Supplementary Material for "Enhancing Hands in 3D Whole-Body Pose Estimation with Conditional Hands Modulator"

Gyeongsik Moon
Korea University
mks0601@korea.ac.kr
https://mks0601.github.io/Hand4Whole-plus-plus

In this supplementary material, we provide more experiments, discussions, and other details that could not be included in the main text due to the lack of pages. The contents are summarized below:

- Sec. S1: Ablation studies
- Sec. S2: Implementation details
- Sec. S3: Running time analysis

S1. Ablation studies

Expressiveness of MANO and SMPL-X shape space. In Sec. 3.3 of the main manuscript, we describe our finger articulation and shape transfer strategy. The shape transfer aims to leverage the more expressive hand shape space of MANO. Fig. S1 and Tab. S1 support this design choice by demonstrating that MANO provides a more expressive hand shape representation. The figure shows that MANO hands better align with the 3D scans, while the table reports lower point-to-point errors compared to SMPL-X hands.

For this comparison, we fit both MANO and SMPL-X models to the 3D scans and keypoints from the MANO test set [3]. Both models are initialized by rigidly aligning the wrist and the four MCP joints (index, middle, ring, pinky), with zero shape parameters and zero finger poses. During SMPL-X optimization, we use only the hand vertices from the full-body mesh. The objective consists of: 1) 3D keypoint loss, 2) point-to-point loss, and 3) shape parameter regularization, weighted by 1, 1, and 0.001, respectively. Each sample is optimized for 500 iterations.

Body root pose regularizer. In Sec. 3.4 of the main manuscript, we describe the body root pose regularizer, which encourages the root pose to remain vertical when only hand annotations are available and full-body annotations are not available. As shown in Fig. S2, without this regularizer, the recovered 3D human often exhibits an incorrect root pose due to the lack of full-body annotations in hand-only datasets. Since hand-only datasets are typically captured with subjects in an upright standing pose, we introduce this prior to regularize the root orientation. Specifically, we enforce alignment between the up-right direction of the world coordinate system and that of the human body by constraining their dot product to be 1.

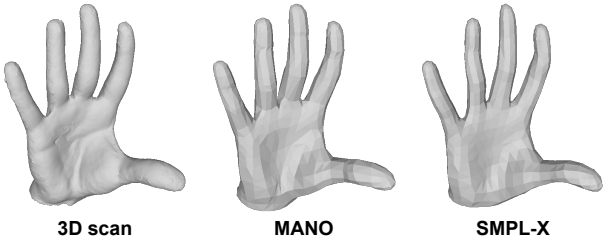

Figure S1. Comparison of hand shape expressiveness between MANO and SMPL-X. MANO produces hand shapes that more closely match the 3D scans compared to SMPL-X.

Table S1. Comparison of point-to-point distances between 3D scans and optimized hand meshes. We report the mean point-to-point distances between 3D scans and hand meshes optimized from 1) SMPL-X and 2) MANO.

Settings	Point-to-point error (mm)
SMPL-X hands	1.98
MANO hands (Ours)	1.34

Cross-attention in CHAM. As described in Sec. 3.2 of the main manuscript, we introduce cross-attention with 2D positional encoding to capture the inter-relationship between the two hands. Tab. S2 demonstrates that our cross-attention effectively reduces both MPVPE and MRRPE in hand-only datasets with challenging two-hand interactions.

S2. Implementation details

We implement our method using PyTorch. The model is trained for 4 epochs with a batch size of 32, using the Adam

Table S2. Comparison of MPVPE/MRRPE with and without cross-attention in CHAM.

Settings	IH26M	ReIH	HIC
Without cross attention	9.77/35.36	9.12/19.42	18.25/30.44
With cross attention (Ours)	**9.40/32.30**	**7.98/16.37**	**17.72/29.09**

Table S3. Running time (in seconds) to process a single image on an RTX A6000 GPU.

Hand detector	Hand pose estimator [2]	CHAM	Whole-body pose estimator [1]	Total
0.01	0.05	0.01	0.03	0.1

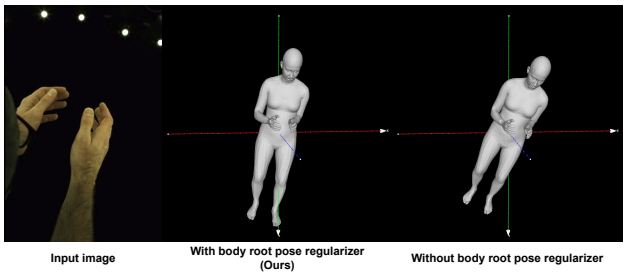

Figure S2. Effectiveness of the body root pose regularizer. The red, green, and blue axes represent the x-, y-, and z-axes of the world coordinate system, respectively. The green axis (y-axis) points downward, which is opposite to the up-right direction in the world coordinate system.

optimizer and an initial learning rate of 1e-4. The learning rate is reduced by a factor of 10 at the 3rd epoch. All loss terms are equally weighted with a weight of 1. Training is performed on a single NVIDIA RTX A6000 GPU and takes approximately 20 hours. All other implementation details are provided in the released code.

S3. Running time analysis

Tab. S3 summarizes the runtime of each component in our Hand4Whole++. Overall, Hand4Whole++ runs at 10 frames per second on a single RTX A6000 GPU. Among all components, the hand pose estimator (WiLoR[2]) is the most time-consuming. In contrast, our CHAM module runs significantly faster than both the hand pose estimator (WiLoR [2]) and the whole-body pose estimator (SMPLer-X [1]), thanks to its lightweight design.

References

[1] Zhongang Cai, Wanqi Yin, Ailing Zeng, Chen Wei, Qingping Sun, Wang Yanjun, Hui En Pang, Haiyi Mei, Mingyuan Zhang, Lei Zhang, et al. SMPLer-X: Scaling up expressive human pose and shape estimation. In *NeurIPS*, 2023. 2

[2] Rolandos Alexandros Potamias, Jinglei Zhang, Jiankang Deng, and Stefanos Zafeiriou. WiLoR: End-to-end 3D hand localization and reconstruction in-the-wild. In *CVPR*, 2025. 2

[3] Javier Romero, Dimitrios Tzionas, and Michael J Black. Embodied Hands: Modeling and capturing hands and bodies together. *ACM TOG*, 2017. 1