Technical report of HCB team for Multiview Egocentric Hand Tracking Challenge on HANDS 2024 Challenge

Haohong Kuang[1], Yang Xiao[1†], Changlong Jiang[1], Jinghong Zheng[1], Hang Xu[1],
Ran Wang[2], Zhiguo Cao[1], Min Du[3], Zhiwen Fang[4,5], Joey Tianyi Zhou[6,7]

[1]Key Laboratory of Image Processing and Intelligent Control, Ministry of Education, School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan 430074, China

[2]School of Journalism and Information Communication, Huazhong University of Science and Technology, Wuhan 430074, China

[3]PICO, ByteDance, Beijing, 100089, China

[4]School of Biomedical Engineering, Southern Medical University, Guangzhou 510515, China

[5]Department of Rehabilitation Medicine, Zhujiang Hospital, Southern Medical University, Guangzhou 510280, China

[6]Centre for Frontier AI Research, Agency for Science, Technology and Research (A*STAR), Singapore

[7]Institute of High Performance Computing, Agency for Science, Technology and Research (A*STAR), Singapore

haohong_kuang, Yang_Xiao, changlongj, deepzheng, hang_xu, rex_wang, zgcao@hust.edu.cn,
bingwen.ai@bytedance.com, fzw310@smu.edu.cn, zhouty@cfar.a-star.edu.sg

Abstract

In this report, we introduce the method proposed for 3D hand pose tracking in HANDS@ECCV2024 challenge based on UmeTrack: multiview egocentric hand tracking challenge, which aims to track hand poses in calibrated stereo videos, utilizing a pre-calibrated hand shapes. We provide a novel method for estimating accurate 3D hand poses from stereo images. Given that stereo images can provide 3D spatial information, the main idea of our method is to leverage this stereo information to guide the estimation of MANO pose and transformation. Specifically, an effective cross-view feature fusion mechanism is utilized to accurately estimate the relative 2D poses, which are then lifted to 3D space and used to calculate the MANO transformation. Besides, an MANO pose optimization method is proposed to alleviate the performance gap between MANO positions and joint coordinates. Finally, our method achieves a FINGERTIP PCK AUC of 70.81% on the UmeTrack dataset, securing the first place in the challenge.

1. Introduction

With the development of the field of computer vision and the rise of XR/VR domains, 3D hand pose and shape estimation has become increasingly important in aiding the understanding of human interaction with the surrounding environment.

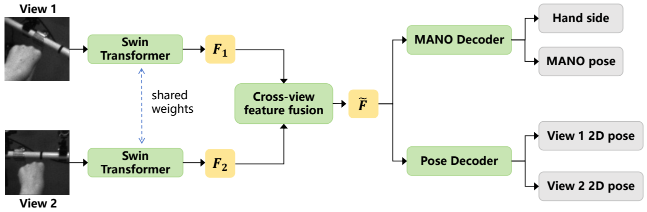

Figure 1. The features from the dual-view images are fused through cross-view feature fusion. The MANO Decoder is used to regress the hand pose parameters and hand side, while the Pose Decoder is used to predict the 2D joint coordinates for both views.

Previous methods[8, 11] primarily focus on hand pose estimation from single image, treating multi-view images as individual inputs. For 3D hand pose recovery, these approaches often neglect the inter-view relationships. In contrast, Remelli et al.[9] propose a method that uses camera parameters to transform image features into a unified latent representation space. Our proposed cross-view feature fusion, however, allows the model to learn the inter-view relationships on its own. The final results demonstrate the effectiveness of our approach. Some methods[2, 5, 12] estimate the MANO[10] parameters and the parameters of a weak-perspective camera, but this approach loses depth information and cannot accurately localize the hand in the world coordinate system.

In this work, we propose a novel method for 3D hand

[†]Yang Xiao is corresponding author (Yang_Xiao@hust.edu.cn).

This technical report has been patented and cannot be used for commercial purposes without our permission.

pose tracking from stereo videos. First, based on the A2J-Transformer[4] network, an effective cross-view feature fusion mechanism is utilized to accurately provide the MANO pose parameters and 2D hand poses from the dual views. Then, by lifting 2D poses to 3D space, the transformation matrix between MANO coordinates and world coordinates can be calculated. Finally, due to the ground-truth performance gap between MANO positions and joint coordinates in UmeTrack dataset, we introduce a MANO pose optimization method to improve performance. Besides, a temporal smoothing process is used to provide the temporal information. Ultimately, our approach prove to be effective, with the first place in pose tracking challenge on UmeTrack dataset.

2. Method

A2J-Transformer is a powerful 3D hand pose estimation method that takes a single RGB images as input and outputs a root-related 3D pose. Based on A2J-Transformer, we explore the potential of extending it to handle dual-view images. Our approach takes synchronized dual-view images as input and performs inter-view feature fusion. The enhanced model then outputs the MANO pose parameters and the 2D joint positions for both views. Through triangulation, we can reconstruct the 3D hand pose in the world coordinate system and subsequently derive the wrist transformation. This process will be explained in detail below.

Cross-view feature fusion. We use the Swin Transformer[6, 7] as our visual feature extractor to capture multi-scale image feature, which shares weights between the two viewpoints. Then, we employ the Deformable Transformer[13] to fuse the multi-scale image features from two different viewpoints to improve the representation of features, which enables information exchange between two perspectives, effectively improving the accuracy of localization hand joints between them.

Back to world system. Taking 2D point $c_1 = [u_1, v_1, 1]$ in one view as an example, its corresponding point in the other view is $c_2 = [u_2, v_2, 1]$,We can obtain the 3D point $C = [X, Y, Z, 1]$ by solving the following over-determined system of equations:

$$A = \begin{pmatrix} v_1p_{31}^1 - p_{21}^1 & v_1p_{32}^1 - p_{22}^1 & v_1p_{33}^1 - p_{23}^1 & v_1p_{34}^1 - p_{24}^1 \\ u_1p_{31}^1 - p_{11}^1 & u_1p_{32}^1 - p_{12}^1 & u_1p_{33}^1 - p_{13}^1 & u_1p_{34}^1 - p_{14}^1 \\ v_2p_{31}^2 - p_{21}^2 & v_2p_{32}^2 - p_{22}^2 & v_2p_{33}^2 - p_{23}^2 & v_2p_{34}^2 - p_{24}^2 \\ u_2p_{31}^2 - p_{11}^2 & u_2p_{32}^2 - p_{12}^2 & u_2p_{33}^2 - p_{13}^2 & u_2p_{34}^2 - p_{14}^2 \end{pmatrix} \tag{1}$$

$$A\mathbf{C} = \mathbf{0} \tag{2}$$

where P^i represents the projection matrix of the camera corresponding to point p_i:

$$P^i = \begin{bmatrix} p_{11}^i & p_{12}^i & p_{13}^i & p_{14}^i \\ p_{21}^i & p_{22}^i & p_{23}^i & p_{24}^i \\ p_{31}^i & p_{32}^i & p_{33}^i & p_{34}^i \end{bmatrix} \tag{3}$$

Because of the noise of prediction, we use SVD to solve Eq. 1 for the 3D point C with the least error.

After obtaining the 3D joints through triangulation, we use the Kabsch algorithm to solve for the wrist transformation based on the triangulation 3D joints and the MANO joints, where pose parameters are estimated by the model and wrist transformation is set to zero.

Decoder. The overall design of the Decoder follows that in A2J-Transformer. For the MANO Decoder, since no anchor prediction offsets are needed, we modify it to predict the MANO pose parameter and hand type. For the Pose Decoder, since 3D anchors are not needed, we set the anchor points on a 2D plane.

Augmentation. In the training phase, we use image flipping as a simple data augmentation method in which the handedness of the hand is also changed. At the time of testing, we use TTA, which considers the inputs of the original image and the flipped image thereby obtaining results with higher accuracy.

Postprocess. Since our network does not leverage temporal information from videos when predicting 2D joint coordinates, the triangulated 3D joint points exhibit temporal instability. To address this, we apply a simple yet effective temporal smoothing method. Specifically, a Gaussian smoothing window is used to perform weighted averaging of 3D keypoint coordinates across neighboring frames within the window, producing the final smoothed result. Then, the smoothed 3D joint coordinates serve as the optimization target. The initial mano theta parameters and wrist transformation output by the network are then used as initial values, and we optimize them to obtain the final mano theta parameters and wrist transformation.

3. Experiment

For model training, we use only the UmeTrack[3] and HOT3D[1] datasets, with supervision provided by the MANO ground truth. The training is conducted on an NVIDIA RTX 3090 GPU, with each minibatch containing 16 image pairs. We employ the Adam optimizer, starting with an initial learning rate of 1e-4, which decays by a factor of 10 every 10 epochs. The model is trained for a total of 20 epochs.

Since the model itself lacks the ability to handle image sequences, we apply a temporal smoothing filter to ensure the temporal consistency of our results. Additionally, we observe that the 3D joint points reconstructed via triangulation exhibit a lower Mean Per-Joint Position Error (MPJPE) on the validation set compared to those derived from the model's predicted MANO parameters. This suggests that the triangulated 3D joint points can be used to further refine the model's predicted MANO parameters, which are used as the initial values in the optimization process.

In Table 1, we present our results on the test set using TTA as well as temporal smoothing and post-optimization (S&O). The table shows that our model has a solid baseline performance, with an improvement of 0.26 when TTA is applied. Additionally, incorporating temporal smoothing and post-optimization further enhances the performance by another 0.5.

ID	Base	TTA	S&O	FINGERTIP PCK AUC(↑)
1	√			70.05
2	√	√		70.31
3	√	√	√	**70.81**

Table 1. Quantitative results on Umetrack test set.

We conduct our experiments on HANDS@ECCV2024 challenge task4: multiview egocentric hand tracking challenge. The result is shown in Table 2. It can be observed that our method achieves first place in the final three metrics.

User	MPJPE(↓)	PCK AUC(↑)	FINGERTIP PCK AUC(↑)
ppjj	18.70	65.54	56.97
JVHANDS	14.21	72.23	67.63
HCB(ours)	**12.87**	**75.66**	**70.81**

Table 2. Performance comparison on HANDS@ECCV2024 challenge task4 Umetrack dataset.

4. Conclusion

In this report, we introduce a highly accurate hand pose estimation method designed for dual-view inputs. Our network model can effectively predict dual-view 2D keypoints and MANO pose parameters with high consistency without utilizing camera parameters and video sequence information, which lays a solid foundation for our subsequent triangulation reconstruction process. We test our method on the UmeTrack dataset, where it demonstrates promising results.

References

[1] Prithviraj Banerjee, Sindi Shkodrani, Pierre Moulon, Shreyas Hampali, Fan Zhang, Jade Fountain, Edward Miller, Selen Basol, Richard Newcombe, Robert Wang, et al. Introducing hot3d: An egocentric dataset for 3d hand and object tracking. *arXiv preprint arXiv:2406.09598*, 2024. 2

[2] Adnane Boukhayma, Rodrigo de Bem, and Philip HS Torr. 3d hand shape and pose from images in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10843–10852, 2019. 1

[3] Shangchen Han, Po-chen Wu, Yubo Zhang, Beibei Liu, Linguang Zhang, Zheng Wang, Weiguang Si, Peizhao Zhang, Yujun Cai, Tomas Hodan, et al. Umetrack: Unified multi-view end-to-end hand tracking for vr. In *SIGGRAPH Asia 2022 conference papers*, pages 1–9, 2022. 2

[4] Changlong Jiang, Yang Xiao, Cunlin Wu, Mingyang Zhang, Jinghong Zheng, Zhiguo Cao, and Joey Tianyi Zhou. A2j-transformer: Anchor-to-joint transformer network for 3d interacting hand pose estimation from a single rgb image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8846–8855, 2023. 2

[5] Dominik Kulon, Riza Alp Guler, Iasonas Kokkinos, Michael M Bronstein, and Stefanos Zafeiriou. Weakly-supervised mesh-convolutional hand reconstruction in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4990–5000, 2020. 1

[6] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 2

[7] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12009–12019, 2022. 2

[8] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation, 2016. 1

[9] Edoardo Remelli, Shangchen Han, Sina Honari, Pascal Fua, and Robert Wang. Lightweight multi-view 3d pose estimation through camera-disentangled representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6040–6049, 2020. 1

[10] Javier Romero, Dimitrios Tzionas, and Michael J Black. Embodied hands: Modeling and capturing hands and bodies together. *arXiv preprint arXiv:2201.02610*, 2022. 1

[11] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5693–5703, 2019. 1

[12] John Yang, Hyung Jin Chang, Seungeui Lee, and Nojun Kwak. Seqhand: Rgb-sequence-based 3d hand pose and shape estimation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*, pages 122–139. Springer, 2020. 1

[13] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 2