Overcoming the Trade-Off Between Accuracy and Plausibility in 3D Hand Shape Reconstruction

Ziwei Yu[1] Chen Li[1] Linlin Yang[1] Xiaoxu Zheng[2] Michael Bi Mi[2] Gim Hee Lee[1] Angela Yao[1]
[1]National University of Singapore [2]Huawei International Pte Ltd, Singapore
{yuziwei, lichen}@u.nus.edu, {mu4yang, zhengxiaoxu66}@gmail.com, michaelbimi@yahoo.com
{gimhee.lee, ayao}@comp.nus.edu.sg

In this supplementary material, we first illustrate more details about MANO parameter estimation architectures in Sec. 1. Then joint prediction from mesh is discussed, including the use of a MANO linear regression matrix, a neural network, and VAE, in Sec. 2. Finally, more qualitative results are shown in Sec. 3.

1. Estimating the MANO Parameters

We design four types of architecture to estimate the MANO parameters, which are used in previous works *i.e.*, MANO CNN [1, 9] and MANO Fit [8]. Please see Table 1 and Fig. 2 for more details. Furthermore, we design two baseline pipelines for our model, *i.e.*, Ours(MANO Fit Joint) and Ours(MANO Fit Mesh), which directly use the non-parametric model joints or mesh to estimate the MANO parameters. This is a critical and naive baseline for our model. The difference between MANO Fit and MANO CNN is that the MANO Fit pipeline uses the 3D estimated joints or vertices to regress the MANO parameters, whereas MANO CNN directly uses image features. The experimental results offer some insights: 1) The MANO CNN pipeline shows better results than the MANO Fit pipeline, which is consistent with the conclusion of [9]. 2) There is a slight difference between the MANO CNN joints pipeline and the MANO CNN pipeline, which verifies the effectiveness of 3D joint information in the MANO parameter estimation. 3) The MANO Fit Mesh results are better than the MANO Fit Joint results. One reason is that mesh vertices contain hand shape information, which is beneficial to shape parameter estimation. 4) Our MANO Fit results are better than those of the previous MANO CNN and MANO Fit, which verifies the effectiveness of the non-parametric model. 5) Our results are better than ours (MANO Fit), which verifies that our twist-swing module is meaningful and essential when combining the non-parametric model and the MANO model. In other words, although combining the non-parametric model and MANO model is intuitive, it is still a challenging task. The results also show our contribution to proposing an effective combined model.

MANO	FreiHAND			
Method	MPJPE	MPVPE	MPJPE	MPJPE
MANO CNN Joints [9]	8.84	9.10	0.55	0.17
MANO CNN [9]	8.69	8.83	0.54	0.16
MANO Fit Joint [8]	9.95	10.08	0.60	0.19
MANO Fit Mesh [8]	8.81	8.95	0.53	0.16
Ours(MANO Fit joint)	8.76	8.71	0.54	0.16
Ours(MANO Fit Mesh)	8.27	8.30	0.53	0.16
Ours	**7.42**	**7.43**	**0.51**	**0.15**

Table 1. Comparisons with the MANO parameter estimation pipeline on the FreiHAND test set. Best scores are highlighted in **Bold**.

2. Joint From Mesh

As mentioned in the main paper, there is a gap of around 2 mm when directly using the joint from mesh. To further understand the joint from mesh approach, we design the two baselines following previous work [5, 6] in Table 2. Here, the baseline (w. $\mathcal{J}$) means that we derive the joints from the mesh using the MANO linear regress matrix, which is a common strategy for non-parametric model-based methods [2, 3, 5, 7]. The other baseline (w. NN) is a model that uses a neural network to regress the joints from hand meshes. Experimental results show that our proposed VAE pipeline is better than the above methods, especially when directly using baseline1. This verifies the effectiveness of our proposed VAE module. In addition, the results of baseline1 are better than baseline2, which shows that using the MANO linear regress matrix is better than neural network estimation. These results reveal that regressing the joints from a hand mesh is challenging and that a predefined regression matrix is more useful than direct neural network fitting. In addition, we notice that even given the ground truth mesh, there is a difference of about 2 mm between

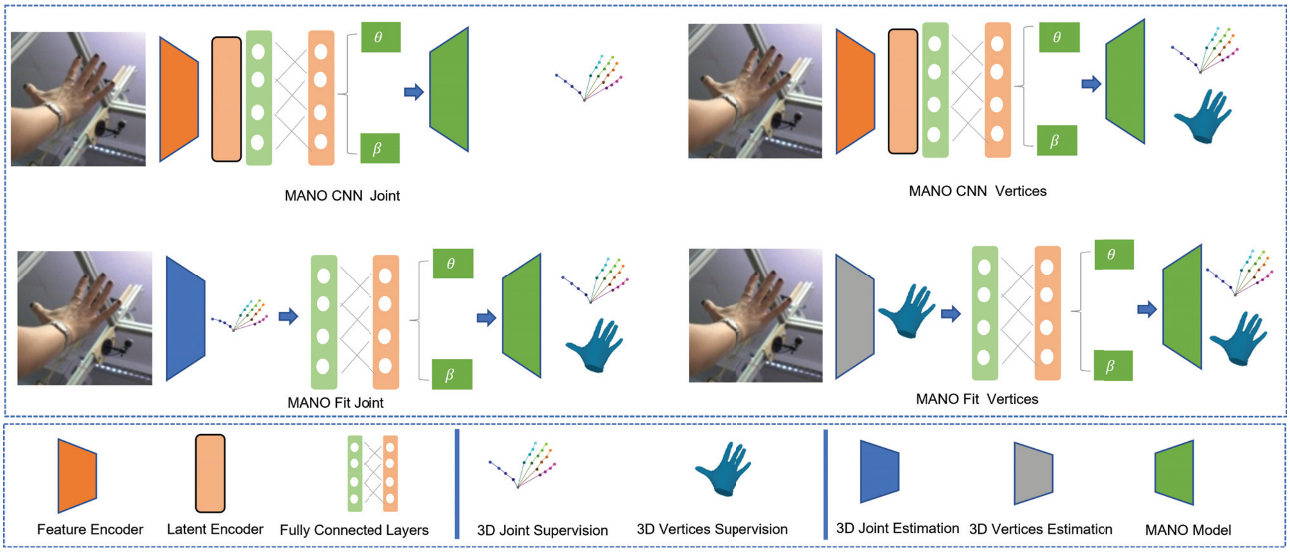

Figure 1. Overview of MANO parameter estimation pipelines.

baseline1 and the joint ground truth – a gap ignored by previous works.

Dataset	DexYCB GT		DexYCB	
Method	MPJPE	MPVPE	MPJPE	MPJPE
Baseline1(w. $\mathcal{J}$)	2.24	0	9.05	9.32
Baseline2 (w. NN)	2.34	0	9.22	9.62
Ours (VAE)	**1.83**	0	**8.96**	**9.33**

Table 2. Joint from mesh discussion on DexYCB test sets. Best scores are highlighted in **Bold**.

3. More Qualitative Results

In addition to the results shown in the main paper's experimental section, we provide more qualitative results for our 3D hand reconstruction in Fig. 2, Fig. 3 and Fig. 4 of this section. In those figures, we provide the multi-view and key points results of the hand reconstruction (See Fig. 2). We can see from the results that our key points and hand meshes are well aligned with the hand joints and surface in the images. The multi-view results indicate that our model is capable of generating a proper estimation of the invisible area with single-view RGB input. In addition, compared with the state-of-the-art works, our hand meshes are well aligned and plausible (See Fig. 3). These results verify the effectiveness of our proposed integrated pipeline. Furthermore, the two-hand interaction and hand-object interaction results show that interaction refinement improves the interaction quality (See Fig. 4), which verifies the effectiveness of our proposed model in the interaction refinement task.

References

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

[2] Xingyu Chen, Yufeng Liu, Dong Yajiao, Xiong Zhang, Chongyang Ma, Yanmin Xiong, Yuan Zhang, and Xiaoyan Guo. Mobrecon: Mobile-friendly hand mesh reconstruction from monocular image. In *CVPR*, 2022. 1

[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. 1

[4] 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. 4

[5] Dominik Kulon, Haoyang Wang, Riza Alp Güler, Michael Bronstein, and Stefanos Zafeiriou. Single image 3d hand reconstruction with mesh convolutions. In *BMVC*, 2019. 1

[6] Jiefeng Li, Chao Xu, Zhicun Chen, Siyuan Bian, Lixin Yang, and Cewu Lu. Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation. In *CVPR*, 2021. 1

[7] Mengcheng Li, Liang An, Hongwen Zhang, Lianpeng Wu, Feng Chen, Tao Yu, and Yebin Liu. Interacting attention graph for single image two-hand reconstruction. In *CVPR*, 2022. 1

[8] Yuxiao Zhou, Marc Habermann, Weipeng Xu, Ikhsanul Habibie, Christian Theobalt, and Feng Xu. Monocular real-time hand shape and motion capture using multi-modal data. In *CVPR*, 2020. 1

[9] 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. 1, 4

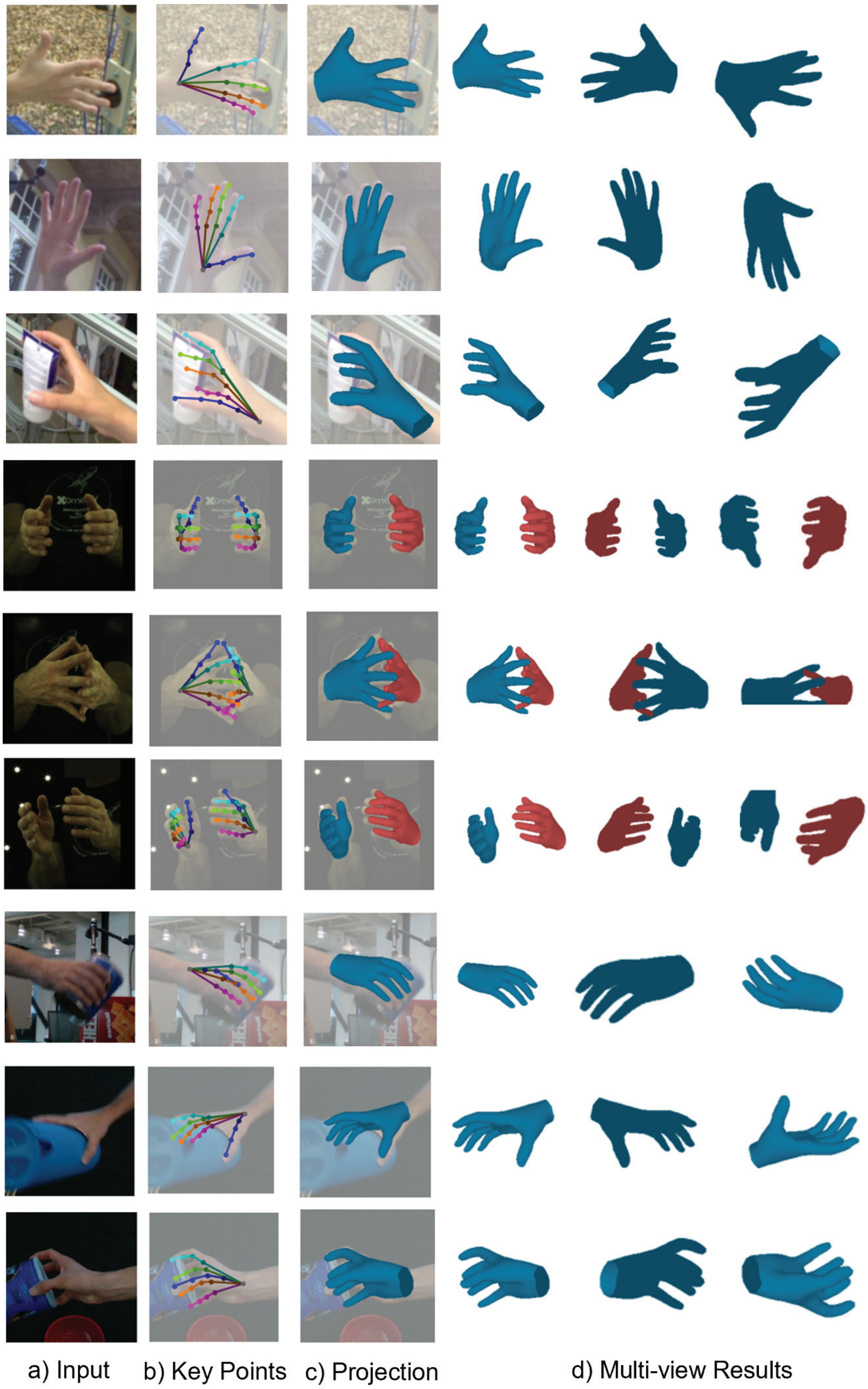

Figure 2. More results from our pipeline. From left to right the columns correspond to RGB input (a), 2D key points results (b), projection of the reconstructed mesh on the original image (c), and the multi-view visualization of reconstructed 3D meshes (d). Our pipeline yields highly accurate and plausible 3D hand meshes.

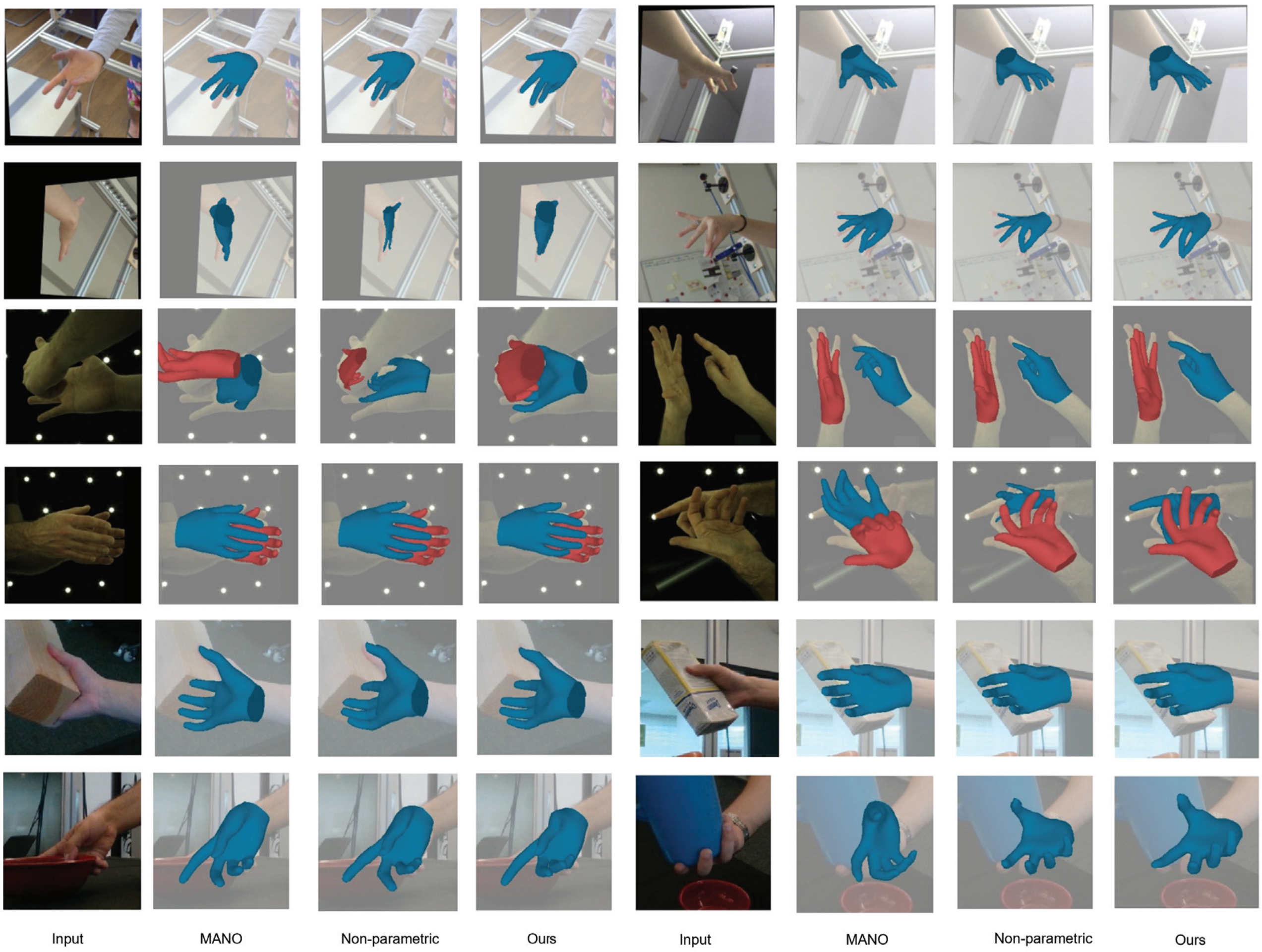

Figure 3. Hand shape reconstruction results. For each quartet, from left to right the columns correspond to RGB input, MANO-based method: MANO CNN [9], non-parametric model-based method: GCN-vert [4] and our method in camera view.

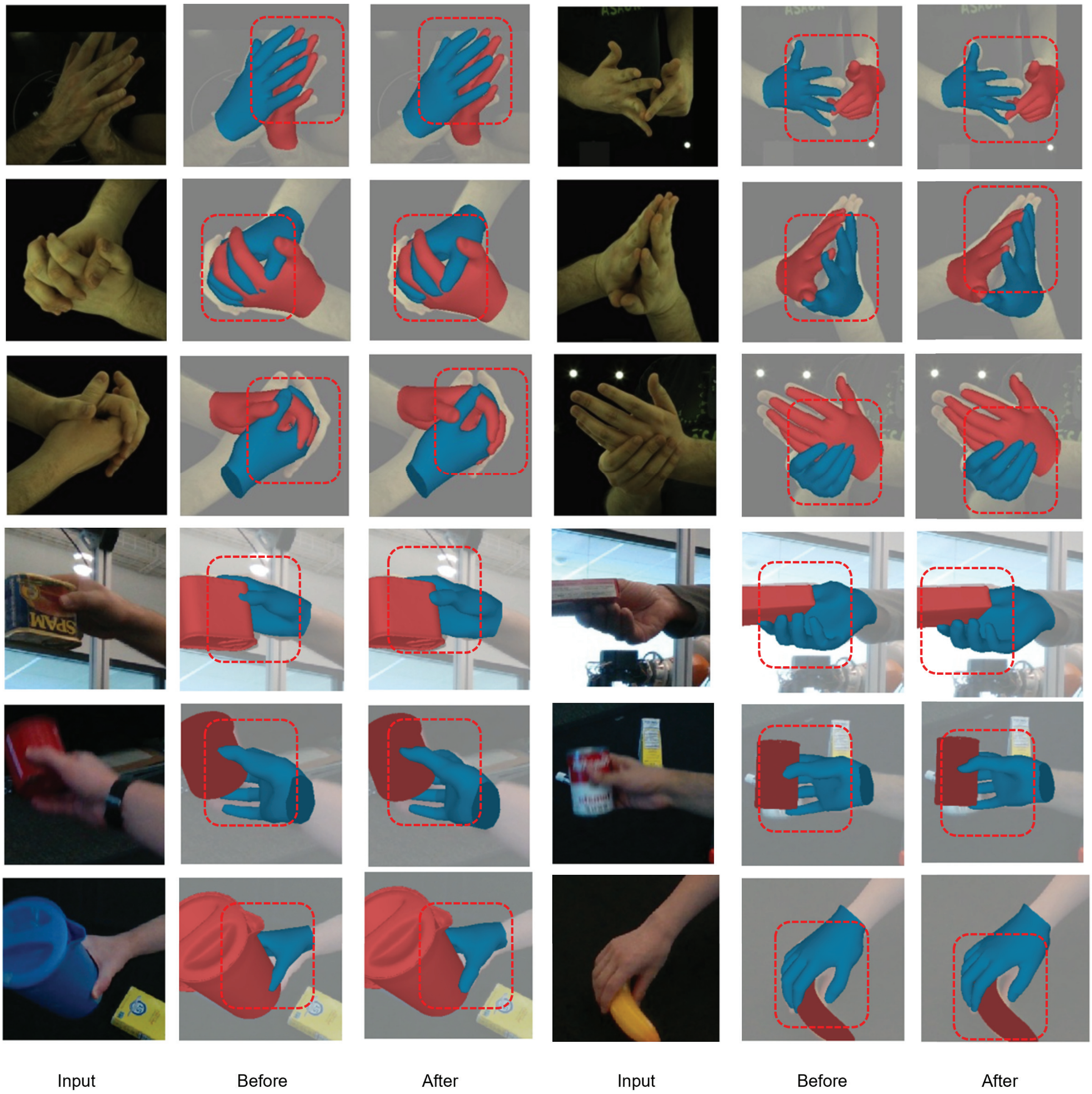

Figure 4. Interaction refinement results. For each triplet, the left to right columns correspond to input RGB images, and our meshes before and after interaction refinement. Red boxes highlight the interaction refinement regions.