







3D Interacting Hand Pose Estimation by Hand De-occlusion and Removal – Supplementary Material –

Hao Meng^{1,3*}, Sheng Jin^{2,3*}, Wentao Liu^{3,4}, Chen Qian³
Mengxiang Lin¹, Wanli Ouyang^{4,5}, and Ping Luo²

¹ Beihang University ² The University of Hong Kong ³ SenseTime
Research and Tetras.AI ⁴ Shanghai AI Lab ⁵ The University of Sydney
hao_meng@163.com, {jinsheng, liuwentao, qianchen}@tetras.ai
linmx@buaa.edu.cn, wanli.ouyang@sydney.edu.au, pluo@cs.hku.hk

1 Video Demo

To justify the generalization ability and the potential of our proposed method in real-world applications, we run our approach on several video clips from the Tzionas dataset [2]. Note that our models are only trained on InterHand2.6M V1.0 dataset [1] and our proposed AIH dataset. Tzionas dataset [2] is *unseen* during training.

In the video demo¹, we compare our approach with ‘Baseline’ which is a Single-Hand Pose Estimator (SHPE). For fair comparisons, both ‘Ours’ and ‘Baseline’ employ the same SHPE [3] trained on the ‘ALL’ branch of InterHand2.6M V1.0 dataset [1]. Note that this model is the same as the one used in Sec. 5.6 (Fig. 5) of the main paper. The pose results are directly obtained from the output of the SHPE model without temporal smoothing.

We first visualize the predicted amodal/visible mask of both hands. Given the segmentation mask, we obtain the corresponding single-hand box (‘red’ for the right hand, and ‘green’ for the left hand). We also demonstrate the results of Hand De-occlusion and Removal Module (HDRM). In order to tackle the severe hand-hand occlusion problem, HDRM applies the hand de-occlusion technique to recover the appearance (texture) in the occluded region. In the meanwhile, it also removes (inpaints) the distracting hand to handle the ambiguity caused by the homogeneous appearance of hands. Finally, our approach obtains better 3d hand pose estimation results.

Although the quality of the image recovery is satisfactory in most cases, there are still some problems in some difficult situations. For example, the boundary of the hand segmentation can be over-smoothed, leading to undesirable artifacts around the hand. This problem can be mitigated by applying an advanced amodal/visible mask segmentation model, which we will explore in the future.

* Equal Contribution.

¹ Our video demo can be downloaded from https://connecthkuhk-my.sharepoint.com/:v:/g/personal/js20_connect_hku_hk/EW_S3kZu97xP1Mk_HQLAJVMBtizU48sGh4jXwwUuyugFRw?e=u2GB6I.

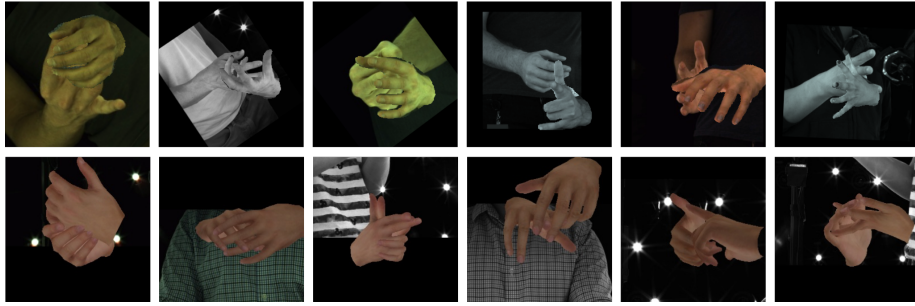


Fig. S1: **Top:** More examples of our proposed AIH_Syn dataset. **Bottom:** More examples of our proposed AIH_Render dataset.

2 More Examples of AIH Dataset

In this section, we present more examples of our proposed Amodal InterHand (AIH) dataset. Our AIH dataset consists of two parts: AIH_Syn and AIH_Render. AIH_Syn is constructed by copy-and-paste while AIH_Render is constructed by rendering the textured interacting hand mesh to the image plane. As shown in Fig S1, both AIH_Syn and AIH_Render have great diversity in hand poses, textures, occlusion and interaction types.

References

1. Moon, G., Yu, S.I., Wen, H., Shiratori, T., Lee, K.M.: Interhand2.6m: A dataset and baseline for 3d interacting hand pose estimation from a single rgb image. In: *Eur. Conf. Comput. Vis.* pp. 548–564. Springer (2020) [1](#)
2. Tzionas, D., Ballan, L., Srikantha, A., Aponte, P., Pollefeys, M., Gall, J.: Capturing hands in action using discriminative salient points and physics simulation. *Int. J. Comput. Vis.* **118**(2), 172–193 (2016) [1](#)
3. Zhou, Y., Habermann, M., Xu, W., Habibie, I., Theobalt, C., Xu, F.: Monocular real-time hand shape and motion capture using multi-modal data. In: *IEEE Conf. Comput. Vis. Pattern Recog.* pp. 5346–5355 (2020) [1](#)