose estimation from a single depth image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 793–802, 2019. 2

[32] Hang Xu, Yang Xiao, Changlong Jiang, Haohong Kuang, Kaidi Zhang, Min Du, and Ran Wang. Moeg-hoi: Mixture of expert groups for one-stage hand-object interaction motion generation with hand-finger-joint semantic guidance. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11223–11231, 2026. 3

[33] Jian Yang, Jiakun Li, Guoming Li, Huai-Yu Wu, Zhen Shen, and Zhaoxin Fan. Mlphand: real time multi-view 3d hand reconstruction via mlp modeling. In *Proceedings of the European Conference on Computer Vision*, pages 407–424. Springer, 2024. 2, 6

[34] Lixin Yang, Kailin Li, Xinyu Zhan, Fei Wu, Anran Xu, Liu Liu, and Cewu Lu. Oakink: A large-scale knowledge repository for understanding hand-object interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20953–20962, 2022. 2, 6

[35] Lixin Yang, Jian Xu, Licheng Zhong, Xinyu Zhan, Zhicheng Wang, Kejian Wu, and Cewu Lu. Poem: reconstructing hand in a point embedded multi-view stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21108–21117, 2023. 2, 5, 6

[36] Jianfeng Zhang, Yujun Cai, Shuicheng Yan, Jiashi Feng, et al. Direct multi-view multi-person 3d pose estimation. In *Proceedings of Advances in Neural Information Processing Systems*, pages 13153–13164, 2021. 2, 6

[37] Yufei Zhang, Jeffrey O Kephart, and Qiang Ji. Weakly-supervised 3d hand reconstruction with knowledge prior and uncertainty guidance. In *Proceedings of the European Conference on Computer Vision*, pages 106–125. Springer, 2024. 3

[38] Jinghong Zheng, Changlong Jiang, Yang Xiao, Jiaqi Li, Haohong Kuang, Hang Xu, Ran Wang, Zhiguo Cao, Min Du, and Joey Tianyi Zhou. Pandapose: 3d human pose lifting from a single image via propagating 2d pose prior to 3d anchor space. *arXiv preprint arXiv:2602.01095*, 2026. 2

[39] Yuxiao Zhou, Marc Habermann, Weipeng Xu, Ikhsanul Habibie, Christian Theobalt, and Feng Xu. 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*, pages 5346–5355, 2020. 2

[40] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 5