2nd Place Solution Technical Report for Hands'24 ARCTIC Challenge from Team ACE

Congsheng Xu[1*], Yitian Liu[1*], Yi Cui[2], Jinfan Liu[1],
Yichao Yan[1], Weiming Zhao[1], Yunhui Liu[3], Xingdong Sheng[3]
[1]Shanghai Jiao Tong University , [2]Fudan University, [3]Lenovo Research
acondaway@sjtu.edu.cn, violetevar@sjtu.edu.cn

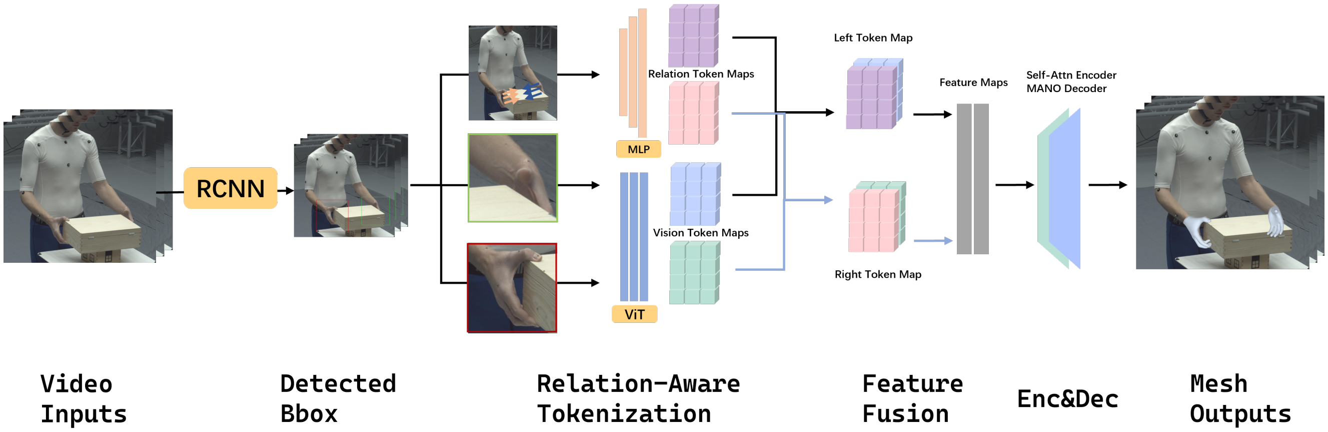


Figure 1. Illustration of the overall pipeline during pre-processing optimization. Both hands are encoded separately using a ViT, while the relation-based information is processed through an MLP. The outputs are concatenated and flattened to compute a feature map, which is then used to estimate the MANO hand poses via a lightweight MANO decoder.

Abstract

Humans frequently interact with various objects using both hands, making the accurate reconstruction of these interactions essential for understanding human-object dynamics. We propose a new method aimed at the bimanual category-agnostic reconstruction task, which specifically focuses on reconstructing interactions between hands and objects in a manner independent of object categories. We leverage the advanced segmentation capabilities of SAM-2, combined with a relation-aware tokenization encoder-decoder mechanism, to achieve a performance improvement of approximately 12.6% *over the baseline on the key metric CD_h, which emphasizes the quality of hand-object interaction during reconstruction.*

1. Introduction

Hand-object interaction reconstruction from monocular temporal-sequential inputs is an intriguing and fundamental task with profound implications for emerging technologies such as Virtual Reality (VR) and Augmented Reality (AR). Several datasets [2, 5, 8] have been proposed to advance the task, and effective algorithms have been developed to tackle the problem. Specifically, algorithms [1] that reconstruct monocular inputs without object templates—so-called category-agnostic approaches—demonstrate strong potential for improved generalization. Although methods like HOLD perform well in HOI reconstruction, they remain limited to single-hand interactions. Our method, therefore, aims to solve this limitation. Thanks to the Hands'24 challenge, our team spent approximately one month implementing the proposed algorithm on the ARCTIC dataset. At the pre-processing step, based on the off-the-shelf hand mesh reconstruction method [6], our method utilizes a bimanual-aware tokenization approach inspired by 4DHands [4] to integrate two-hand information, enabling

[*] Equal Contribution.
[†] The code is available at https://github.com/VioletEvar/ACE

better modeling of HOI with both hands. We also proposed a self-enhancing method based on a sequence window to reinforce spatio-temporal properties during the mano registration process, and utilized the powerful SAM-2 segmentation tool [7] to provide more accurate interaction cues.

2. Method

Our method for enhancing the data pre-processing pipeline consists of three parts: acquiring more accurate mask data using SAM-2, improving hand mesh generation by adding inter-hand information to the original hand tracker, and registering better MANO parameters using a self-boosting method.

2.1. SAM-2-based Mask Extraction

HOLD uses SAM-Track to extract hand and object masks from the input video, where it identifies a selected entity of interest and extracts corresponding masks throughout the video. However, we noticed that SAM-Track occasionally confuses the left and right hands and performs poorly when the entity of interest is heavily occluded by objects or other hands. To address these issues, we used SAM-2, which offers higher accuracy and supports multi-entity tracking, to extract precise hand and object masks.

2.2. Relation-aware Two-Hand Tokenization

The HaMeR hand-tracker used in the original HOLD pre-processing pipeline is quite powerful. However, it estimates left and right hand poses separately, without leveraging information from the other hand, which can limit performance in bimanual tasks. Therefore, inspired by 4DHands, we incorporated Relation-aware Two-Hand Tokenization (RAT) into HaMeR to improve its performance on bimanual HOI datasets, as shown in Figure 1. First, for each frame of the input sequence, we detect hands using a Region-based Convolutional Neural Network (R-CNN) and acquire their bounding box. Next, we compute the relative distance and overlapping maps between the two hands, concatenate them along the last dimension, and use an MLP to generate relation token maps. The relation token maps are then concatenated with vision token maps generated by a ViT in the original HaMeR pipeline to create comprehensive token maps that encode both hands' interactions. Finally, we apply a self-attention transformer to encode the token maps and utilize a MANO decoder to regress MANO parameters of the two hands. For training data, we mainly used H2O3D [3], as it provides detailed hand-object interaction data, which is crucial for improving the accuracy of bimanual pose estimation.

2.3. Self Boost MANO Registration

The R-CNN can sometimes fail to detect heavily occluded hands, resulting in missing hand data. HOLD addresses this issue by linearly interpolating missing frames. However, when multiple consecutive frames are interpolated, the hand poses tend to become inaccurate due to the lack of direct observation and cumulative errors. Also, in cases of heavy occlusion, the estimates made by HaMeR in adjacent frames can also be inaccurate, compounding the errors in interpolated frames. To mitigate these errors, we propose a self-boosting method to more accurately register MANO parameters. When R-CNN fails to detect a hand in the current frame, we use the stored 3D vertices from the previous frame as input for the current frame, then verify the result's accuracy against a predefined loss threshold. If the result meets the threshold, we update the stored 3D vertices for use in subsequent missing frames.

3. Result

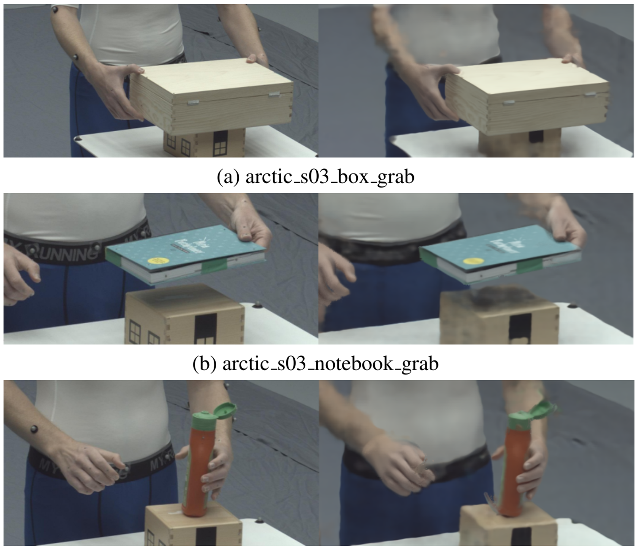
The results of the evaluated metrics and the comparison with the baseline are presented in Table 1. As shown, our method demonstrates a significant improvement over the baseline. Notably, on the core challenge metric CD_h, our method surpasses the baseline by **12.6%**. These results indicate that the RAT encoding technique and Self-boost MANO Registration effectively reduce errors in bimanual pose estimation and mitigate temporal inconsistencies. We also compared the performance of SAM-1 and SAM-2, with results showing that SAM-2 significantly outperforms SAM-1 in the segmentation task.

Table 1. Comparison of key performance metrics across methods

Models	MPJPE↓	CD_h↓
Baseline	25.91	114.73
Ours w/ SAM-1	25.80	108.39
Ours w/ SAM-2	**25.29**	**100.33**

4. Visualization

Figure 2 presents visualized results produced by our method. In the comparison graphs, the left side shows the ground truth, which corresponds to the frames from the ARCTIC dataset, while the right side displays our reconstructed results.



(a) arctic_s03_box_grab

(b) arctic_s03_notebook_grab

(c) arctic_s03_ketchup_grab

Figure 2. Visual comparison of g.t. and reconstructed results

References

[1] Zicong Fan, Maria Parelli, Maria Eleni Kadoglou, Muhammed Kocabas, Xu Chen, Michael J Black, and Otmar Hilliges. HOLD: Category-agnostic 3d reconstruction of interacting hands and objects from video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 494–504, 2024. 1

[2] Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J. Black, and Otmar Hilliges. ARCTIC: A dataset for dexterous bimanual hand-object manipulation. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1

[3] Shreyas Hampali, Sayan Deb Sarkar, Mahdi Rad, and Vincent Lepetit. Keypoint transformer: Solving joint identification in challenging hands and object interactions for accurate 3d pose estimation. In *CVPR*, 2022. 2

[4] Dixuan Lin, Yuxiang Zhang, Mengcheng Li, Yebin Liu, Wei Jing, Qi Yan, Qianying Wang, and Hongwen Zhang. 4dhands: Reconstructing interactive hands in 4d with transformers, 2024. 1

[5] Xintao Lv, Liang Xu, Yichao Yan, Xin Jin, Congsheng Xu, Shuwen Wu, Yifan Liu, Lincheng Li, Mengxiao Bi, Wenjun Zeng, and Xiaokang Yang. Himo: A new benchmark for full-body human interacting with multiple objects, 2024. 1

[6] Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. Reconstructing hands in 3D with transformers. In *CVPR*, 2024. 1

[7] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 2

[8] Omid Taheri, Nima Ghorbani, Michael J. Black, and Dimitrios Tzionas. GRAB: A dataset of whole-body human grasping of objects. In *European Conference on Computer Vision (ECCV)*, 2020. 1