MS-MANO: Enabling Hand Pose Tracking with Biomechanical Constraints

Supplementary Material

1. Overview

In the supplementary materials, we will illustrate the full upper extremity of the MS-MANO in Sec. 2, which is based on SMPL-X [4], report the quantitative results on OakInk dataset [5] and more qualitative results in Sec. 3. Finally, the sensitivity of muscle insertion points discussed in the ablative study is more detailed and elaborated in the attached video, along with more qualitative results.

2. Upper Extremity of Musculoskeletal SMPL-X

As mentioned in the main paper, the hand musculoskeletal system is actually a part of the whole upper extremity, and thus, it is a forearm-wrist-hand structure. When we fit the musculoskeletal system into the MANO, which does not have a forearm, we integrate the MANO model with the human body model SMPLX [4]. We illustrate the full upper extremity in Fig. 1. We append the configuration file of the full upper extremity in the accompanying "upper_extremity.xml" file.

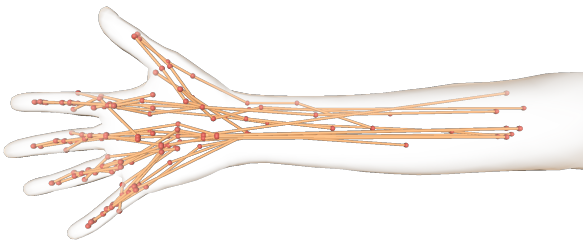


Figure 1. Full Upper Extremity of MS-MANO.

Note that when the MS-MANO is actuated in the simulation, the muscles in the forearm are also used. However, when we report the quantitative results and give the qualitative results, only the MANO part is adopted for a fair comparison.

3. Results, Extended

In the main paper, we have already reported the quantitative results on DexYCB [1] with two baseline methods, gSDF [2], and Deformer [3], and with the biomechanical regularizer, BioPR. Both the baseline methods can enjoy a performance boost with minimal computational overhead.

Methods	MPJPE↓	AUC↑	AE↓
gSDF [2]	8.22	92.3	27.9
gSDF + ours	**6.51**	**94.2**	**27.4**
Deformer [3]	11.15	91.0	29.9
Deformer + ours	**10.36**	**92.0**	**28.8**

Table 1. Quantitative Results with other baseline methods on OakInk.

To further prove the performance gain is ubiquitous, we report the quantitative results on the OakInk dataset [5] in Table 1. Previous baselines faced major problems primarily caused by time-related inconsistencies, leading to extensive jitters in many sequences. In scenarios demanding precise finger pose estimation, such as for the index and middle fingers, our method aligns more accurately with the input gestures. This accuracy is especially evident when transferring objects like a box; our predictions maintain fidelity to the ground truth across nearly every finger, while baseline methods tend to generate looser representations.

The advantage extends to complex situations involving severe occlusions where our approach demonstrates a clear edge. For example, the reconstruction of middle and ring finger poses is more accurate, and thumb tracking is markedly improved even when the thumb is heavily occluded.

When analyzing the act of gripping a mustard bottle with force, our method's predictions conform more closely to the actual finger tightness, particularly in the middle, ring, and little fingers. This fidelity is also evident in the thumb's posture, where our approach avoids the odd distortions seen in the gSDF method, offering a representation that is true to the observed posture.

4. Video Demo

Please kindly refer to the supplementary video which is submitted along with the supplementary file.

References

[1] 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. 1

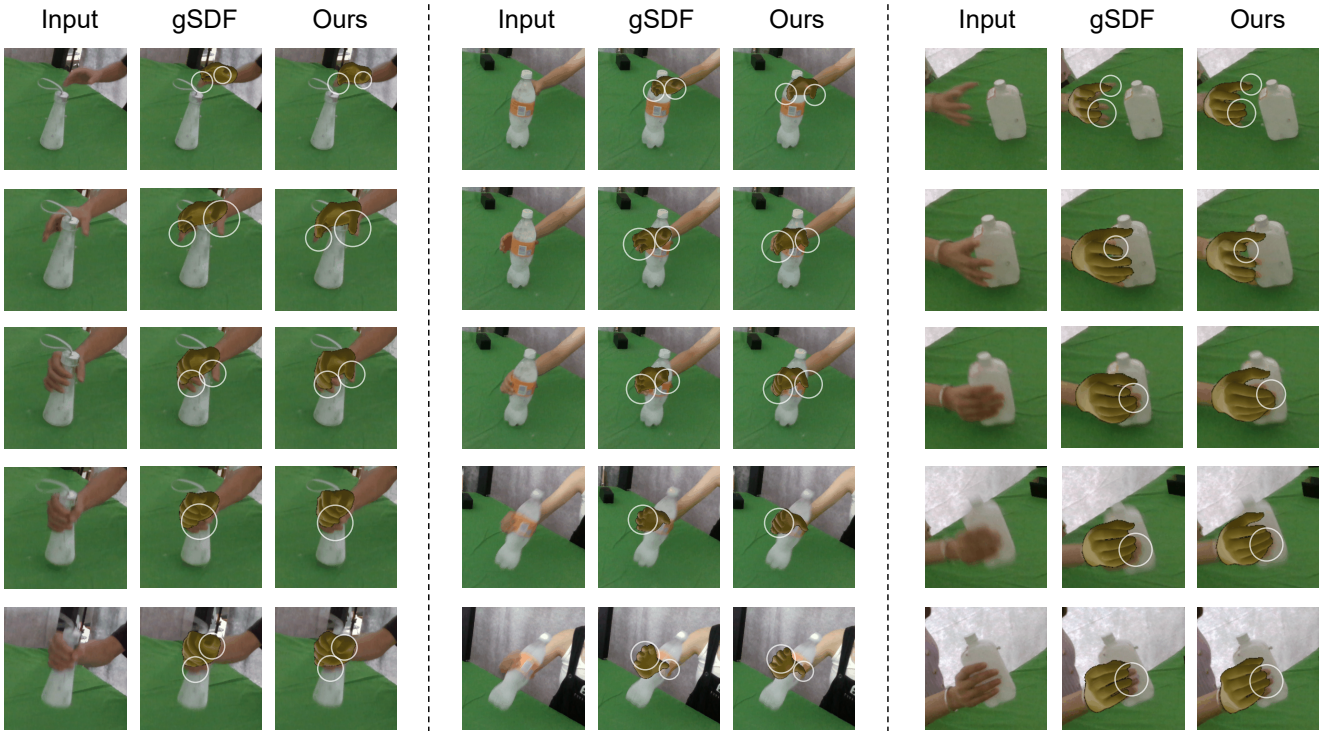
[2] Zerui Chen, Shizhe Chen, Cordelia Schmid, and Ivan Laptev. gsdf: Geometry-driven signed distance functions for 3d hand-

object reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12890–12900, 2023. 1

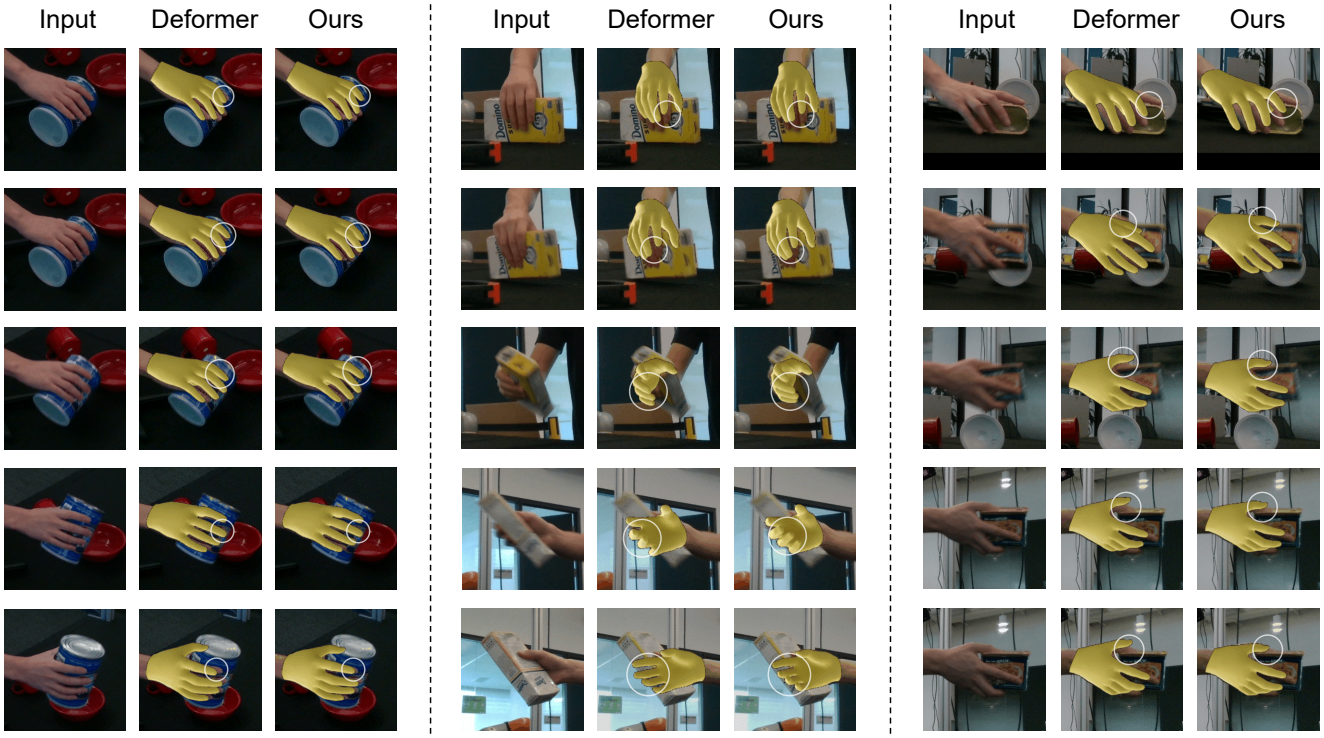
[3] Qichen Fu, Xingyu Liu, Ran Xu, Juan Carlos Niebles, and Kris M Kitani. Deformer: Dynamic fusion transformer for robust hand pose estimation. *arXiv preprint arXiv:2303.04991*, 2023. 1

[4] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *CVPR*, 2019. 1

[5] 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



(a) Qualitative results on OakInk. **Left**: Compared to Deformer, our method has results that are more in line with the input gesture for index and middle finger poses. **Middle**: When one picks up a box and hands it out, the Deformer's prediction appears to be looser, and we are closer to the ground truth on nearly every finger. **Right**: The poses of the middle and ring fingers are better reconstructed with our method, and we have better results on thumbs when encountering severe occlusion.



(b) Qualitative results on DexYCB. **Left**: When a person is forcefully grasping a mustard bottle, there is a difference in the tightness of the middle, ring, and little fingers, comparing gSDF to our method. The projected results of our method better align with the input image. **Middle**: The thumb posture predicted by the gSDF method exhibits some odd distortion, which is not observed in our approach. **Right**: The results of the thumb posture predicted by Deformer were far from the real situation, while our prediction is better aligned with the ground truth.

Figure 2. Extended qualitative results