Domain Adaptive Hand Keypoint and Pixel Localization in the Wild
– Supplementary Material –

In this supplementary material, we present details of datasets, preprocessing, and augmentation, training and evaluation protocol, and additional qualitative analysis.

1.1 Dataset Details

- **DexYCB** [2] contains 582K RGB-D frames captured by 10 subjects interacting 20 different YCB objects [1] from eight different views. In our experiment, we split the dataset by the subject IDs to create train, validation, and test sets with 212K, 71K, and 80K images, respectively.
- **HO3D** [7] contains 103K RGB-D frames captured by 10 subjects interacting 10 different YCB objects [1] from a single third-person view. In our experiment, we randomly split the video sequences to train, validation, and test sets with 51K, 12K, and 8K images, respectively.
- **HanCo** [13] is an extended FreiHAND [14] dataset captured in a multi-view camera setup with eight cameras, which consists of 518K, 106K, and 104K RGB images for training, validation, and testing, respectively. The backgrounds are randomly synthesized using diverse scenery images.
- **FPHA** [4] is an egocentric video dataset capturing users' actions in daily indoor environments from a first-person perspective, and their hand poses are tracked by hand magnetic sensors. It contains 69K training images and 16K validation images. Due to lacking hand mask annotation, we annotated 50 hand masks in the validation set.
- **Ego4D** [6] is a collection of daily-life egocentric activity videos lasting over 3,000 hours and gathered across the world. Due to the lack of annotation for the two tasks, we show qualitative examples in our experiments. We treated each video sequence as the domain to adapt.

1.2 Preprocessing and Augmentation

For creating an input of a training network, we assumed to have hand center positions, cropped hand regions of the original images, and resized them to 128×128 pixels. To extract hand centers and regions in Ego4D videos without ground truth, we used an off-the-shelf hand detector [11]. Inspired by [9,3,8], we used two different augmentation sets: strong augmentation for the student's learning (Equation 4) and weak augmentation for the teacher's learning (Equation 5). We used horizontal flip, rotation, transition, gaussian blur, brightness/contrast jitter, hue/saturation/input value jitter, and cutout as the strong augmentation. In contrast, we adopted horizontal flip, rotation, transition, and gaussian blur as the weak augmentation.

1.3 Network Architecture and Evaluation

For the design of our multi-task baseline model, we employed an hourglass network [10] as the backbone and the keypoint regression branch. We added 1d-convolution to its intermediate features to predict hand pixel labels. Following hand keypoint regression methods [12,10,5], we optimized 2D joint heatmaps for each 2D ground-truth joint location instead of joint coordinates.

We also provide the details of out evaluation, namely, MPE, PCK, and IoU. MPE (px) indicates the euclidean error per joint in the image coordinate. PCK (%) represents the percentage of joints whose MPE is smaller than a given joint error threshold, which is calculated by the area under the curve (AUC) over the joint error range [0, 20 px]. IoU (%) measures the overlap over two masks. We report the average score (Avg.) over PCK and IoU to evaluate multi-task performance.

1.4 Qualitative Analysis

In Figs. 6, 7, 8, and 9, we show additional qualitative results of our proposed method. As shown in Fig. 6, our method performed well when complex hand-object interactions occur on HO3D and FPHA and when the backgrounds are diverse on HanCo. In Fig. 7, we show qualitative comparison between GAC and C-GAC (Ours-Full). Our full method particularly improved keypoint regression compared to the simple consistency baseline, GAC. Our method (right) corrected the keypoint prediction of the GAC (left), which contains incorrect predictions on the position of the thumb (red).

Our method also demonstrated improved performance on Ego4D, an egocentric video dataset collected across various countries, cultures, ages, indoors/outdoors, and performing tasks with hands. In particular, we observed that our method successfully adapted to various imaging conditions, such as outdoor environments (rows 1 and 2 in Fig. 8), extremely dark environments (rows 3 to 6 in Fig. 8), the second person's hands in social interactions (row 7 in Fig. 8), *e.g.*, playing board games, and indoor environments (Fig. 9), *e.g.*, where people perform cooking, cleaning, fitness, DIY, painting, and crafting.

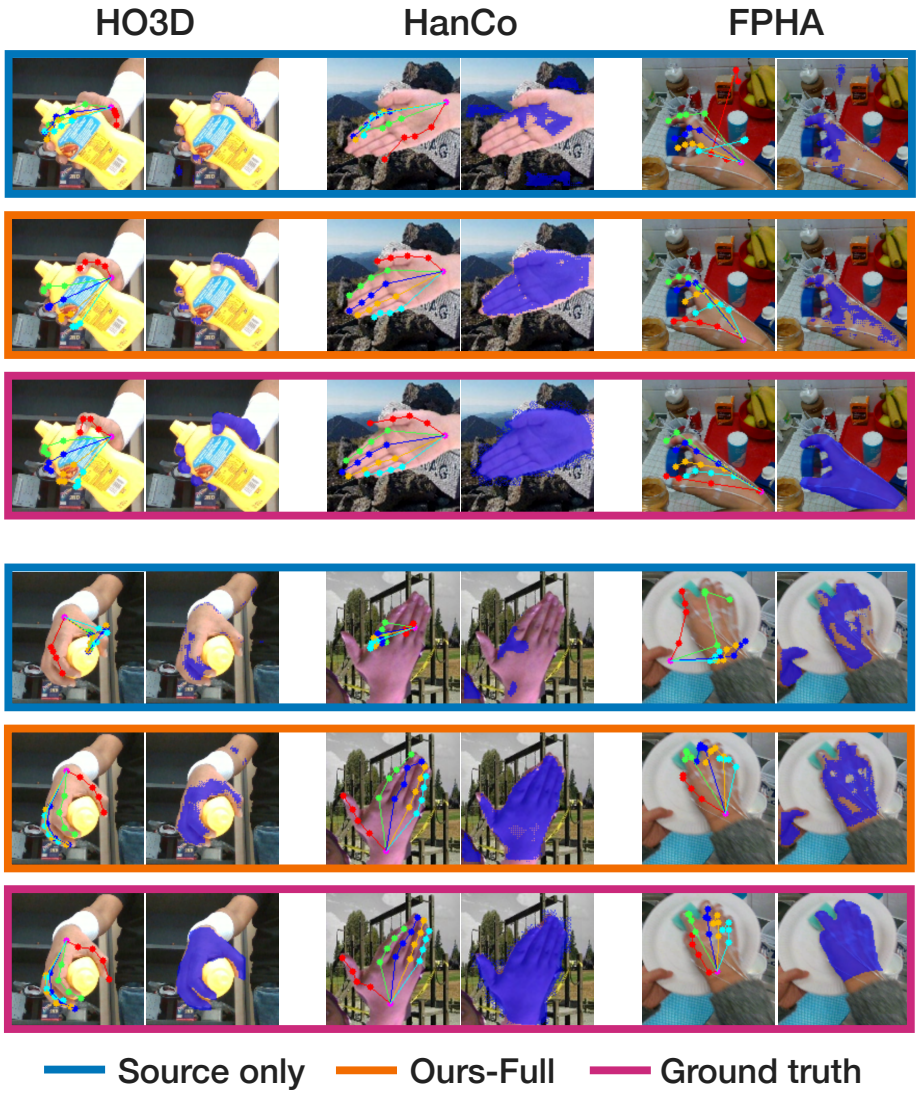

Fig. 6: **Additional qualitative results on HO3D [7], HanCo [13], and FPHA [4].**

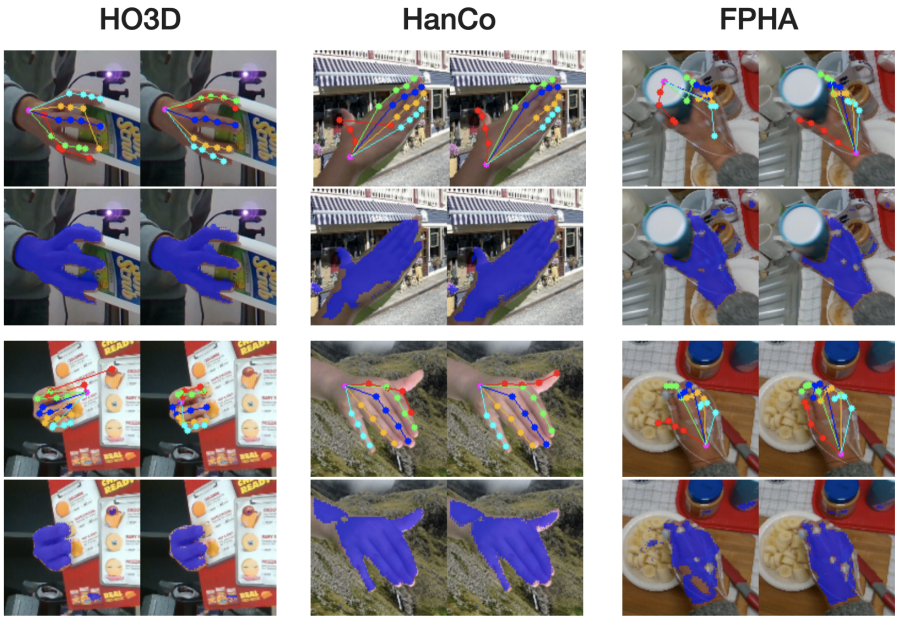

Fig. 7: **Comparison between GAC and C-GAC (Ours-Full).** Left: GAC, Right: C-GAC (Ours-Full).

Fig. 8: **Additional qualitative results on Ego4D [6].**

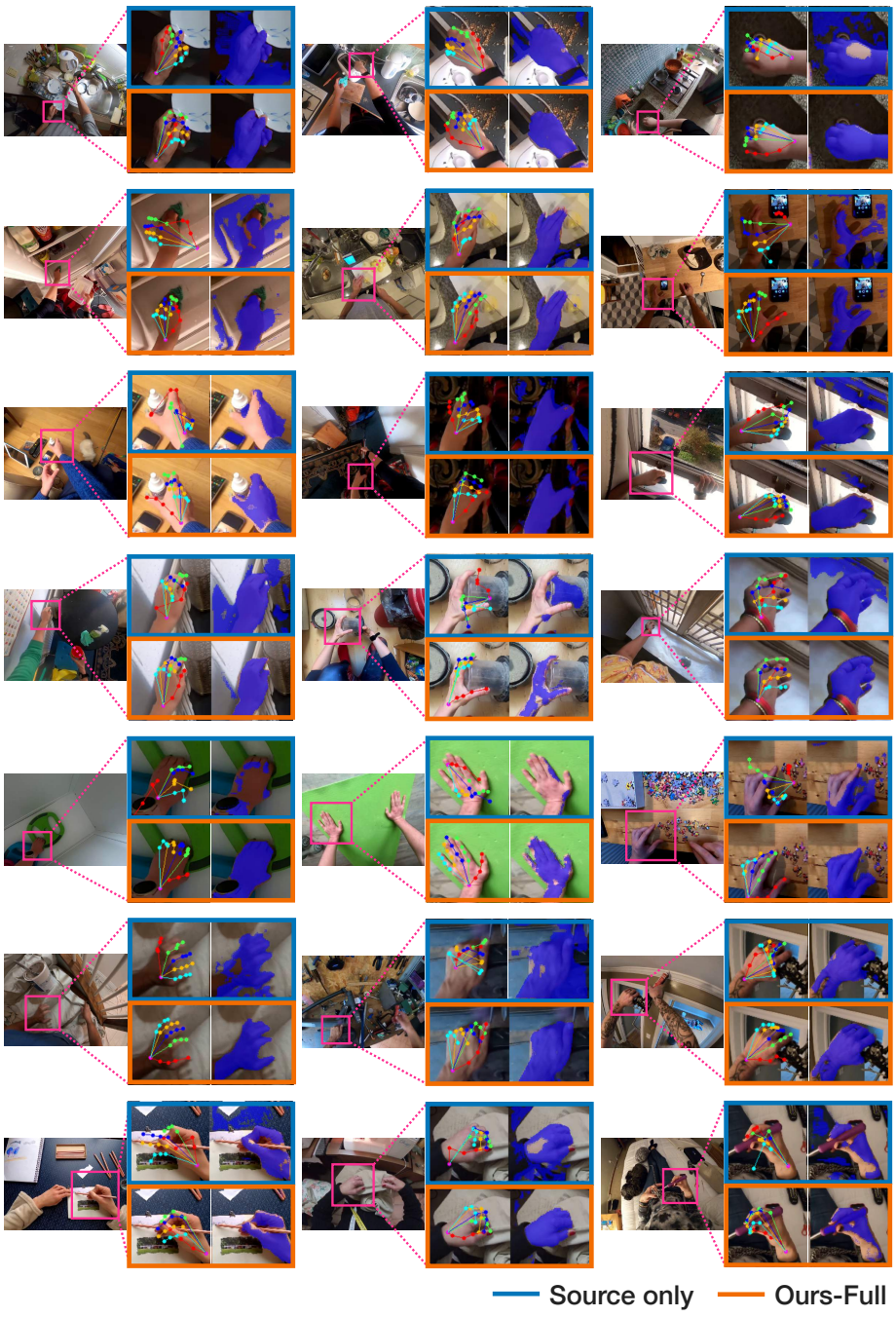

Fig. 9: **Additional qualitative results on Ego4D [6].**

References

1. B. Çalli, A. Walsman, A. Singh, S. S. Srinivasa, P. Abbeel, and A. M. Dollar. Benchmarking in manipulation research: Using the yale-cmu-berkeley object and model set. *IEEE Robotics Automation Magazine*, 22(3):36–52, 2015. 1
2. Y.-W. Chao, W. Yang, Y. Xiang, P. Molchanov, A. Handa, J. Tremblay, Y. S. Narang, K. Van Wyk, U. Iqbal, S. Birchfield, J. Kautz, and D. Fox. DexYCB: A benchmark for capturing hand grasping of objects. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9044–9053, 2021. 1
3. J. Deng, W. Li, Y. Chen, and L. Duan. Unbiased mean teacher for cross-domain object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4091–4101, 2021. 1
4. G. Garcia-Hernando, S. Yuan, S. Baek, and T.-K. Kim. First-person hand action benchmark with RGB-D videos and 3D hand pose annotations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 409–419, 2018. 1, 3
5. L. Ge, Z. Ren, Y. Li, Z. Xue, Y. Wang, J. Cai, and J. Yuan. 3D hand shape and pose estimation from a single RGB image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10833–10842, 2019. 2
6. K. Grauman, A. Westbury, E. Byrne, Z. Chavis, A. Furnari, R. Girdhar, J. Hamburger, H. Jiang, M. Liu, X. Liu, M. Martin, T. Nagarajan, I. Radosavovic, S. K. Ramakrishnan, F. Ryan, J. Sharma, M. Wray, M.g Xu, E. Zhongcong Xu, C. Zhao, S. Bansal, D. Batra, V. Cartillier, S. Crane, T. Do, M. Doulaty, A. Erapalli, C. Feichtenhofer, A. Fragomeni, Q. Fu, C. Fuegen, A. Gebreselasie, C. Gonzalez, J. Hillis, X. Huang, Y. Huang, W. Jia, W. Khoo, J. Kolar, S. Kottur, A. Kumar, F. Landini, C. Li, Y. Li, Z. Li, K. Mangalam, R. Modhugu, J. Munro, T. Murrell, T. Nishiyasu, W. Price, P. R. Puentes, M. Ramazanova, L. Sari, K. Somasundaram, A. Southerland, Y. Sugano, R. Tao, M. Vo, Y. Wang, X. Wu, T. Yagi, Y. Zhu, P. Arbelaez, D. Crandall, D. Damen, G. M. Farinella, B. Ghanem, V. K. Ithapu, C. V. Jawahar, H. Joo, K. Kitani, H. Li, R. Newcombe, A. Oliva, H. Soo Park, J. M. Rehg, Y. Sato, J. Shi, M. Z. Shou, A. Torralba, Lo Torresani, M.i Yan, and J. Malik. Ego4D: Around the world in 3, 000 hours of egocentric video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18995–19012, 2022. 1, 5, 6
7. S. Hampali, M. Rad, M. Oberweger, and V. Lepetit. Honnotate: A method for 3D annotation of hand and object poses. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3196–3206, 2020. 1, 3
8. Y.-J. Li, X. Dai, C.-Y. Ma, Y.-C. Liu, K. Chen, B. Wu, Z. He, K. Kitani, and P. Vajda. Cross-domain object detection via adaptive self-training. *CoRR*, abs/2111.13216, 2021. 1
9. Y.-C. Liu, C.-Y. Ma, Z. He, C.-W. Kuo, K. Chen, P. Zhang, B. Wu, Z. Kira, and P. Vajda. Unbiased teacher for semi-supervised object detection. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021. 1
10. A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, volume 9912, pages 483–499, 2016. 2
11. D. Shan, J. Geng, M. Shu, and D. Fouhey. Understanding human hands in contact at internet scale. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9866–9875, 2020. 1

12. S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. Convolutional pose machines. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4724–4732, 2016. 2
13. C. Zimmermann, M. Argus, and T. Brox. Contrastive representation learning for hand shape estimation. *CoRR*, abs/2106.04324, 2021. 1, 3
14. C. Zimmermann, D. Ceylan, J. Yang, B. Russell, M. J. Argus, and T. Brox. FreiHAND: A dataset for markerless capture of hand pose and shape from single RGB images. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 813–822, 2019. 1