

# HaWoR: World-Space Hand Motion Reconstruction from Egocentric Videos

## Supplementary Material

### 1. Generalization on In-the-Wild Videos

To evaluate the generalization of HaWoR on in-the-wild video, we show qualitative results of the camera and hand reconstruction on sequences from EPIC-KITCHENS [3] in Fig. 1. Although the proposed model has not been trained on these in-the-wild data, it can still recover hands and cameras that are consistent with the input videos. We include additional in-the-wild cases in the supplementary video, where the generalization of HaWoR can be easily observed. It is worth noting that the input videos include numerous frames where the hands are outside the view frustum. Despite this, HaWoR achieves accurate reconstructions by leveraging the proposed motion priors and the in-filling network.

### 2. Comparison on In-the-Wild Videos

To compare HaWoR with other state-of-the-art methods on in-the-wild video, we show qualitative results in the camera view on sequences from EPIC-KITCHENS [3] in Fig. 2. It is evident that our method achieves significantly better results compared to HaMeR [7] and WiLoR [8] when hand truncation occurs at the boundary.

### 3. Implementation Details

In this section we provide the training and evaluation details about the network of hand motion estimation and hand motion infiller.

### 3.1. Hand Motion Estimation Network

To train the hand motion estimation network we use a combination of multiple hand video datasets for generalization of the model. In particular, we use 4 video datasets with both 3D and 2D hand annotations, constructing of 1M training frames totally:

- HOT3D [1] is an egocentric video dataset that contains daily hand activity, and we partitioned it to use 573K frames as training set.
- ARCTIC [4] dataset that contains two hands dexterously manipulating objects, focusing on hand-object interaction dynamics and 165K video frames are utilized for training.
- DexYCB [2] is a dataset focused on capturing hand grasping of objects, designed to support tasks in hand-object interaction and robotics, which provides 169K frames to train.
- HO3D [5] is a markerless dataset of color images with hands and objects involving 10 persons and 10 objects, and there are 66K frames for training.

We train the hand motion estimation network with AdamW [6] optimizer for 250K iterations with a learning rate of  $1e-5$ . The model is trained by freezing the pre-trained ViT backbone of WiLoR [8] for 2 days, using four NVIDIA A800 and a total batch size of 32. Regarding the loss weighting factor, we set  $\lambda_1 = 0.05$  for the 3D keypoint loss,  $\lambda_2 = 0.01$  for the 2D keypoint loss and  $\lambda_3 = 0.001$  for the MANO pose loss.

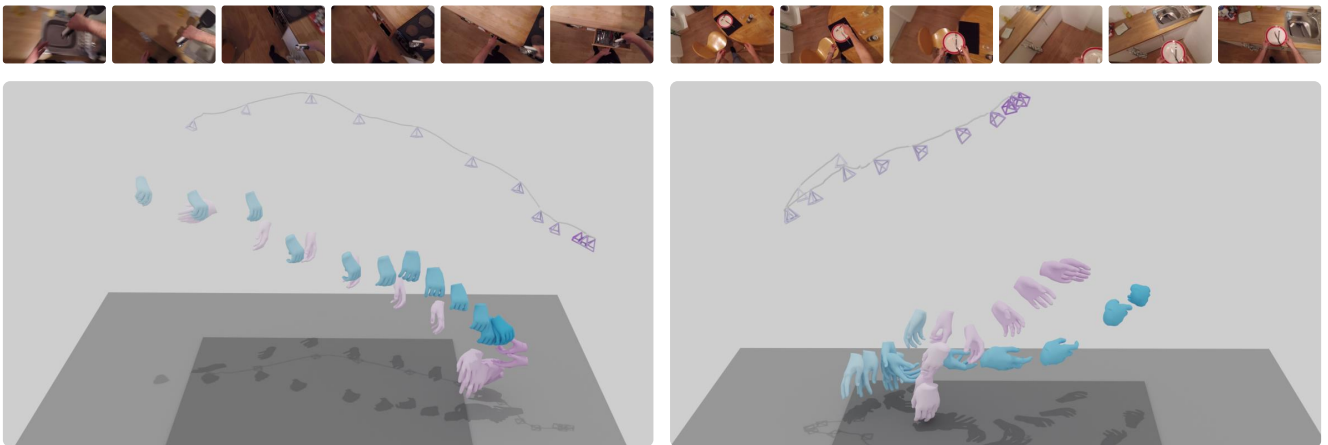


Figure 1. **Qualitative Evaluation** of the reconstructed world-space hands on in-the-wild videos from EPIC-KITCHENS [3]. Refer to the supplementary video.

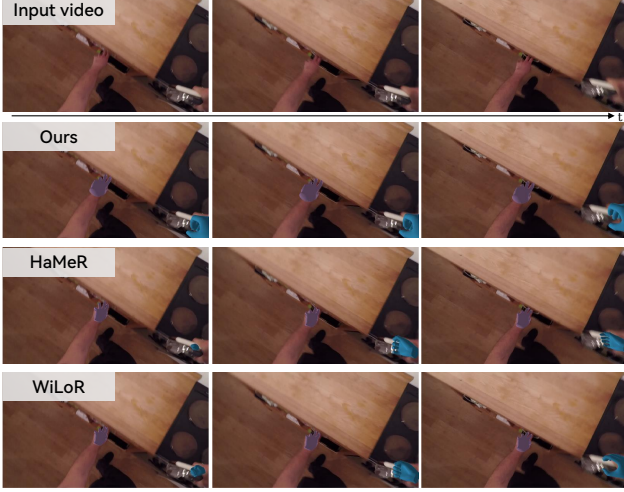


Figure 2. **Qualitative Comparison** with other state-of-the-art methods on in-the-wild videos from EPIC-KITCHENS [3]. Refer to the supplementary video.

### 3.2. Hand Motion Infiller

To facilitate the training, we transform the input sequence from camera space to canonical space, as shown in Fig. 3. Specifically, we define the first frame of each sequence as the canonical frame, and we compute the canonical transformation by aligning the first ( $0^{th}$ ) frame’s hand rotation and offset the hand translation to zero:

$$\begin{aligned} R_{c_0 2cano,i}^{c_t} &= (R_{c_0} \times \Phi_0^{c_0,i})^{-1} \times R_{c_t}, \\ t_{c_0 2cano,i}^{c_t} &= (R_{c_0} \times \Phi_0^{c_0,i})^{-1} (t_{c_t} - t_{c_0} - R_{c_0} \times \Gamma_0^{c_0,i}), \end{aligned} \quad (1)$$

where  $R_{c_t}$  denotes the rotation of  $t^{th}$  frame camera to world,  $t_{c_t}$  is the translation of  $t^{th}$  frame camera to world,  $\Phi_0^{c_0,i}$  and  $\Gamma_0^{c_0,i}$  are the hand rotation and translation in  $0^{th}$  frame camera space.

To train the hand motion infiller we use the HOT3D [1] dataset that provides 3D hand annotations of all frames, including the frames with missing hands. We first collect the non-visible hand segments to create training sequences. To increase the data scale, other sequences are sampled from the dataset and randomly masked. We keep the start and end frames as context for the infiller and randomly mask middle continuous frames. We train the hand motion infiller with AdamW [6] optimizer for 1500K iterations. The learning rate is initialized with 0.0001 and decreased by a factor of 0.9 every 100 steps. We trained the model for 1 day using one NVIDIA A800 and a batch size of 32. For weighting the losses, we set  $\gamma_1 = 0.05$  for the translation loss,  $\gamma_2 = 2.0$  for the rotation loss,  $\gamma_3 = 2.0$  for the pose loss and  $\gamma_4 = 0.05$  for the shape loss.

### 3.3. Evaluation Details

We evaluate the reconstructed world-space camera trajectories and hand motions using HOT3D dataset [1], that con-

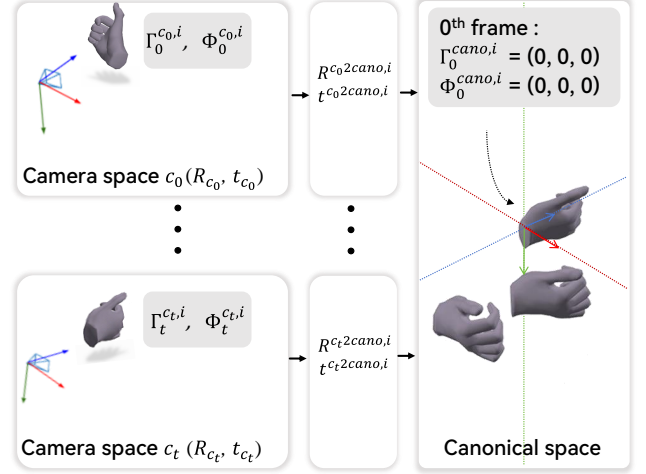


Figure 3. **Illustration of camera to canonical space transform.** We transform the sequence into canonical space that decouples the hand motion from the dynamic camera and aligns the sequence start state to zero translation and zero rotation.

tains egocentric videos from Aria glasses accompanied with moving camera trajectories and hand MANO annotations in the world-coordinates. HOT3D is also used to evaluate the infiller network since it provides accurate hand annotations, even when hands are out of the egocentric camera frustum. We use 110 videos as the training set and 27 videos as the validation set.

To evaluate HaWoR we use the following metrics:

- **PA-MPJPE** and **AUC**. To evaluate 3D hand pose in the camera-frame, we compute the Procrustes-Aligned Mean Per Joint Position Error (PA-MPJPE) measured in millimeters ( $mm$ ) and Area Under the Curve (AUC) to assess the 3D joint accuracy.
- **W-MPJPE** and **WA-MPJPE** that measure the MPJPE in  $mm$ , for a sliced sequence of 100-frame segments, after aligning the first frames and aligning the entire trajectories, respectively.
- **RTE**. We evaluate the Root Translation Error (RTE in %) normalized by the displacement of the hand trajectories after rigid alignment.
- **Accel**. We compute Acceleration error (Accel, in  $m/s^2$ ) that measures the inter-frame smoothness of the reconstructed motion.
- **FID** is the Frchet Inception Distance that calculates the distribution distance between MANO space of the estimated and GT sequence.
- **ATE** and **ATE-S**. We compute the Average Trajectory Error (ATE), which uses Procrustes analysis to align the scale to GT. ATS-S is adopted to report the Average Trajectory Error with the estimated scale.



Figure 4. Failure cases of HaWoR in hand motion reconstruction.

## 4. Limitations

One limitation of our approach is its reliance on hand-tracking outputs from an off-the-shelf method [8], which can propagate erroneous detections to HaWoR, particularly in cases of tracking identity failures. As illustrated in Fig. 4, such issues can lead to reconstruction errors, for example, when left/right hand tracking is incorrect.

It is also important to note that HaWoR models each hand independently, without any inter-penetration constraints. This can cause self-penetrations when the two hands interact, as shown in Fig. 4. In the future we plan to explore both hand interactions and further constrain the penetration between the two hands.

## 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. 1, 2
- [2] Yu-Wei Chao, Wei Yang, Yu Xiang, Pavlo Molchanov, Ankur Handa, Jonathan Tremblay, Yashraj S Narang, Karl Van Wyk, Umar Iqbal, Stan Birchfield, et al. Dexycb: A benchmark for capturing hand grasping of objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9044–9053, 2021. 1
- [3] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European conference on computer vision (ECCV)*, pages 720–736, 2018. 1, 2
- [4] 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 of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 1
- [5] Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vin-

cent Lepetit. Honnotate: A method for 3d annotation of hand and object poses. In *CVPR*, 2020. 1

- [6] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 1, 2
- [7] Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. Reconstructing hands in 3d with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9826–9836, 2024. 1
- [8] Rolandos Alexandros Potamias, Jinglei Zhang, Jiankang Deng, and Stefanos Zafeiriou. Wilor: End-to-end 3d hand localization and reconstruction in-the-wild. *arXiv preprint arXiv:2409.12259*, 2024. 1, 3