Camera-Space Hand Mesh Recovery via Semantic Aggregation and Adaptive 2D-1D Registration – Supplementary Materials

Xingyu Chen[1*] Yufeng Liu[1,3] Chongyang Ma[1] Jianlong Chang[2] Huayan Wang[1]
Tian Chen[1] Xiaoyan Guo[1] Pengfei Wan[1] Wen Zheng[1]
[1]Y-tech, Kuaishou Technology [2]Huawei Cloud & AI
[3]SEU-ALLEN Joint Center, Institute for Brain and Intelligence, Southeast University, China.

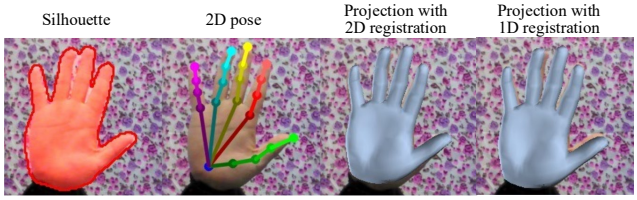

Figure 1. Visualization of 2D/1D registration results.

ISM	Multi-scale	Self-regression	PA-MPJPE ↓	PA-MPJPE ↓
			8.82	9.06
✓			8.73	8.89
✓	✓		8.47	8.56
✓	✓	✓	8.46	8.54

Table 1. Effects of our designs for the spiral decoder.

Visualization of 2D/1D registration. As shown in Figure 1, the 2D registration is instructive for finger alignment while the 1D process is beneficial to holistic shape alignment. Thus, joint landmarks and silhouette have different effects on the task of root recovery, and both of them can be effectively leveraged by our adaptive 2D-1D registration scheme.

Effects of our designs for spiral decoder. We improve the spiral decoder with ISM, multi-scale mechanism, and self-regression. As shown in Table 1, all of these three design choices are beneficial to 2D-to-3D decoding, where ISM and multi-scale mechanism has relatively more significant impact.

Full result comparison on FreiHAND dataset. As shown in Table 2 and 3. For root-relative and camera-space tasks, CMR achieves state-of-the-art performance on all the merits. Figure 2 plots PCK curves of 3D joints, which can serve as a complement of Figure 9 of our main text.

Full-feature representation after 2D cues. Figure 4 serves as a complement of Figure 8 in the main text. $\mathbf{H}_s$ and $sum(\mathbf{H}_p)$ induce holistic shape and pose. In contrast, $\mathbf{H}_p$

*Corresponding author, chenxingyu@kuaishou.com

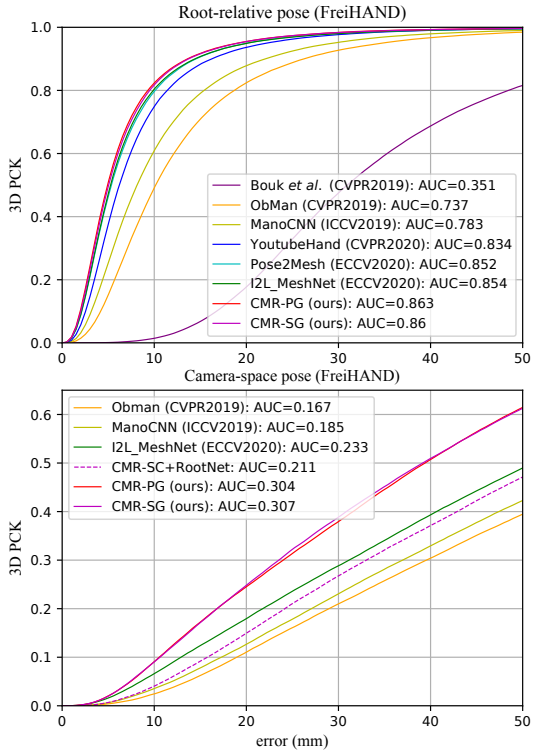

Figure 2. Pose PCK vs. error thresholds. Our CMR outperforms existing methods by a large margin.

invites simultaneously activated joint representation that essentially implies semantic relation. However, this relation representation is not comprehensive. We design $group(\mathbf{H}_p)$ for explicitly exploring known high-level semantics so that more comprehensive joint relations can be captured.

More qualitative results. Figure 5 and 6 illustrate comprehensive qualitative results of our predicted silhouette, 2D pose, projection of mesh, side-view mesh, camera-space mesh and pose in meter. Different from most methods, 3D roots required by image-mesh alignment are provided by CMR itself rather than ground truth.

Referring to Figure 5, FreiHAND's challenges include

Method	Backbone	PA-MPJPE ↓	J-AUC ↑	PA-MPVPE ↓	V-AUC ↑	F@5 mm ↑	F@15 mm ↑
Boukhayma *et al.* (CVPR2019) [1]	ResNet50	35.0	0.351	13.2	0.738	0.427	0.895
ObMan (CVPR2019) [4]	ResNet18	13.3	0.737	13.3	0.736	0.429	0.907
MANO CNN (ICCV2019) [7]	ResNet50	11.0	0.783	10.9	0.783	0.516	0.934
YotubeHand (CVPR2020) [5]	ResNet50	8.4	0.834	8.6	0.830	0.614	0.966
Pose2Mesh (ECCV2020) [2]	–	7.7	0.852	7.8	0.850	0.674	0.969
I2L-MeshNet (ECCV2020) [6]	ResNet50	7.4	0.854	7.6	0.850	0.681	0.973
CMR-SG	ResNet18	7.5	0.851	7.6	0.850	0.685	0.971
CMR-PG	ResNet18	7.4	0.853	7.5	0.851	0.687	0.973
CMR-SG	ResNet50	7.0	0.860	7.1	0.858	0.706	0.976
CMR-PG	ResNet50	**6.9**	**0.863**	**7.0**	**0.861**	**0.715**	**0.977**

Table 2. Results of root-relative mesh recovery on FreiHAND. "J-AUC, V-AUC" denote AUC of 3D joint and mesh vertex, respectively. "F@5 mm, F@15 mm" is the harmonic mean between recall and precision between two meshes *w.r.t.* a specific distance threshold.

Method	Backbone	CS-MPJPE ↓	J-AUC ↑	CS-MPVPE ↓	V-AUC ↑	F@5 mm ↑	F@15 mm ↑
ObMan (CVPR2019) [4]	ResNet18	85.2	0.168	85.4	0.167	0.087	0.305
ManoCNN (ICCV2019) [7]	ResNet50	71.3	0.185	71.5	0.184	0.102	0.345
I2L-MeshNet (ECCV2020) [6]	ResNet50	60.3	0.233	60.4	0.232	0.132	0.394
CMR-PG	ResNet18	50.6	0.290	50.6	0.289	0.155	0.474
CMR-SG	ResNet18	49.7	0.295	49.8	0.294	0.159	0.481
CMR-PG	ResNet50	48.9	0.304	49.0	0.303	0.163	0.488
CMR-SG	ResNet50	**48.8**	**0.307**	**48.9**	**0.306**	**0.166**	**0.492**

Table 3. Results of camera-space mesh recovery on FreiHAND. Please refer to Table 2 for metrics explanation.

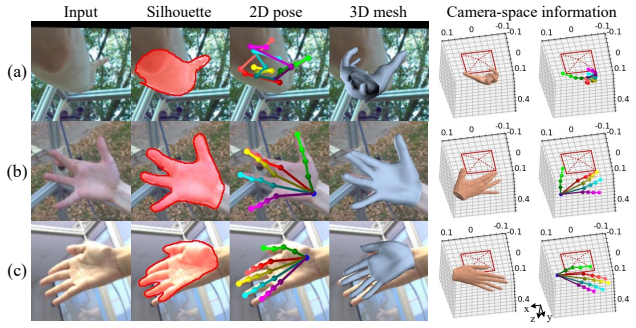

Figure 3. Failure cases of CMR-SG.

hard poses, object interactions, and truncation. Overcoming these difficulties, CMR can generate accurate silhouette, 2D pose, and camera-space 3D information.

Referring to Figure 6, samples of RHD, STB, and real-world dataset released by [3] are illustrated. We directly use the FreiHAND model for these datasets, and equally accurate predictions are obtained. Thus, CMR demonstrates superior capability of cross-domain generalization. Figure 6 also shows examples on Human3.6M and COCO. It can be seen that our CMR achieves reasonable results in the task of human body recovery.

Failure case analysis. Figure 3 shows three typical failure cases of CMR-SG. When only a small portion of the hand is visible in the input (Figure 3(a)), CMR-SG predicts wrong silhouette and 2D pose. Consequently, the camera-space information is not accurate. For cases of occlusion (*e.g.*, Figure 3(b), in which the forefinger is completely occluded by the middle finger), although 2D pose and silhouette prediction results are still reasonable, it is difficult to obtain accurate 3D mesh since self-occlusion is challenging for the mesh recovery stage. Referring to Figure 3(c), strong contrast and extreme illumination change in the RGB input leads to large but consistent errors in silhouette, 2D pose, and 3D mesh prediction results.

References

[1] Adnane Boukhayma, Rodrigo de Bem, and Philip HS Torr. 3D hand shape and pose from images in the wild. In *CVPR*, 2019. 2

[2] 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 *ECCV*, 2020. 2

[3] 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 *CVPR*, 2019. 2

[4] 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 *CVPR*, 2019. 2

[5] Dominik Kulon, Riza Alp Guler, Iasonas Kokkinos, Michael M Bronstein, and Stefanos Zafeiriou. Weakly-supervised mesh-convolutional hand reconstruction in the wild. In *CVPR*, 2020. 2

[6] 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 *ECCV*, 2020. 2

[7] 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 *ICCV*, 2019. 2

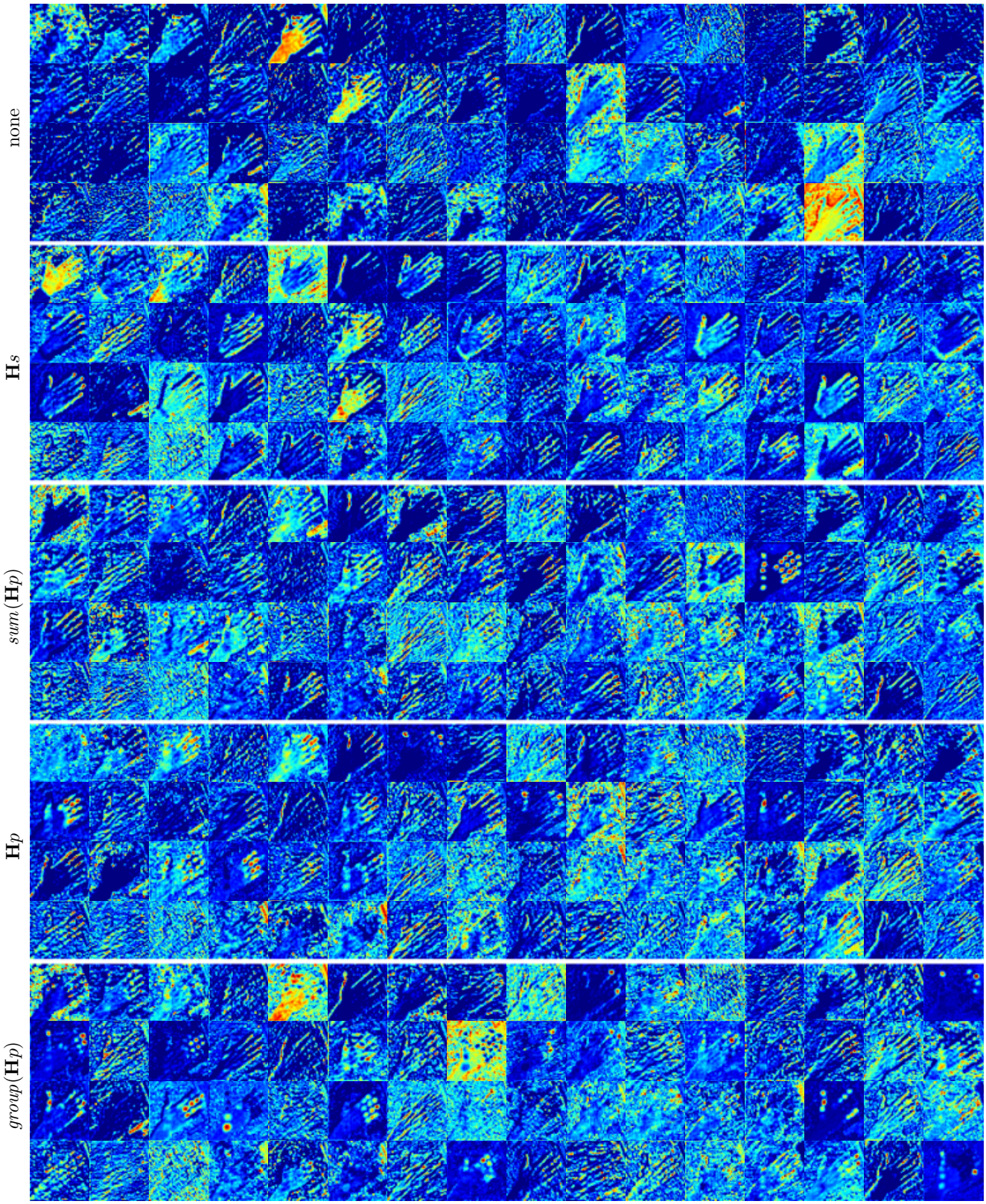

Figure 4. Comparison of full feature representation after various 2D cues. *group*($\mathbf{H}_p$) induces more semantic relation of joints. Input image is the first line of Figure 1 in the main paper.

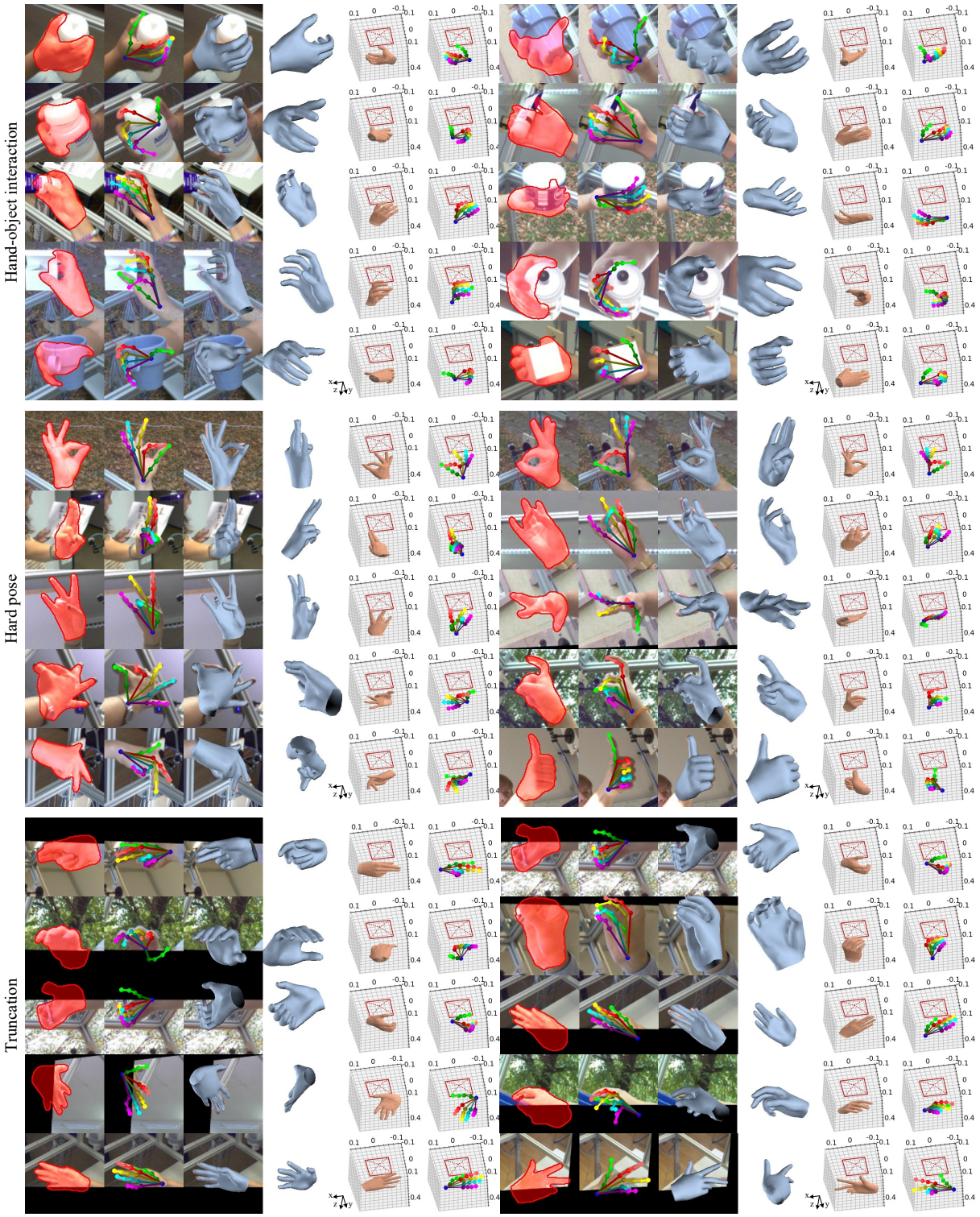

Figure 5. Qualitative results on FreiHAND dataset. We can handle challenging cases of object occlusion, hard poses, and truncation.

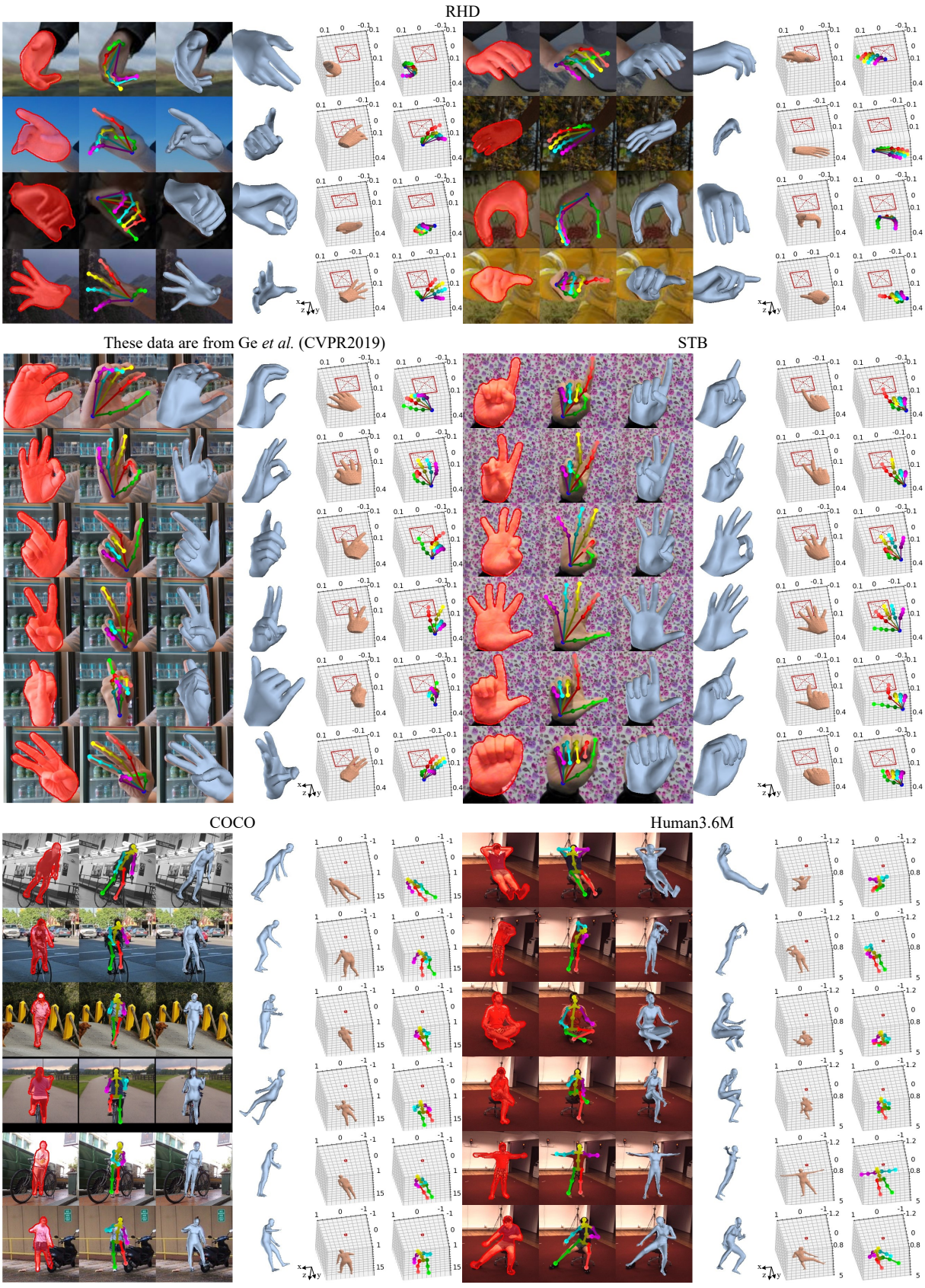

Figure 6. Qualitative results on other datasets. The hand model is trained with FreiHAND only.