Supplementary Material for Harmonious Feature Learning for Interactive Hand-Object Pose Estimation

Zhifeng Lin[1] Changxing Ding[1,2*] Huan Yao[1] Zengsheng Kuang[1] Shaoli Huang[3]
[1] South China University of Technology [2] Pazhou Lab, Guangzhou [3] Tencent AI-Lab, Shenzhen
eezhifenglin@mail.scut.edu.cn, chxding@scut.edu.cn
{mehuanyao,ftkuangzs}@mail.scut.edu.cn, shaolihuang@tencent.com

Methods	Joint	Mesh	F@5	F@15	Object
Pose2Mesh et al. [8]	12.5	12.7	44.1	90.9	No
Hasson et al. [4]	11.4	11.4	42.8	93.2	Yes
I2L-MeshNet [9]	11.2	13.9	40.9	93.2	No
Hasson et al. [5]	11.0	11.2	46.4	93.9	Yes
Hampali et al. [10]	10.7	10.6	50.6	94.2	Yes
METRO [3]	10.4	11.1	48.4	94.6	No
Liu et al. [1]	10.1	9.7	53.2	95.2	Yes
ArtiBoost [7]	11.4	10.9	48.8	94.4	Yes
Keypoint Trans. [6]	10.8	-	-	-	Yes
HandOccNet [2]	9.1	8.8	56.4	96.3	No
Ours(-)	9.1	9.0	56.1	96.1	Yes
Ours	**8.9**	**8.7**	**57.5**	**96.5**	Yes

Table 1. Performance comparison with state-of-the-art methods on the hand pose estimation task on the HO3D dataset. The last column indicates whether a method performs the object 6D pose estimation task. (-) denotes HFL-Net with the smaller backbone model.

In this supplementary material, we further demonstrate the effectiveness of HFL-Net by adopting a smaller backbone model. Specifically, we reduce the number of stage-2 and stage-3 layers by one half, respectively. Considering that there are two independent sets of stage-2 and stage-3 layers, the model size of our backbone is the same as the common ResNet-50 [11] model adopted in [1, 2, 6].

Performance comparisons on the HO3D [10] database are summarized in Table 1 and Table 2. The PAMPJPE and PAMPVPE performance of the smaller HFL-Net are 9.1mm and 9.0mm, respectively. They are only slightly lower than the performance achieved with our original backbone by 0.2mm and 0.3mm, respectively.

The performance of HFL-Net with the smaller backbone still significantly outperforms its baseline model [1] by 1.0mm and 0.7mm on PAMPJPE and PAMPVPE, respectively. It is worth noting that we adopt smaller image size (256×256 pixels) than that in [1] (512×512 pixels). The time cost of our smaller backbone and that in [1] are 3.16ms and 7.5ms per image on a Titan V GPU. This comparison means that HFL-Net can be more efficiently utilized in practice.

*Corresponding author.

Methods	cleanser	bottle	can	average
Liu et al. [1]	88.1	61.9	**53.0**	67.7
Ours(-)	**89.2**	75.8	50.7	71.9
Ours	81.4	**87.5**	52.2	**73.3**

Table 2. Performance comparison with state-of-the-art methods on the object 6D pose estimation task on the HO3D dataset. (-) denotes HFL-Net with the smaller backbone model.

Moreover, with the smaller backbone, our HFL-Net still achieves comparable performance with [2] over all metrics. It is worth noting that [2] performs the hand pose estimation task only, which means it is free from the interference caused by the object pose estimation task. We also conduct comparisons on the object 6D pose estimation task in Table 2. With the smaller backbone model, the average ADD-0.1 score of HFL-Net is 71.9%, outperforming [1] by 4.2%.

The above rigours comparisons further justify the effectiveness of HFL-Net. We attribute the advantage of HFL-Net to its harmonious feature learning scheme, which not only significantly relieves the competition between the hand and object pose estimation tasks, but also facilitates the mutual enhancement between the hand and object features.

References

[1] S. Liu, H. Jiang, J. Xu, S. Liu, and X. Wang. Semi-supervised 3d hand-object poses estimation with interactions in time. In *CVPR*, 2021. 1

[2] J. Park, Y. Oh, G. Moon, H. Choi, and K. Lee. HandOccNet: Occlusion-Robust 3D Hand Mesh Estimation Network. In *CVPR*, 2022. 1

[3] K. Lin, L. Wang, and Z. Liu. End-to-end human pose and mesh reconstruction with transformers. In *CVPR*, 2021. 1

[4] Y. Hasson, B. Tekin, F. Bogo, I. Laptev, M. Pollefeys, and C. Schmid. Leveraging photometric consistency over time for sparsely supervised hand-object reconstruction. In *CVPR*, 2020. 1

[5] Y. Hasson, G. Varol, D. Tzionas, I. Kalevatykh, M. Black, I. Laptev, and C. Schmid. Learning joint reconstruction of hands and manipulated objects. In *CVPR*, 2019. 1

[6] S. Hampali, S. Sarkar, M. Rad, and V. Lepetit. Keypoint Transformer: Solving Joint Identification in Challenging Hands and Object Interactions for Accurate 3D Pose Estimation. In *CVPR*, 2022. 1

[7] L. Yang, K. Li, X. Zhan, J. Lv, W. Xu, J. Li, and C. Lu. ArtiBoost: Boosting Articulated 3D Hand-Object Pose Estimation via Online Exploration and Synthesis. In *CVPR*, 2022. 1

[8] H. Choi, G. Moon, and K. Lee. Pose2mesh: Graph convolutional network for 3d human pose and mesh recovery from a 2d human pose. In *ECCV*, 2020. 1

[9] G. Moon, and K. Lee. I2l-meshnet: Image-to-lixel prediction network for accurate 3d human pose and mesh estimation from a single rgb image. In *ECCV*, 2020. 1

[10] S. Hampali, M. Rad, M. Oberweger, and V. Lepetit. Honnotate: A method for 3d annotation of hand and object poses. In *CVPR*, 2020. 1

[11] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 1