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

Supplementary Material

1. Implementation Details

Network in uncertainty modeling. Our uncertainty estimation employs a multi-scale Feature Pyramid Network (FPN) [9] built on a ResNet-34 backbone [6]. Multi-scale features from the four ResNet stages (with spatial resolutions of 64×64, 32×32, 16×16, and 8×8) are first projected into a unified 256-channel space via 1×1 convolutions followed by GroupNorm [11]. The FPN then fuses these features through three top-down upsampling stages. At each stage, the higher-level feature is bilinearly upsampled by $2\times$, concatenated with the corresponding lower-level feature, and refined by a 3×3 ConvBlock (256 channels, BN, ReLU). The resulting feature map $\mathbf{F}_{\text{fpn}} \in \mathbb{R}^{256\times32\times32}$ serves as the shared representation for both 2D pose prediction and uncertainty estimation.

The heatmap head $\mathcal{H}$ applies a 1×1 convolution on $\mathbf{F}_{\text{fpn}}$ to produce 21-channel heatmaps, followed by sigmoid and softmax normalization across spatial locations to form a probability distribution $p_j^{(h,w)}$. The mean 2D joint location $\boldsymbol{\mu}_j = (\mu_{j,x}, \mu_{j,y})$ is then computed via soft-argmax:

$$\boldsymbol{\mu}_j = \sum_{h,w} (h, w) \cdot p_j^{(h,w)}.$$

For uncertainty estimation, the variance head $\mathcal{V}$ aggregates spatial features using the heatmap as weights:

$$\mathbf{f}_j = \sum_{h,w} \mathbf{F}_{\text{fpn}}^{(h,w)} \cdot p_j^{(h,w)} \in \mathbb{R}^{256},$$

yielding a joint-specific feature vector $\mathbf{f}_j$. A linear layer then maps $\mathbf{f}_j$ to pre-activation uncertainty values $\mathbf{s}_j = (s_{j,x}, s_{j,y}) \in \mathbb{R}^2$, which are converted into positive variances through a Softplus [2] activation:

$$\boldsymbol{\sigma}_j^2 = \text{Softplus}(\mathbf{s}_j).$$

To ensure stable training, the bias of the variance head is initialized to 1.0, and the weights are initialized with Kaiming normal initialization [5] using slope $a = 0.01$.

Loss weights. In the final loss function, we set $\lambda_1 = 0.1$, $\lambda_2 = 1$, $\lambda_3 = 1$, and $\lambda_4 = 10$ to balance the individual loss terms.

2. More Evaluations

Calibration analysis of uncertainty. To validate the physical interpretability of our predicted uncertainty, we analyze the correlation between the model-predicted standard deviation σ and the actual reprojection error between the estimated 2D poses and ground-truth poses on the OakInk-MV [12] test set. As shown in Fig. 1, a Pearson correlation coefficient of 0.62 indicates a clear positive correlation, confirming that our uncertainty estimates are reasonably well calibrated. This calibration is important for the uncertainty-guided gating mechanism: regions with low predicted uncertainty tend to exhibit low actual errors and thus receive high gating weights, while high-uncertainty regions are appropriately down-weighted.

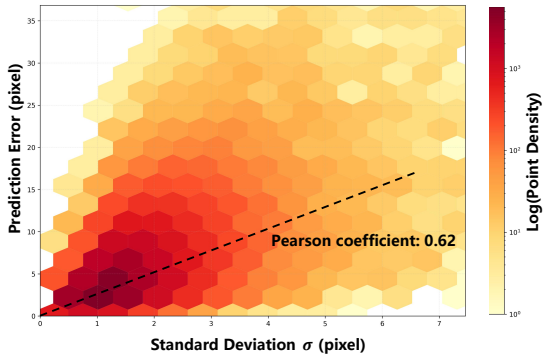

Figure 1. Calibration analysis of predicted uncertainty. The scatter plot illustrates the correlation between the predicted uncertainty (standard deviation σ) and the actual reprojection errors on the OakInk-MV test set, with each point representing a joint.

Number of Decoder Layers. We study the impact of the number of decoder layers in the refinement stage on 3D hand mesh reconstruction. As shown in Table 1, performance improves as the number of decoder layers increases. Considering the trade-off between accuracy and computational cost, we adopt a 6-layer decoder in the final model.

Decoder layers	MPJPE↓	PA-J↓	MPVPE↓	PA-V↓
2	15.38	8.99	15.53	9.44
4	14.20	8.47	14.49	8.95
6	13.10	8.30	13.39	8.78
8	12.99	8.16	13.27	8.66

Table 1. Performance of JUMP-Hand with different numbers of decoder layers on the HO3D-MV test set.

Comparison of Learnable DLT. To validate the effectiveness of our uncertainty-guided triangulation, we com-

pare several variants: vanilla DLT [4], Learnable DLT [7], and our proposed UG-DLT. Although Learnable DLT shares the same mathematical formulation as UG-DLT, its weights are learned directly through triangulation supervision, making them less interpretable than our uncertainty-derived weights. In contrast, UG-DLT derives its weights from our probabilistic joint-wise uncertainty estimation, making the weighting process physically interpretable and consistent with the overall uncertainty-guided design. As shown in Table 2, UG-DLT achieves the best performance among all variants, demonstrating that uncertainty-derived weighting provides a more reliable coarse geometric initialization than both vanilla DLT and purely data-driven Learnable DLT.

Method	MPJPE↓	PA-J↓
vanilla DLT	6.55	4.59
Learnable DLT	6.16	4.45
UG-DLT (Ours)	5.82	3.97

Table 2. Comparison of different triangulation methods on OakInk-MV test set.

3. More Visualizations

To further illustrate the effectiveness and robustness of JUMP-Hand, we provide additional qualitative visualizations across multiple datasets and under various challenging real-world conditions. As shown in Fig. 2, our method consistently outperforms existing approaches [13, 14] in scenarios such as motion blur, severe occlusions, and missing or incomplete camera views. Furthermore, Figs. 3–5 present qualitative results on all three benchmark datasets. For each sample, we display five representative views. In each view, we show the predicted 2D pose together with its associated uncertainty, along with the reprojected 3D hand mesh. One view is visualized at full resolution, while the remaining four are shown at half scale for compactness.

4. Additional Evaluations

Generalization to larger-scale datasets. We further evaluate JUMP-Hand on two larger-scale multi-view benchmarks, ARCTIC-MV [3] and InterHand-MV [10], to examine its generalization ability beyond the three main datasets. Following the same protocol as POEM, all methods are trained using 20% of the available training data. As shown in Table 3, JUMP-Hand consistently outperforms POEM on both datasets. On the full test sets, our method reduces MPVPE from 8.33 to 6.37 on ARCTIC-MV and from 11.35 to 8.34 on InterHand-MV. The gains are even larger on the 10% hard subsets, where the improvement reaches 6.87 mm on ARCTIC-MV and 7.86 mm on InterHand-MV. These results indicate that the proposed uncertainty-guided fusion generalizes well across larger-scale and more diverse data domains, while remaining particularly effective under challenging conditions.

Dataset	Method	Full set	10% hard subset
ARCTIC-MV	POEM	8.33	19.79
	Ours	**6.37**	**12.92**
InterHand-MV	POEM	11.35	31.56
	Ours	**8.34**	**23.70**

Table 3. MPVPE comparison on ARCTIC-MV and InterHand-MV. All models are trained using 20% of the training data.

Performance on non-challenging subsets. In addition to the challenging subsets reported in the main paper, we also evaluate the remaining non-challenging samples. As shown in Table 4, JUMP-Hand consistently performs better than POEM across all three datasets. On HO3D, our method improves MPVPE from 15.77 to 12.15 and AUC-V from 0.33 to 0.43. On DexYCB and OakInk, the gains are smaller but remain consistent across MPVPE, PA-V, and AUC-V. These results show that the uncertainty-guided gating improves robustness in difficult cases without sacrificing precision on easier samples.

Set	Method	MPVPE↓	PA-V↓	AUC-V↑
HO3D	POEM	15.77	9.20	0.33
	Ours	**12.15**	**8.20**	**0.43**
DexYCB	POEM	5.30	3.57	0.74
	Ours	**4.91**	**3.45**	**0.76**
OakInk	POEM	5.39	3.74	0.73
	Ours	**5.30**	**3.70**	**0.74**

Table 4. Comparison on non-challenging subsets.

Off-the-shelf confidence versus learned uncertainty. We compare the learned uncertainty in JUMP-Hand with the keypoint confidence produced by an off-the-shelf 2D detector, RTMPose-m [1, 8]. As shown in Table 5, replacing the learned uncertainty with RTMPose confidence consistently degrades performance on all three datasets. In particular, under the same 2D predictions, the learned uncertainty reduces MPVPE from 16.11 to 13.39 on HO3D, from 6.38 to 5.45 on DexYCB, and from 6.45 to 5.94 on OakInk.

We further analyze the entropy of the resulting joint-wise view weights. RTMPose confidence produces nearly uniform view weights, with entropy 1.59, close to the uniform upper bound 1.61, while the learned uncertainty yields substantially lower entropy of 1.38. This indicates that the

learned uncertainty provides more discriminative view selection. We attribute this difference to end-to-end optimization with 3D supervision: while off-the-shelf detector confidence is designed for 2D keypoint estimation quality, it is not necessarily calibrated for cross-view fusion in 3D reconstruction.

We additionally evaluate a combined-training setting by jointly training on multiple datasets. Under this setting, the performance can be further improved on HO3D and OakInk, reducing MPVPE to 8.62 and 5.52, respectively, while maintaining competitive performance on DexYCB. This suggests that the proposed uncertainty-guided fusion can also benefit from increased training diversity and scale.

Gating source	HO3D	DexYCB	OakInk
RTMPose confidence	16.11	6.38	6.45
Learned uncertainty (Ours)	**13.39**	**5.45**	**5.94**
Learned uncertainty (Ours)*	**8.62**	**5.51**	**5.52**

Table 5. MPVPE comparison of different gating sources. * denotes the model trained with combined datasets.

References

[1] MMPose Contributors. Openmmlab pose estimation toolbox and benchmark. https://github.com/open-mmlab/mmpose, 2020. 2

[2] 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. 1

[3] 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. 2

[4] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003. 2

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1026–1034, 2015. 1

[6] 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. 1

[7] 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

[8] Tao Jiang, Peng Lu, Li Zhang, Ningsheng Ma, Rui Han, Chengqi Lyu, Yining Li, and Kai Chen. Rtmpose: Real-time multi-person pose estimation based on mmpose, 2023. 2

[9] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2117–2125, 2017. 1

[10] 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 *Proceedings of the European Conference on Computer Vision*, pages 548–564. Springer, 2020. 2

[11] Yuxin Wu and Kaiming He. Group normalization. In *Proceedings of the European Conference on Computer Vision*, pages 3–19, 2018. 1

[12] 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. 1

[13] 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

[14] 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

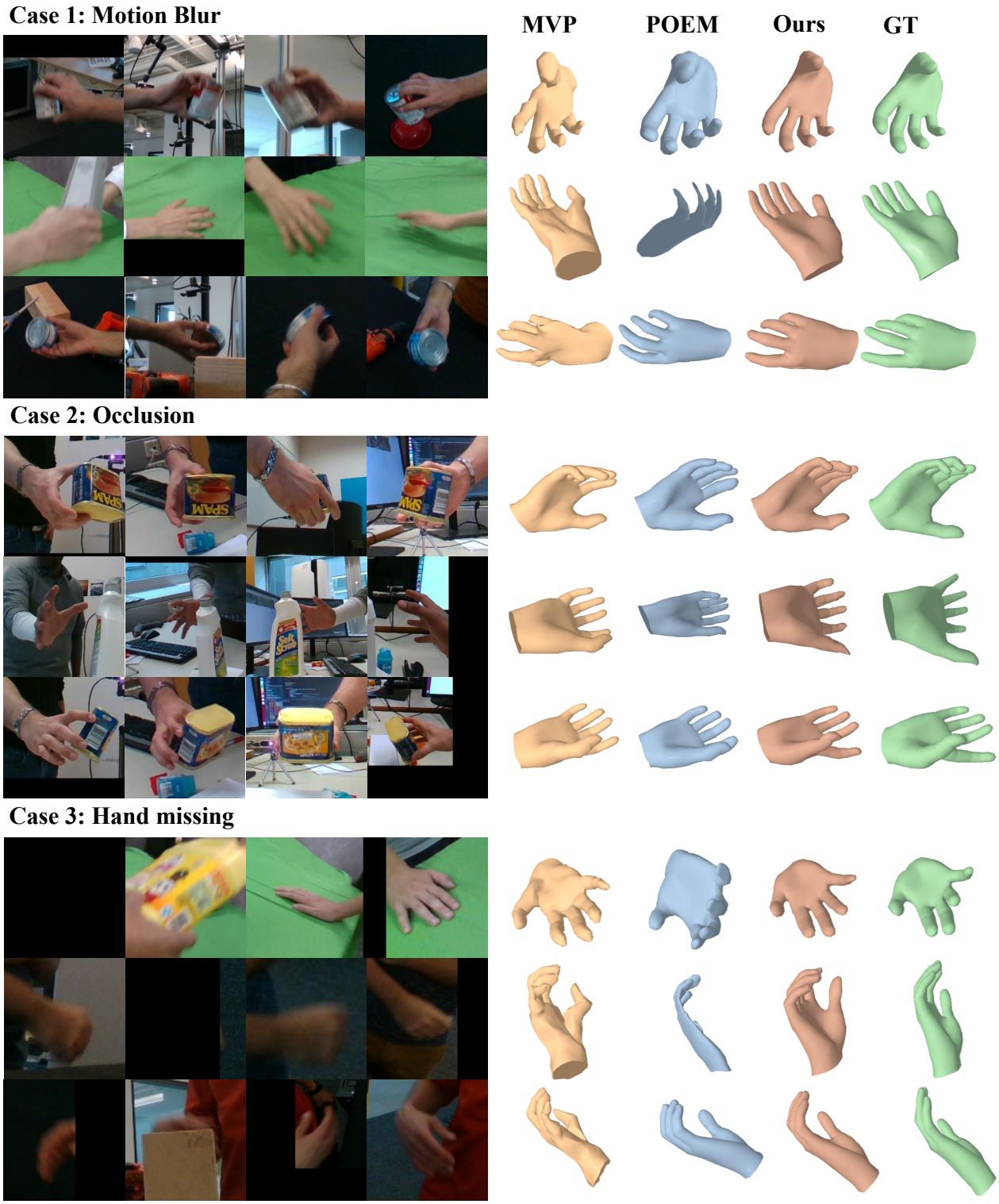

Figure 2. More qualitative results under challenging scenarios in testing set, including (1) motion blur, (2) occlusion, and (3) hand missing.

Figure 3. Qualitative results on the DexYCB-MV test set.

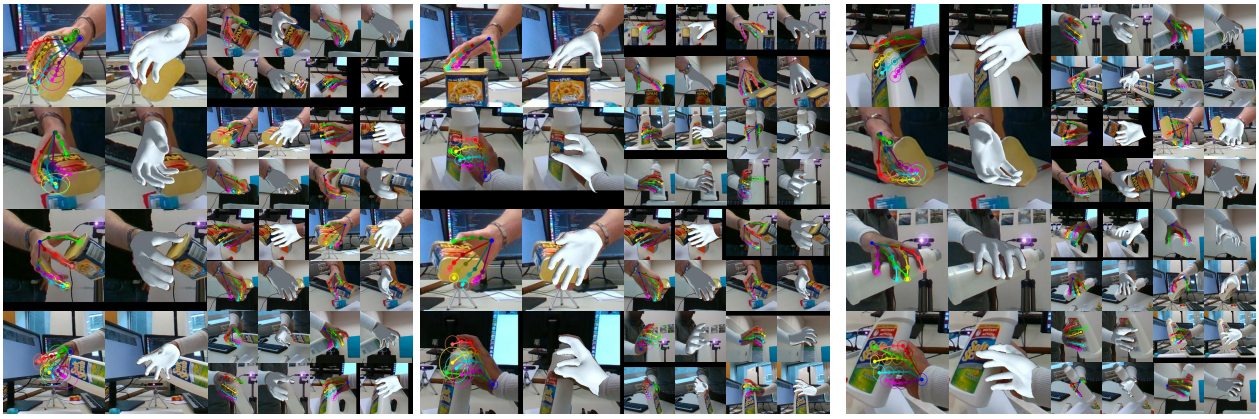

Figure 4. Qualitative results on the HO3D-MV test set.

Figure 5. Qualitative results on the OakInk-MV test set.