

UST-Hand: An Uncertainty-aware Spatiotemporal Point Cloud Interaction Network for 3D Self-supervised Hand Pose Estimation

Supplementary Material

A. Video Demo

We provide sequential visualizations in the attached video to illustrate our method’s performance.

B. Additional Ablation Study

To further validate our approach, we conduct fine-grained ablation studies on the individual components within the Confidence-aware feature interaction module and the Spatiotemporal Point Transformer (STPT). As shown in Tab. 1, in the Confidence-aware feature interaction module, removing the adaptive-GCN or CASA mechanism leads to performance degradation, demonstrating that both structural topology modeling and confidence-aware feature aggregation contribute to robust cross-view feature interaction. For the STPT module, temporal attention shows the largest impact, highlighting the importance of exploiting temporal consistency for video-based hand pose estimation. These results confirm that each component plays an indispensable role in achieving accurate pose refinement.

Table 1. Additional Ablation Study of UST-Hand on DexYCB-MV dataset. We report the MPVPE (mm), PA-V (mm), MPJPE (mm), PA-J (mm).

Method	MPVPE	PA-V	MPJPE	PA-J
UST-Hand	8.16	4.81	7.84	5.31
w/o GCN	8.22	4.83	7.90	5.41
w/o CASA	8.28	4.87	7.95	5.43
w/o SpatialAttn	8.46	5.02	8.18	5.58
w/o TemporalAttn	8.55	5.09	8.20	5.64
w/o CrossAttn	8.50	4.99	8.15	5.59

C. Model Analysis

Different Temporal Length. We examine the performance of UST-Hand with varying temporal lengths in the video sequence. The results on HanCo dataset are shown in Tab. 2 rows t1-t7. We find that using 5 frames achieves the best performance. Increasing temporal length from 1 to 5 frames enables the model to capture hand motion patterns and reduce frame-wise jittering. However, extending to 7 frames shows marginal degradation, suggesting that excessively long temporal windows may introduce optimization challenges. We use a temporal length of 5 to balance modeling capacity and computational efficiency.

Table 2. Performance of UST-Hand on varying temporal length, the number of STPT blocks and cameras in the multi-view setting. The best results are highlighted in **bold**.

	Exp	MPVPE	PA-V	MPJPE	PA-J
HanCo	t1	5.94	4.15	5.38	3.55
	t3	5.91	4.10	5.27	3.53
	t5	5.82	4.13	5.19	3.50
	t7	5.87	4.07	5.25	3.51
	b1	6.01	4.19	5.45	3.62
	b2	5.95	4.17	5.40	3.55
	b3	5.88	4.13	5.34	3.51
	b4	5.82	4.13	5.19	3.50
DexYCB	c2	10.90	7.06	10.48	7.12
	c4	9.63	6.02	9.33	6.19
	c6	8.41	5.16	8.17	5.61
	c8	8.16	4.81	7.84	5.31

Different Number of STPT Blocks. We evaluate the performance of UST-Hand on varying the number of STPT blocks. The results on HanCo dataset are shown in Tab. 2 rows b1-b4. The results show that stacking more STPT blocks consistently improves performance. This validates that iterative refinement through multiple blocks enables the model to progressively correct pose errors by repeatedly integrating spatiotemporal information and distributional uncertainty. We use 4 STPT blocks in our final model to achieve optimal performance.

Different Number of Cameras. We evaluate UST-Hand across varying numbers of cameras in the multi-view setting. The results on the DexYCB dataset are shown in Tab. 2 rows c2-c8, where "c2" indicates the use of two cameras. The results demonstrate that UST-Hand effectively fuses complementary features from diverse viewpoints, boosting overall performance as the number of cameras increases.

D. More Qualitative Results

We provide comprehensive qualitative results across all three evaluation datasets to further validate our method’s superiority. Specifically, Figs. 1 to 3 compare our method with the SOTA approach HaMuCo, where both models utilize 2D keypoints generated by Wilor for self-supervision. Furthermore, Figs. 4 to 6 explicitly highlight the robustness of UST-Hand in maintaining strict multi-view geometric consistency, even when encountering highly diverse hand poses and complex camera view configurations.

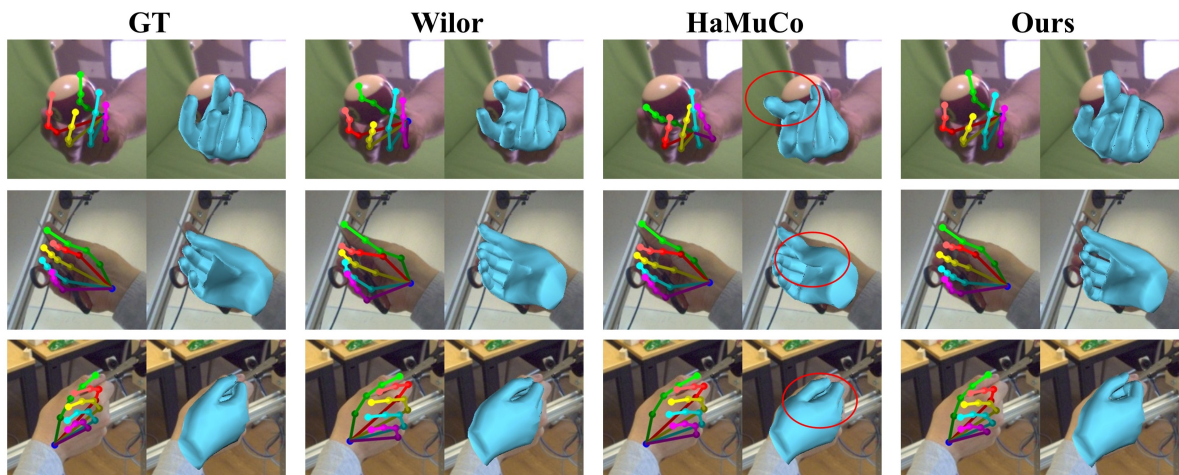


Figure 1. 2D joints prediction and 3D mesh prediction between ground-truth, Wilor, HaMuCo, and ours on HanCo dataset.

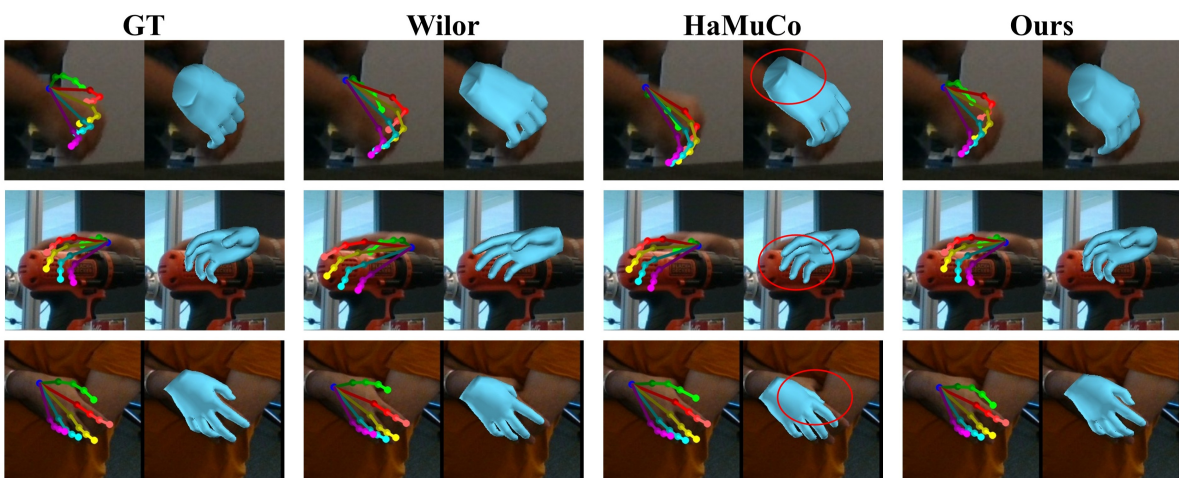


Figure 2. 2D joints prediction and 3D mesh prediction between ground-truth, Wilor, HaMuCo, and ours on DexYCB-MV dataset.

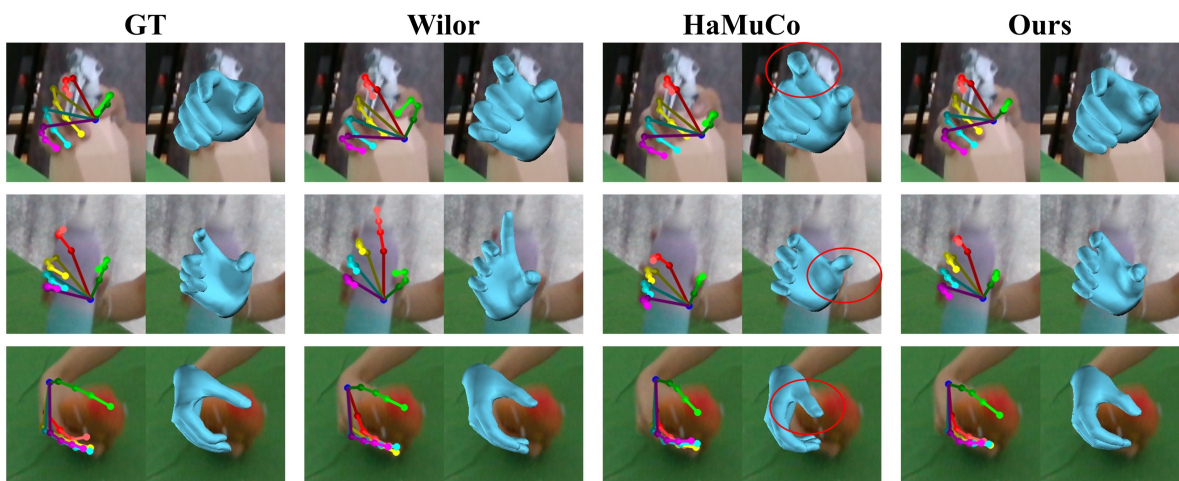


Figure 3. 2D joints prediction and 3D mesh prediction between ground-truth, Wilor, HaMuCo, and ours on OakInk-MV dataset.

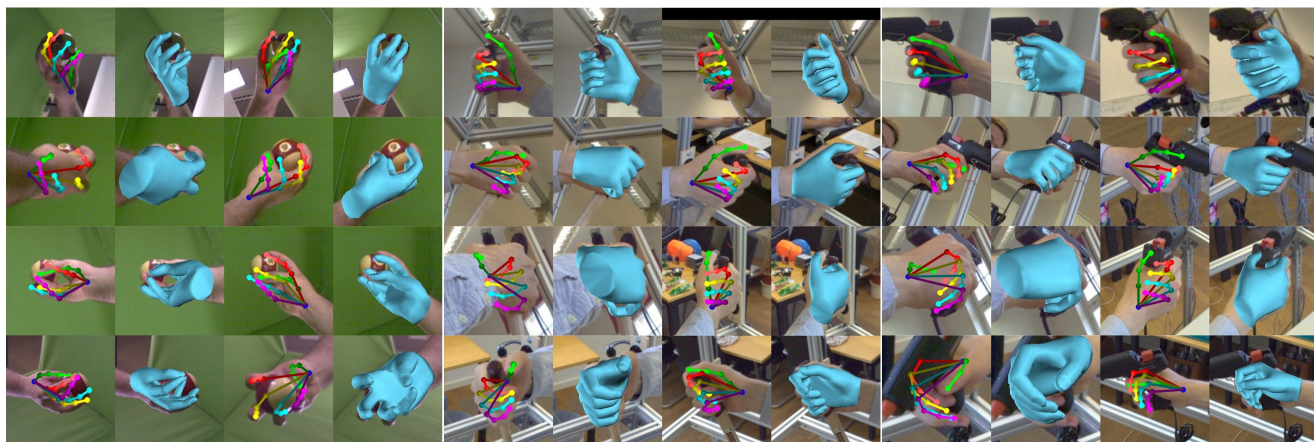


Figure 4. Qualitative results of each view on HanCo dataset.



Figure 5. Qualitative results of each view on DexYCB-MV dataset.

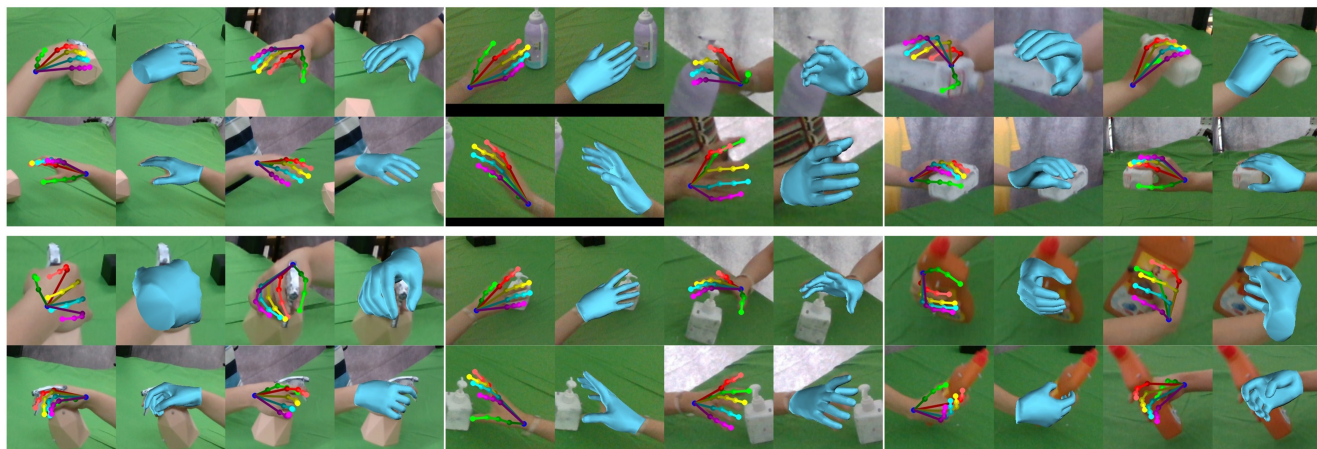


Figure 6. Qualitative results of each view on OakInk-MV dataset.