Learning Correlation-aware Aleatoric Uncertainty for 3D Hand Pose Estimation

Lee Chae-Yeon[1]
chaeyeon.lee@postech.ac.kr
Nam Hyeon-Woo[2]
hyeonw.nam@postech.ac.kr
Tae-Hyun Oh[3]
thoh.kaist.ac.kr@gmail.com

[1] Grad. School of AI
POSTECH, South Korea
[2] Dept. of Electrical Engineering
POSTECH, South Korea
[3] School of Computing
KAIST, South Korea

Abstract

3D hand pose estimation is a fundamental task in understanding human hands. However, accurately estimating 3D hand poses remains challenging due to the complex movement of hands, self-similarity, and frequent occlusions. In this work, we address two limitations: the inability of existing 3D hand pose estimation methods to estimate aleatoric (data) uncertainty, and the lack of uncertainty modeling that incorporates joint correlation knowledge, which has not been thoroughly investigated. To this end, we introduce aleatoric uncertainty modeling into the 3D hand pose estimation framework, aiming to achieve a better trade-off between modeling joint correlations and computational efficiency. We propose a novel parameterization that leverages a single linear layer to capture intrinsic correlations among hand joints. This is enabled by formulating the hand joint output space as a probabilistic distribution, allowing the linear layer to capture joint correlations. Our proposed parameterization is used as a task head layer, and can be applied as an add-on module on top of the existing models. Our experiments demonstrate that our parameterization for uncertainty modeling outperforms existing approaches. Furthermore, the 3D hand pose estimation model equipped with our uncertainty head achieves favorable accuracy in 3D hand pose estimation while introducing new uncertainty modeling capability to the model. The project page is available at https://hand-uncertainty.github.io/.

1 Introduction

Understanding human hands is fundamental for applications ranging from robotics to AR/VR [9, 19]. The ability to perceive and interpret human hands significantly enhances the dexterity of robot-assisted tasks and ensures seamless human-like robot-object interactions. In learning from human demonstrations, robots can acquire human hand behavior by observing these demonstrations [13, 19, 44]. In this work, we address two key limitations found in state-of-the-art methods for estimating 3D hand pose.

(1) Inability to estimate the aleatoric (data) uncertainty. While recent works in learning-based 3D hand pose estimation [20, 28, 30, 35, 42, 53, 54, 55, 60] have made significant progress, accurate 3D hand-pose estimation inherently faces uncertainty from in-the-wild video. These uncertainty arise from the high number of degrees of freedom present in

hands [2], frequent occurrence of self-similarity and occlusion [37, 49], and motion blur due to their dynamic nature [39, 41]. Against these observation noises, quantifying aleatoric uncertainty enhances the confidence of the hand estimation models to be deployed in real-world applications. *(2) Uncertainty modeling in the absence of joint correlation knowledge.* Human hand offers many individual degrees of freedom, yet joint movements are correlated [17, 23, 33, 34, 47]. Although the uncertainty of one hand joint can influence the uncertainty of another joint, previous works [27, 38] model uncertainty entry-wise independently under the independent assumption due to computational and parameter efficiency.

In this work, we introduce aleatoric uncertainty modeling into the 3D hand pose estimation framework, achieving a better trade-off between correlation modeling and efficiency. We propose a novel parametrization that leverages a single linear layer to capture the intrinsic correlations among hand joints. To enable this, we design a probabilistic hand joint output space that facilitates uncertainty modeling with consideration of joint dependencies. Specifically, we begin by training a pre-trained large model [42] for aleatoric uncertainty estimation by adding a transformer head that regresses the per-joint variance. We use a Gaussian negative log-likelihood loss with a diagonal covariance matrix, which enables the network to predict variances that represent the uncertainty of hand joints under the independence assumption. The estimated uncertainty defines a probabilistic output space, from which we draw samples and feed them into a single linear layer to transform the output space into a correlation-aware space. Our parametrization serves as a mid-representation between diagonal and full covariance matrix parametrizations. It provides higher expressiveness for capturing joint correlations than the diagonal form, yet requires significantly fewer parameters than the full covariance parametrization.

We demonstrate the effectiveness of our method on two standard benchmarks for 3D hand pose estimation, FreiHAND [59] and HO3Dv2 [18]. Our method outperforms existing aleatoric uncertainty modeling methods on uncertainty estimation. The key to our method's effectiveness is the formulation of the hand joint output space as a probabilistic distribution, which enables the linear layer to effectively learn hand joint correlations. This approach allows for an analytic representation of a structured covariance matrix, facilitating direct estimation of uncertainty. Furthermore, our method maintains competitive accuracy in 3D hand pose estimation, demonstrating that modeling uncertainty does not compromise pose estimation performance. Our main contributions are summarized as follows:

- We introduce aleatoric uncertainty modeling into 3D hand pose estimation framework.
- We propose a novel parameterization that leverages a single linear layer to effectively model inherent hand joint correlations, which is enabled by formulating the hand joint output space as a probabilistic distribution.
- Comprehensive experiments demonstrate that our proposed method significantly outperforms existing aleatoric uncertainty modeling methods in uncertainty estimation, while maintaining accurate 3D hand pose estimation performance.

2 Related Work

3D hand pose estimation. 3D hand pose estimation from a single RGB image has received a great attention for understanding complex hand interaction. The advent of deep learning has especially improved the performance of 3D hand pose estimation compared to hand-crafted geometric features. Most existing works [1, 14, 42, 46, 55] leverage the MANO parametric

hand model [45] and regress the hand pose and shape parameters directly from an RGB image. Other works [10, 16, 29, 30] follow a non-parametric approach and regress mesh vertex coordinates directly from images for a more fine-grained reconstruction of hand surfaces. Another line of work [4, 28, 35, 37] infer 3D hand joint positions of x, y and z, which serve as a skeletal representation of hand posture. More recently, HaMeR [42] exploits transformer networks [50] and train on a large dataset, achieving robust 3D hand reconstruction and strong generalization to in-the-wild images. Our work addresses aleatoric uncertainty in 3D hand pose estimation, which has been underexplored.

Uncertainty in deep learning. Uncertainty in deep learning can be categorized into aleatoric and epistemic uncertainties [11]. Aleatoric uncertainty is attributed to the non-deterministic nature known as data uncertainty. Epistemic uncertainty arises from model uncertainty due to insufficient knowledge learned from the training data. In this work, we focus on aleatoric uncertainty in 3D hand joint positions. We assume that the uncertainty is heteroscedastic [26], indicating that it depends on the inputs to the model, as certain inputs may inherently have higher uncertainty than others. A commonly used approach is to estimate the probability distribution over the output, and train the network by minimizing the negative log-likelihood (NLL) of the ground truth [3, 26, 56]. Caramalau *et al*. [6] demonstrates that jointly modeling aleatoric and epistemic uncertainties is effective when applying an active learning framework to estimating 3D hand pose from a single depth image. However, existing approaches for estimating 3D hand pose from a single RGB image either overlook the heteroscedastic uncertainty in input images or rely on latent distributions to implicitly capture aleatoric uncertainty through sampling [51, 60], and they are not publicly available. Zhang *et al*. [57] applies the NLL loss for 3D hand reconstruction to model output-space uncertainty; however, their method is limited to modeling uncertainty in 2D hand joint positions, which may not accurately reflect the uncertainty in 3D joint space. AMVUR [24] adopts a probabilistic framework for 3D hand pose and shape estimation, inherently supporting uncertainty quantification in 3D hand joint positions. In this work, we explicitly model joint-wise uncertainty in the output space by learning a Gaussian distribution over 3D hand joint positions and incorporating it into the training objective through a simple yet effective NLL loss. Furthermore, we model inter-joint dependencies using a single linear transformation, enabling an analytic formulation of the structured uncertainty without relying on sampling-based approximations.

3 Method

3.1 Preliminary

Notation. We denote $\mathbf{x} \in \mathbb{R}^{d_i}$ as the input image, $f : \mathbb{R}^{d_i} \to \mathbb{R}^{d_f}$ as the feature extractor, $g : \mathbb{R}^{d_f} \to \mathbb{R}^{d_o}$ as the parameter regressor, and $\text{diag}(\cdot)$ outputs a diagonal matrix given a vector. As shown in Fig. 1a, the model without uncertainty modeling is denoted as $g(f(\mathbf{x}))$. The uncertainty modeling [27] modifies the regressor g such that $g(f(\mathbf{x})) \sim \mathcal{N}(\mu, \Sigma)$ where $\mathcal{N}$ is the Gaussian distribution with a mean $\mu \in \mathbb{R}^{d_o}$ and a covariance $\Sigma \in \mathbb{R}^{d_o \times d_o}$. The subsequent paragraphs describe two approaches, categorized based on modeling the covariance matrix Σ: diagonal and full covariance matrix.

Diagonal Covariance Matrix. The most widely used way for uncertainty modeling is a diagonal covariance matrix. It is parameter-efficient because the regressor only needs to output the mean and diagonal variance vector. The diagonal Gaussian modeling needs twice the output dimensions, *i.e*., $g_{diag} : \mathbb{R}^{d_f} \to \mathbb{R}^{2 \cdot d_o}$ (See Fig. 1b). The half dimension is for the

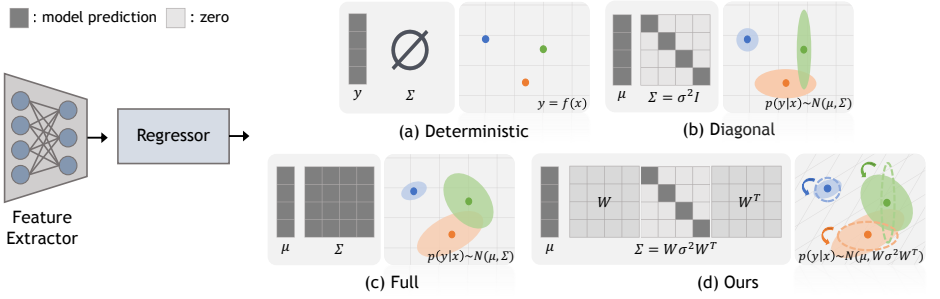

Figure 1: **Illustration of deterministic modeling and correlation modelings of uncertainty.** (a) Deterministic modeling produces a single deterministic output, represented as point embedding in the output space; (b) Diagonal and (c) Full correlation modeling of output produce means and covariances of a Gaussian distribution, where the uncertainty of the prediction is modeled by its variance. (d) Ours learns the variation of each dimension in the output space and shared weight $\mathbf{W}$ which captures intrinsic hand joint dependencies, so that our covariance matrix is represented by $\Sigma = \mathbf{W}\sigma^2\mathbf{W}^\top$.

mean vector, $\mu \in \mathbb{R}^{d_o}$, and the other half is for the variance vector, $\sigma^2 \in \mathbb{R}^{d_o}$ as follows:

$$g_{diag}(f(\mathbf{x})) \sim \mathcal{N}(\mu, \mathrm{diag}(\sigma^2)). \tag{1}$$

In implementation, instead of using direct σ^2, the logarithm of σ^2 is used with the exponential function mapping to ensure the positive values. The model is trained by minimizing the negative log-likelihood (NLL) loss with ground truth $\mathbf{y}$ as

$$\mathcal{L}_{NLL} = \log p(\mathbf{y}|\mu, \sigma^2) = \frac{||\mathbf{y}-\mu||^2}{2\sigma^2} + 0.5 \log \sigma^2. \tag{2}$$

Full Covariance Matrix. In a full covariance matrix, the regressor outputs all covariance matrix elements as

$$g_{full}(f(\mathbf{x})) \sim \mathcal{N}(\mu, \Sigma), \tag{3}$$

where $\mu \in \mathbb{R}^{d_o}$ and $\Sigma \in \mathbb{R}^{d_o \times d_o}$ (See Fig. 1c). Since Σ should be a positive definite matrix, we construct Σ as $\mathbf{A}\mathbf{A}^\top$, which ensures Σ to be positive definite. Although it is less parameter-efficient compared to the diagonal covariance matrix, this approach has a higher capacity for capturing correlations. The model is also trained by minimizing the NLL loss as follows:

Parametrization	Uncertainty modeling	# Params.	Example [# Params]
Deterministic	✗	$d_f d_o$	0.065M
Diagonal covariance	✓(independent)	$2d_f d_o$	0.129M
Full covariance	✓	$d_f(d_o+d_o^2)$	4.129M
Ours	✓	$2d_f d_o + d_o^2$	0.133M

Table 1: **The number of parameters.** The dimensions of feature and output are d_f and d_o, respectively. We list the required number of parameters for each parameterization. We set d_f=1024, d_o=63, and k=21 for example.

$$\mathcal{L}_{NLL} = \log p(\mathbf{y}|\mu, \Sigma) = \tfrac{1}{2}(\mathbf{y}-\mu)^\top \Sigma^{-1}(\mathbf{y}-\mu) + \tfrac{1}{2}\log\det\Sigma. \tag{4}$$

The objective function involves the inverse covariance matrix. This introduces optimization instability. Thus, it is more stable to directly estimate the precision matrix as $\Psi = \Sigma^{-1}$.

In summary, the diagonal covariance matrix is parameter-efficient but does not capture the correlation; the full covariance matrix has opposite properties. We propose a new mid-representation that leverages a single linear layer which is parameter-efficient and effectively captures the hand joint correlation in a probabilistic manner.

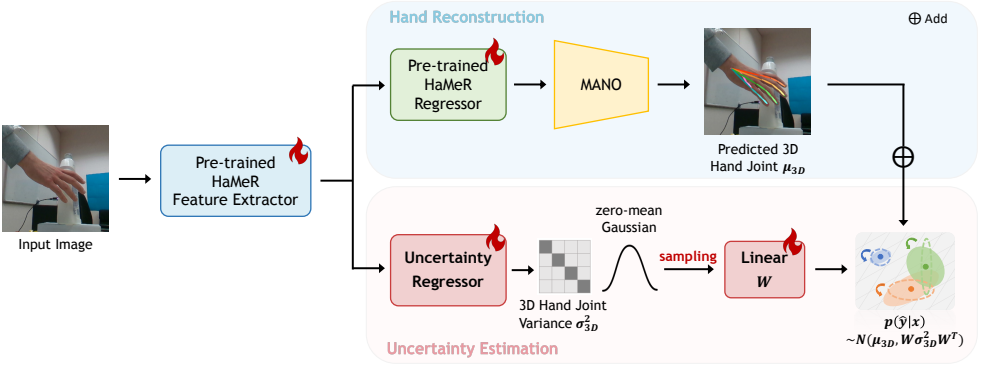

Figure 2: **Pipeline of our proposed method.** We train a pre-trained large model [42] by introducing an additional transformer head that estimates the variance for each joint under the independent assumption. This estimated uncertainty then defines a probabilistic hand joint output space, from which we sample and pass the samples through a single linear layer to model the correlations between hand joints.

3.2 Correlation-Aware Aleatoric Uncertainty Estimation in Hand Joints

Our goal is to estimate the aleatoric uncertainty of hand joints with the incorporation of joint correlation knowledge in an efficient yet expressive manner. The key idea is to define a probabilistic output space based on per-joint uncertainties learned under an independence assumption, and then transform it into a correlation-aware space using a single linear layer.

Aleatoric Uncertainty Estimation with Diagonal Covariance Matrix. We adapt HaMeR [42], a recently proposed transformer-based model for hand pose estimation pretrained on large-scale datasets to enable uncertainty modeling. In HaMeR, the Vision Transformer (ViT) [12] backbone serves as a feature extractor and the mean 3D hand joint positions $\mu_{\mathbf{3D}}$ are regressed by a transformer head followed by the MANO [45]. To enable uncertainty modeling, we incorporate an additional transformer head, denoted as uncertainty regressor, to regress the variance of each joint $\sigma^2_{\mathbf{3D}}$. The model is trained by minimizing the NLL loss (Eq. (2)) with ground truth 3D hand joint positions $\mathbf{y}_{\mathbf{3D}}$. Up to here, the uncertainty modeling is same with the diagonal covariance matrix.

Probabilistic Hand Joint Output Space. We define a probabilistic hand joint output space under a zero-mean Gaussian assumption, where per-joint variances learned under an independence assumption serve as the diagonal elements of the covariance matrix as $p(\mathbf{z}|\mathbf{x}) \sim \mathcal{N}(\mathbf{0}, \mathrm{diag}(\sigma^2_{\mathbf{3D}}))$. We draw N samples from $p(\mathbf{z}|\mathbf{x})$ and feed-forward these samples to single linear layer of weights $\mathbf{W}$, ensuring that the final output follows $p(\hat{\mathbf{y}}|\mathbf{z}) \sim \mathcal{N}(\mathbf{0}, \mathbf{W}\,\mathrm{diag}(\sigma^2_{\mathbf{3D}})\mathbf{W}^\top)$ by the linearity of the Gaussian. By adopting the zero-mean assumption, we remove the dependency between the weights $\mathbf{W}$ and the mean $\mu_{\mathbf{3D}}$, enabling the model to focus solely on capturing correlations. Afterwards, the mean 3D hand joint positions $\mu_{\mathbf{3D}}$ are added to this output space. As a result, we obtain:

$$p(\hat{\mathbf{y}}|\mathbf{z}) \sim \mathcal{N}(\mu_{\mathbf{3D}}, \mathbf{W}\,\mathrm{diag}(\sigma^2_{\mathbf{3D}})\mathbf{W}^\top). \tag{5}$$

We minimize the mean squared error (MSE) between the predictions and the ground truth:

$$\mathcal{L}_{MSE} = \mathbb{E}||\mathbf{y} - \hat{\mathbf{y}}||_2^2, \tag{6}$$

where $\mathbb{E}$ is the expectation. The overall training loss is defined as follows:

$$\mathcal{L} = \mathcal{L}_{DETER} + \lambda_{NLL}\mathcal{L}_{NLL} + \lambda_{MSE}\mathcal{L}_{MSE}, \tag{7}$$

where $\mathcal{L}_{DETER}$ denotes the loss functions used for deterministic 3D hand pose estimation in HaMeR [42], and λ_{NLL} and λ_{MSE} are the weight factors for the uncertainty modeling terms.

By sampling from the probabilistic hand joint output space and applying a single linear transformation, we naturally capture the correlation between hand joints, which can also be formulated analytically. We position our method as a mid-representation between diagonal and full covariance modeling. First, it offers higher expressiveness than diagonal covariance modeling in capturing inherent correlations between hand joints. By adopting probabilistic space, our output distribution models hand joint correlation structure as full covariance modeling. Second, our method requires less number of parameters compared to full covariance modeling, which involves the square of output dimension—making it impractical for high-dimensional outputs and prone to optimization instability. In contrast, our approach introduces only a single linear transformation matrix on top of diagonal covariance modeling, providing both computational efficiency (see Table 1) and improved optimization stability.

4 Experiments

4.1 Experimental Setup

Datasets. We train our method on 2.7M training examples from multiple datasets that provide 2D or 3D hand annotations as in HaMeR [42]. This includes FreiHAND [59], HO3D [18], MTC [52], RHD [58], InterHand2.6M [37], H2O3D [18], DEX YCB [6], COCO WholeBody [25], Halpe [15] and MPII NZSL [48]. To evaluate the quality of the estimated uncertainty and 3D hand pose estimation accuracy of our method, we use two standard benchmarks for 3D hand pose estimation, FreiHAND [59] and HO3Dv2 [18] which provide ground truth 3D hand annotations.

Metrics on 3D hand pose estimation. We follow the typical protocols used in previous works [32, 40, 42], and report PA-MPJPE and AUC_J for evaluating estimated 3D hand joints and PA-MPVPE, AUC_V, F@5mm and F@15mm for evaluating estimated 3D hand mesh. PA-MPJPE and PA-MPVPE are measured in mm.

Metrics on uncertainty estimation. To evaluate the quality of the estimated uncertainty, we use sparsification curves [22, 43]. The predicted hand joints are sorted based on the estimated uncertainty. Given an error metric ε, we evaluate the top x% most certain joints. Ideally, if the uncertainty estimates are well-calibrated, the error is supposed to decrease as uncertain predictions are removed. We vary x from 2 to 100, incrementing by 2, and report the area under the sparsification curve (AUSC) as in previous work [21]. This metric is affected by both prediction accuracy and how similar the uncertainty-based sorting is to the actual error-based

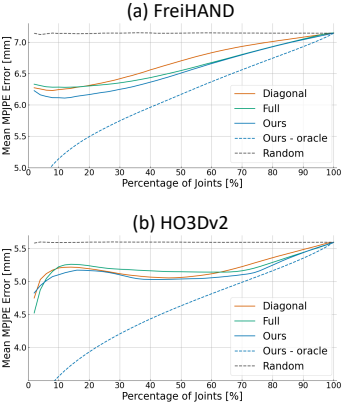

Figure 3: **Sparsification curves.** We compare sparsification curves obtained by different methods of estimating the uncertainty of 3D hand joints.

Method	FreiHAND [59]			HO3Dv2 [18]		
	AUSC ↓	AUSE ↓	Pearson's ρ ↑	AUSC ↓	AUSE ↓	Pearson's ρ ↑
Diagonal	655	54.6	0.393	511	63.3	0.448
Full	648	47.6	0.453	512	64.3	0.493
Ours	**642**	**42.2**	**0.569**	**505**	**57.6**	**0.600**

Table 2: **Quantitative evaluation on uncertainty estimation.** We demonstrate the effectiveness of our parametrization in consistently enhancing all three metrics.

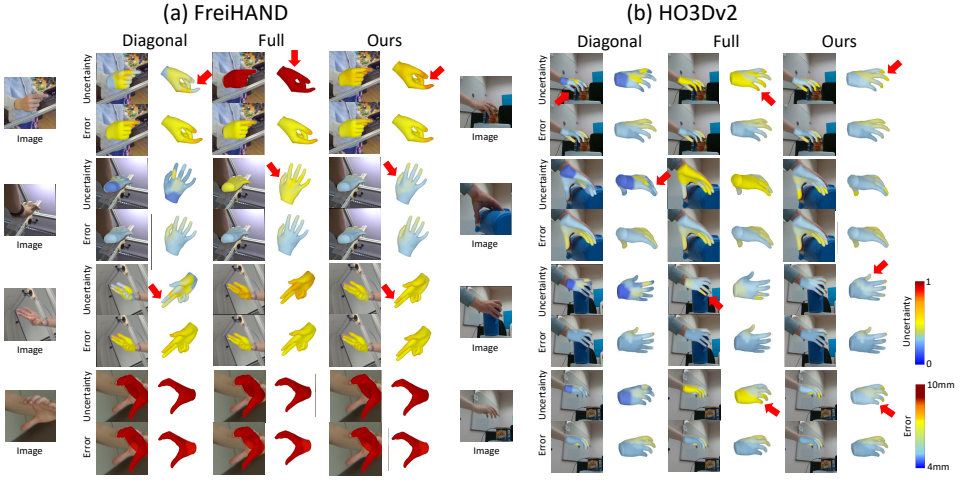

Figure 4: **Qualitative results of uncertainty estimation.** We evaluate the quality of uncertainty estimates by comparing our method with existing uncertainty modeling methods. Specifically, we visualize the prediction errors alongside the corresponding uncertainty values. A desirable property of an uncertainty estimator is its proportionality to the actual prediction error—i.e., higher uncertainty values should correspond to higher prediction errors. The uncertainty estimated by our method shows stronger correlation with the prediction error, indicating its effectiveness in capturing model confidence.

sorting. To only evaluate the latter, we also report the area under the sparsification error (AUSE) [22] by subtracting the oracle sparsification, which is obtained by sorting joints based on the ε magnitude, from the estimated sparsification curve. We assume MPJPE as ε. The uncertainty of each joint is measured as the trace of its estimated covariance matrix. Additionally, we report Pearson's correlation coefficient, ρ, to quantify the degree to which the predicted uncertainty correlates with the true error. Pearson's ρ measures the strength and direction of the linear relationship between two continuous variables, providing a value between -1 and 1.

Baselines. We implement two conventional uncertainty modeling methods based on Gaussian negative log-likelihood (NLL) as baselines and compare them with our method. *(1) Diagonal:* The uncertainty regressor outputs per-joint variances under the independent assumption; the uncertainty is modeled using a diagonal covariance matrix. *(2) Full:* The uncertainty regressor outputs all elements of the covariance matrix, modeling the uncertainty with a full covariance structure (see Sec. 3.1).

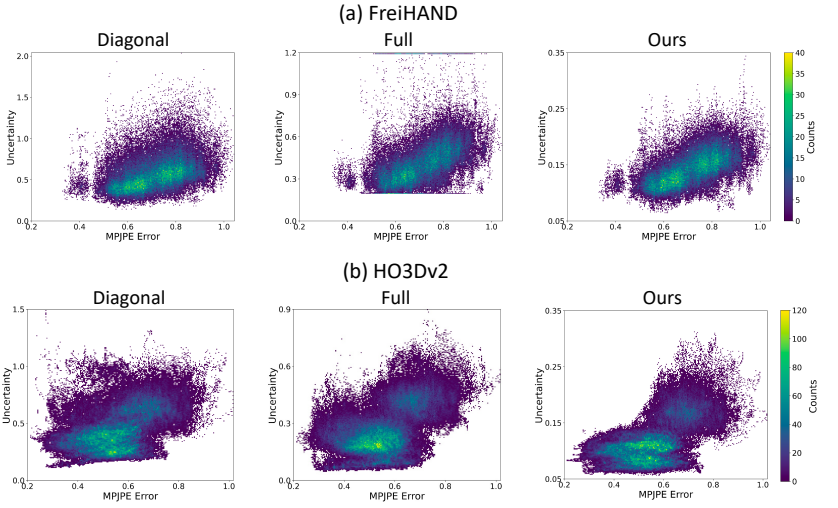

Figure 5: **2D histograms of model error and estimated uncertainty.** We show 2D histograms of the x-axis representing the model's MPJPE error and the y-axis representing the estimated uncertainty on (a) FreiHAND [59] and (b) HO3Dv2 [18] datasets.

Implementation details. In all experiments, including both the baselines and our method, we set $N = 25$, $\lambda_{NLL} = 5e\text{-}4$. We additionally set $\lambda_{MSE} = 5e\text{-}4$ in our method. We train the models on a single NVIDIA A100 GPU for 550k iterations using the AdamW optimizer with a weight decay of $1e\text{-}4$, $\beta_1 = 0.9$, and $\beta_2 = 0.999$. The learning rate is initialized as $1e\text{-}6$, and the mini-batch size is set to 64. All other experimental settings follow those used in HaMeR [42].

4.2 Uncertainty Estimation

We conduct evaluations to assess the quality of the estimated uncertainty. The quantitative comparison with existing uncertainty modeling methods is presented in Table 2. The results demonstrate that our method outperforms the baselines across all metrics. The sparsification curves in Fig. 3 show that although all methods perform similarly when evaluated on all 3D joints, our method achieves significantly higher accuracy than the others as 3D joints with high uncertainty are removed. This indicates that our uncertainty better reflects true 3D joint prediction errors. We additionally validate this by a qualitative comparision in Fig. 4, wherein our uncertainty correlates better with the prediction error. Furthermore, we present qualitative results of the Pearson correlation by visualizing a 2D histogram of the model's MPJPE error and its estimated uncertainty. As shown in Fig. 5, the uncertainty estimates from our proposed method correlate with the model's error better than those from the *Diagonal* and *Full* baselines, while exhibiting fewer outliers.

4.3 3D Hand Pose Estimation

We evaluate the 3D hand pose estimation capability of our method. Table 3 compares our approach with existing 3D hand pose estimation methods that do not support uncertainty estimation, as well as with uncertainty modeling baselines. For a fair comparison, we also

Method	FreiHAND [57]				HO3Dv2 [18]					
	PA-MPJPE ↓	PA-MPVPE ↓	F@5 ↑	F@15 ↑	AUC_J ↑	PA-MPJPE ↓	AUC_V ↑	PA-MPVPE ↓	F@5 ↑	F@15 ↑
I2L-MeshNet [40]	7.4	7.6	0.681	0.973	0.775	11.2	0.722	13.9	0.409	0.932
Pose2Mesh [11]	7.7	7.8	0.674	0.969	0.754	12.5	0.749	12.7	0.441	0.909
I2UV-HandNet [8]	6.7	6.9	0.707	0.977	0.804	9.9	0.799	10.1	0.500	0.943
METRO [31]	6.5	6.3	0.731	0.984	0.792	10.4	0.779	11.1	0.484	0.946
MobRecon [3]	**5.7**	5.8	0.784	0.986	-	9.2	-	9.4	0.538	0.957
AMVUR [29]	6.2	6.1	0.767	0.987	0.835	8.3	0.836	8.2	0.608	0.965
HaMeR [35]	6.0	**5.7**	**0.785**	**0.990**	0.846	7.7	0.841	7.9	0.635	0.980
Deterministic	6.1	**5.7**	**0.785**	**0.990**	0.845	7.8	0.840	8.0	0.629	0.980
Diagonal	6.1	5.8	0.782	**0.990**	0.847	7.7	0.842	7.9	0.638	**0.981**
Full	6.1	5.9	0.773	0.989	**0.849**	**7.6**	**0.844**	**7.8**	**0.644**	**0.981**
Ours	6.0	**5.7**	0.784	**0.990**	0.847	**7.6**	0.843	7.9	0.640	**0.981**

Table 3: **Quantitative evaluation on 3D hand pose estimation.** Our method maintains comparable performance compared to 3D hand pose estimation methods and uncertainty modeling baselines.

Method	FreiHAND [57]			HO3Dv2 [18]		
	AUSC ↓	AUSE ↓	Pearson's ρ ↑	AUSC ↓	AUSE ↓	Pearson's ρ ↑
Ours w/o linear layer	648	47.9	0.544	509	61.4	0.524
Ours	**642**	**42.2**	**0.569**	**505**	**57.6**	**0.600**

Table 4: **Ablation study on joint correlation in uncertainty estimation.** We show that our joint correlation modeling consistently improves uncertainty estimation performance across all three metrics.

Method	FreiHAND [57]				HO3Dv2 [18]					
	PA-MPJPE ↓	PA-MPVPE ↓	F@5 ↑	F@15 ↑	AUC_J ↑	PA-MPJPE ↓	AUC_V ↑	PA-MPVPE ↓	F@5 ↑	F@15 ↑
Ours w/o linear layer	6.1	5.8	0.782	**0.990**	**0.847**	7.7	**0.843**	**7.9**	0.638	**0.981**
Ours	**6.0**	**5.7**	**0.784**	**0.990**	**0.847**	**7.6**	**0.843**	**7.9**	**0.640**	**0.981**

Table 5: **Ablation study on joint correlation in 3D hand pose estimation.** We demonstrate that our joint correlation modeling shows improvement in 3D hand pose estimation accuracy.

assess our method against a deterministic baseline denoted as *Deterministic*, which fine-tunes the pre-trained network without incorporating uncertainty modeling. Compared to 3D hand pose estimation methods and deterministic baseline, our model achieves competitive performance on both benchmarks, while additionally providing uncertainty estimation. In the comparison of uncertainty modeling baselines, although the full baseline achieves slightly better 3D hand pose estimation performance on the HO3Dv2 [18] evaluation dataset, our method offers far more reliable uncertainty estimates through efficient joint correlation modeling (see Tables 1 & 2).

4.4 Ablation Study

Ablation study on joint correlation. We conduct an ablation study to investigate whether modeling joint correlations using a linear layer improves uncertainty estimation. Specifically, we remove the linear layer, resulting in a probabilistic hand joint output space constructed under an independence assumption, where uncertainty is modeled with per-joint variance. The quantitative results in Tables 4 and 5 support our conclusion that incorporating the linear layer to capture joint correlations enhances the performance of uncertainty estimation and 3D hand pose estimation.

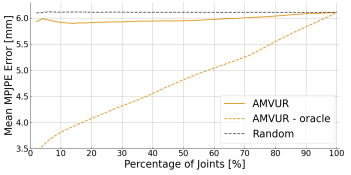

Figure 6: **Sparsification curve of AMVUR [24].** The competing method exhibits a large discrepancy between uncertainty-based and error-based sortings.

Method	AUSC ↓	AUSE ↓	Pearson's ρ ↑
AMVUR [24]	586	113	0.127
Ours	**505**	**57.6**	**0.600**

Table 6: **Quantitative comparison with AMVUR [24].** Our method demonstrates consistently better performance than the competing method across all uncertainty evaluation metrics.

# Samples	FreiHAND [65]			HO3Dv2 [18]		
	AUSC ↓	AUSE ↓	Pearson's ρ ↑	AUSC ↓	AUSE ↓	Pearson's ρ ↑
1	653	52.7	0.399	507	58.6	0.557
5	655	54.3	0.414	510	62.4	0.484
10	648	47.1	0.551	507	59.1	0.536
25	**642**	**42.2**	**0.569**	**505**	**57.6**	**0.600**

Table 7: **Effect of the number of samples on uncertainty estimation.** We change the number of samples in output space from 1 to 25 and evaluate the quality of estimated uncertainty.

Comparison with related work. We compare the uncertainty estimation quality of our method on the HO3Dv2 [18] dataset with an open-sourced competing method [24], which models the 3D hand joint output space as a probabilistic distribution and inherently supports uncertainty measurement. As shown in Table 6, our method outperforms the competing approach in terms of uncertainty estimation performance. Furthermore, the sparsification curve of AMVUR [24] in Fig. 6 indicates that it struggles to align uncertainty-based sorting with actual error-based sorting.

Ablation study on number of samples. Our method draws the samples in the output space and feed-forward these samples to the regressor. We conduct an ablation study to investigate the effect of the number of samples on uncertainty estimation by varying this number. Table 7 presents the results. In general, the larger the number of samples, the better the performance. We select the default number as 25, considering the trade-off between performance and computational complexity.

5 Conclusion

This paper addresses two key challenges in existing 3D hand pose estimation models: the inability to estimate aleatoric uncertainty, and the lack of uncertainty modeling that incorporates joint correlation knowledge. We estimate and evaluate the aleatoric uncertainty in 3D hand pose estimation. We propose a new parametrization that leverages a single linear layer, which effectively captures the inherent hand joint correlation and achieves a better trade-off between modeling joint correlations and computational efficiency. To enable this, we formulate a probabilistic hand joint output space and then transform it into the correlation-aware space using the single linear layer. Experimental results show that the estimated aleatoric uncertainty of our method correlates well with the prediction error while maintaining 3D hand pose estimation performance.

Acknowledgments This research was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.RS-2025-25443318, Physically-grounded Intelligence: A Dual Competency Approach to Embodied AGI through Constructing and Reasoning in the Real World; No.RS-2019-II191906, Artificial Intelligence Graduate School Program(POSTECH)), and Culture, Sports and Tourism R&D Program through the Korea Creative Content Agency grant funded by the Ministry of Culture, Sports and Tourism in 2024 (Project Name: Development of barrier-free experiential XR contents technology to improve accessibility to online activities for the physically disabled, Project Number: RS-2024-00396700, Contribution Rate: 20%). It was also supported by the KAIST Cross-Generation Collaborative Lab Project, and the 'Ministry of Science and ICT' and NIPA ("HPC Support" Project).

References

[1] Seungryul Baek, Kwang In Kim, and Tae-Kyun Kim. 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*, pages 1067–1076, 2019.

[2] Sara Bilal, Rini Akmeliawati, Momoh Jimoh El Salami, and Amir A Shafie. Vision-based hand posture detection and recognition for sign language—a study. In *2011 4th International Conference on Mechatronics (ICOM)*, pages 1–6. IEEE, 2011.

[3] Lennart Bramlage, Michelle Karg, and Cristóbal Curio. Plausible uncertainties for human pose regression. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15133–15142, 2023.

[4] Yujun Cai, Liuhao Ge, Jianfei Cai, and Junsong Yuan. Weakly-supervised 3d hand pose estimation from monocular rgb images. In *Proceedings of the European conference on computer vision (ECCV)*, pages 666–682, 2018.

[5] Razvan Caramalau, Binod Bhattarai, and Tae-Kyun Kim. Active learning for bayesian 3d hand pose estimation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3419–3428, 2021.

[6] 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.

[7] Ping Chen, Yujin Chen, Dong Yang, Fangyin Wu, Qin Li, Qingpei Xia, and Yong Tan. I2uv-handnet: Image-to-uv prediction network for accurate and high-fidelity 3d hand mesh modeling. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12929–12938, 2021.

[8] Xingyu Chen, Yufeng Liu, Yajiao Dong, Xiong Zhang, Chongyang Ma, Yanmin Xiong, Yuan Zhang, and Xiaoyan Guo. Mobrecon: Mobile-friendly hand mesh reconstruction from monocular image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20544–20554, 2022.

[9] Seunggeun Chi, Pin-Hao Huang, Enna Sachdeva, Hengbo Ma, Karthik Ramani, and Kwonjoon Lee. Estimating ego-body pose from doubly sparse egocentric video data. *arXiv preprint arXiv:2411.03561*, 2024.

[10] Hongsuk Choi, Gyeongsik Moon, and Kyoung Mu Lee. Pose2mesh: Graph convolutional network for 3d human pose and mesh recovery from a 2d human pose. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16*, pages 769–787. Springer, 2020.

[11] Armen Der Kiureghian and Ove Ditlevsen. Aleatory or epistemic? does it matter? *Structural safety*, 31(2):105–112, 2009.

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[13] Yuming Du, Philippe Weinzaepfel, Vincent Lepetit, and Romain Brégier. Multi-finger grasping like humans. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1564–1570. IEEE, 2022.

[14] Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J Black, and Otmar Hilliges. Arctic: A dataset for dexterous bimanual hand-object manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12943–12954, 2023.

[15] Hao-Shu Fang, Jiefeng Li, Hongyang Tang, Chao Xu, Haoyi Zhu, Yuliang Xiu, Yong-Lu Li, and Cewu Lu. Alphapose: Whole-body regional multi-person pose estimation and tracking in real-time. *IEEE transactions on pattern analysis and machine intelligence*, 45(6):7157–7173, 2022.

[16] Liuhao Ge, Zhou Ren, Yuncheng Li, Zehao Xue, Yingying Wang, Jianfei Cai, and Junsong Yuan. 3d hand shape and pose estimation from a single rgb image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10833–10842, 2019.

[17] Charlotte Häger-Ross and Marc H Schieber. Quantifying the independence of human finger movements: comparisons of digits, hands, and movement frequencies. *Journal of Neuroscience*, 20(22):8542–8550, 2000.

[18] Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. Honnotate: A method for 3d annotation of hand and object poses. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3196–3206, 2020.

[19] Ankur Handa, Karl Van Wyk, Wei Yang, Jacky Liang, Yu-Wei Chao, Qian Wan, Stan Birchfield, Nathan Ratliff, and Dieter Fox. Dexpilot: Vision-based teleoperation of dexterous robotic hand-arm system. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9164–9170. IEEE, 2020.

[20] 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.

[21] Xiaoyan Hu and Philippos Mordohai. A quantitative evaluation of confidence measures for stereo vision. *IEEE transactions on pattern analysis and machine intelligence*, 34 (11):2121–2133, 2012.

[22] Eddy Ilg, Ozgun Cicek, Silvio Galesso, Aaron Klein, Osama Makansi, Frank Hutter, and Thomas Brox. Uncertainty estimates and multi-hypotheses networks for optical flow. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 652–667, 2018.

[23] James N Ingram, Konrad P Körding, Ian S Howard, and Daniel M Wolpert. The statistics of natural hand movements. *Experimental brain research*, 188:223–236, 2008.

[24] 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.

[25] Sheng Jin, Lumin Xu, Jin Xu, Can Wang, Wentao Liu, Chen Qian, Wanli Ouyang, and Ping Luo. Whole-body human pose estimation in the wild. In *European Conference on Computer Vision*, pages 196–214. Springer, 2020.

[26] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017.

[27] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *NIPS*, 2017.

[28] Dong Uk Kim, Kwang In Kim, and Seungryul Baek. End-to-end detection and pose estimation of two interacting hands. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11189–11198, 2021.

[29] Dominik Kulon, Riza Alp Guler, Iasonas Kokkinos, Michael M Bronstein, and Stefanos Zafeiriou. Weakly-supervised mesh-convolutional hand reconstruction in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4990–5000, 2020.

[30] Mengcheng Li, Liang An, Hongwen Zhang, Lianpeng Wu, Feng Chen, Tao Yu, and Yebin Liu. Interacting attention graph for single image two-hand reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2761–2770, 2022.

[31] 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.

[32] Kevin Lin, Lijuan Wang, and Zicheng Liu. Mesh graphormer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12939–12948, 2021.

[33] Yuan Liu, Li Jiang, Dapeng Yang, and Hong Liu. Analysis of hand and wrist postural synergies in tolerance grasping of various objects. *PloS one*, 11(8):e0161772, 2016.

[34] Yuan Liu, Bo Zeng, Ting Zhang, Li Jiang, Hong Liu, and Dong Ming. Quantitative investigation of hand grasp functionality: hand joint motion correlation, independence, and grasping behavior. *Applied Bionics and Biomechanics*, 2021(1):2787832, 2021.

[35] Hao Meng, Sheng Jin, Wentao Liu, Chen Qian, Mengxiang Lin, Wanli Ouyang, and Ping Luo. 3d interacting hand pose estimation by hand de-occlusion and removal. In *European Conference on Computer Vision*, pages 380–397. Springer, 2022.

[36] Gyeongsik Moon and Kyoung Mu Lee. 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*, pages 752–768. Springer, 2020.

[37] Gyeongsik Moon, Shoou-I Yu, He Wen, Takaaki Shiratori, and Kyoung Mu Lee. Interhand2. 6m: A dataset and baseline for 3d interacting hand pose estimation from a single rgb image. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16*, pages 548–564. Springer, 2020.

[38] Seong Joon Oh, Andrew C. Gallagher, Kevin P. Murphy, Florian Schroff, Jiyan Pan, and Joseph Roth. Modeling uncertainty with hedged instance embeddings. In *ICLR*, 2019.

[39] Yeonguk Oh, JoonKyu Park, Jaeha Kim, Gyeongsik Moon, and Kyoung Mu Lee. Recovering 3d hand mesh sequence from a single blurry image: A new dataset and temporal unfolding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 554–563, 2023.

[40] JoonKyu Park, Yeonguk Oh, Gyeongsik Moon, Hongsuk Choi, and Kyoung Mu Lee. Handoccnet: Occlusion-robust 3d hand mesh estimation network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1496–1505, 2022.

[41] Joonkyu Park, Gyeongsik Moon, Weipeng Xu, Evan Kaseman, Takaaki Shiratori, and Kyoung Mu Lee. 3d hand sequence recovery from real blurry images and event stream. In *European Conference on Computer Vision*, pages 343–359. Springer, 2024.

[42] Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. Reconstructing hands in 3d with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9826–9836, 2024.

[43] Matteo Poggi, Filippo Aleotti, Fabio Tosi, and Stefano Mattoccia. On the uncertainty of self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3227–3237, 2020.

[44] Yuzhe Qin, Yueh-Hua Wu, Shaowei Liu, Hanwen Jiang, Ruihan Yang, Yang Fu, and Xiaolong Wang. Dexmv: Imitation learning for dexterous manipulation from human videos. In *European Conference on Computer Vision*, pages 570–587. Springer, 2022.

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

[46] Yu Rong, Takaaki Shiratori, and Hanbyul Joo. Frankmocap: A monocular 3d whole-body pose estimation system via regression and integration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1749–1759, 2021.

[47] Marc H Schieber. Muscular production of individuated finger movements: the roles of extrinsic finger muscles. *Journal of Neuroscience*, 15(1):284–297, 1995.

[48] Tomas Simon, Hanbyul Joo, Iain Matthews, and Yaser Sheikh. Hand keypoint detection in single images using multiview bootstrapping. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1145–1153, 2017.

[49] Breannan Smith, Chenglei Wu, He Wen, Patrick Peluse, Yaser Sheikh, Jessica K Hodgins, and Takaaki Shiratori. Constraining dense hand surface tracking with elasticity. *ACM Transactions on Graphics (ToG)*, 39(6):1–14, 2020.

[50] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[51] Jiayi Wang, Diogo Luvizon, Franziska Mueller, Florian Bernard, Adam Kortylewski, Dan Casas, and Christian Theobalt. Handflow: Quantifying view-dependent 3d ambiguity in two-hand reconstruction with normalizing flow. *arXiv preprint arXiv:2210.01692*, 2022.

[52] Donglai Xiang, Hanbyul Joo, and Yaser Sheikh. Monocular total capture: Posing face, body, and hands in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10965–10974, 2019.

[53] Lixin Yang, Kailin Li, Xinyu Zhan, Jun Lv, Wenqiang Xu, Jiefeng Li, and Cewu Lu. 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*, pages 2750–2760, 2022.

[54] Zhengdi Yu, Shaoli Huang, Chen Fang, Toby P Breckon, and Jue Wang. Acr: Attention collaboration-based regressor for arbitrary two-hand reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12955–12964, 2023.

[55] Baowen Zhang, Yangang Wang, Xiaoming Deng, Yinda Zhang, Ping Tan, Cuixia Ma, and Hongan Wang. Interacting two-hand 3d pose and shape reconstruction from single color image. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11354–11363, 2021.

[56] Xiaoran Zhang, Daniel H Pak, Shawn S Ahn, Xiaoxiao Li, Chenyu You, Lawrence H Staib, Albert J Sinusas, Alex Wong, and James S Duncan. Heteroscedastic uncertainty estimation framework for unsupervised registration. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 651–661. Springer, 2024.

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

[58] Christian Zimmermann and Thomas Brox. Learning to estimate 3d hand pose from single rgb images. In *Proceedings of the IEEE international conference on computer vision*, pages 4903–4911, 2017.

[59] Christian Zimmermann, Duygu Ceylan, Jimei Yang, Bryan Russell, Max Argus, and Thomas Brox. Freihand: A dataset for markerless capture of hand pose and shape from single rgb images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 813–822, 2019.

[60] Binghui Zuo, Zimeng Zhao, Wenqian Sun, Wei Xie, Zhou Xue, and Yangang Wang. Reconstructing interacting hands with interaction prior from monocular images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9054–9064, 2023.