HaWoR: World-Space Hand Motion Reconstruction from Egocentric Videos

Jinglei Zhang[1], Jiankang Deng[2], Chao Ma[1*], Rolandos Alexandros Potamias[2*]
[1]Shanghai Jiao Tong University, [2]Imperial College London
{zhangjinglei168, chaoma}@sjtu.edu.cn, {j.deng16, r.potamias}@imperial.ac.uk

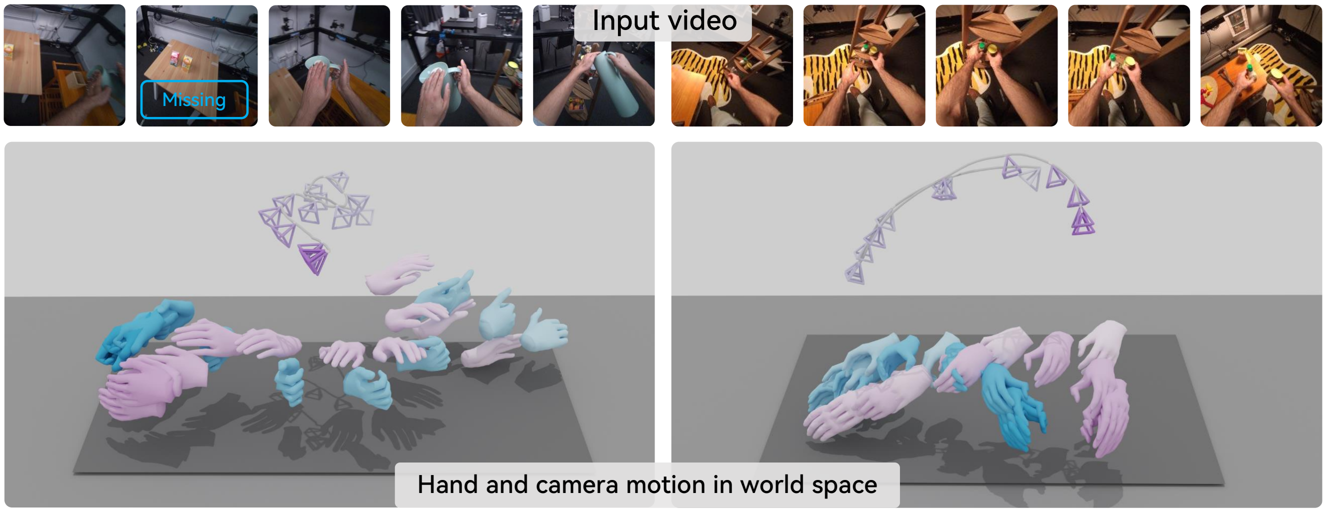

Figure 1. We propose **HaWoR**, a world-space 3D hand motion estimation method for egocentric videos. We decouple world-space hand motion estimation by combining camera-frame motions and world-space camera trajectories. HaWoR achieves state-of-the-art performance on both camera pose estimation and hand motion reconstruction, even under challenging cases where hands are out of the view frustum.

Abstract

Despite the advent in 3D hand pose estimation, current methods predominantly focus on single-image 3D hand reconstruction in the camera frame, overlooking the world-space motion of the hands. Such limitation prohibits their direct use in egocentric video settings, where hands and camera are continuously in motion. In this work, we propose HaWoR, a high-fidelity method for hand motion reconstruction in world coordinates from egocentric videos. We propose to decouple the task by reconstructing the hand motion in the camera space and estimating the camera trajectory in the world coordinate system. To achieve precise camera trajectory estimation, we propose an adaptive egocentric SLAM framework that addresses the shortcomings of traditional SLAM methods, providing robust performance under challenging camera dynamics. To ensure robust hand motion trajectories, even when the hands move out of view frustum, we devise a novel motion infiller network that effectively completes the missing frames of the sequence. Through extensive quantitative and qualitative evaluations, we demonstrate that HaWoR achieves state-of-the-art performance on both hand motion reconstruction and world-frame camera trajectory estimation under different egocentric benchmark datasets. Code and models are available on https://hawor-project.github.io/.

1. Introduction

Recovering fine-grained 3D hand motion estimation from monocular videos has garnered significant attention, given its critical role in various applications such as augmented/virtual reality (AR/VR) and human behavior analysis [3, 5, 31, 50]. Despite the progress of 3D hand pose estimation from monocular images and videos [14, 22, 23, 28–30], existing approaches predominantly focus on camera-space reconstructions, often overlooking the hands' trajectories in world-space. Neglecting the camera motion restricts the ability of hand reconstruction methods to accu-

[*] Corresponding authors

rately interpret the human movements, posing a significant burden in advancing the understanding of human actions.

Estimating hand motion on world coordinates and capturing the global motion trajectory in dynamic environments is non-trivial. This is particularly pronounced in egocentric scenarios, where both the hands and the camera are simultaneously in motion, complicating the estimation of the scale of hand movements, resulting in trajectories that fail to reflect the true motion in world coordinates. Finding a direct mapping between egocentric videos and 3D world coordinates of the hands is extremely challenging due to frequent occlusions, rapid hand movements, and the dynamic interactions between the hands and the surrounding environment [4]. In particular, in contrast to human motion recovery, reconstructing hand motion poses challenges for two reasons. Firstly, the scale of hand trajectories in egocentric views is inherently more complex compared to third-person perspectives. Secondly, in egocentric scenarios, hands frequently fall outside the field of view or experience severe occlusions, making motion estimation particularly challenging. While human motion estimation can benefit from the use of motion priors, developing such priors of the hand motion is non-trivial due to the intricate nature of hand displacements and articulations, compounded by the limited availability of large-scale hand mocap datasets.

Early approaches in world-space human mesh recovery depend on multi-view camera setups and visual odometry systems, which often struggle to generalize beyond controlled capture environments [26, 40]. Although Simultaneous Localization and Mapping (SLAM) methods [36] have made considerable strides in capturing unstructured environments with dynamic camera movements, they often struggle when dealing with dynamic scenes that involve complex human motions. To tackle this, several methods have approached world-frame reconstruction by leveraging heavy optimization schemes to align the human motion to estimated SLAM camera trajectories [43, 46]. To alleviate the costly optimization process, follow-up works have attempted to utilize camera-space motion recovery methods and directly predict the camera-to-world transform [33, 41].

Given that accurately reconstructing 3D hand motions in the world-coordinate system is significantly challenging, we propose to decompose the problem into two simpler tasks: the 3D hand motion reconstruction in the camera space and the camera trajectory estimation in the world space. For the first task, we train a high-fidelity transformer-based 3D hand motion reconstruction model to effectively capture hand motions in the camera space. However, reconstructing the 3D hand motions from egocentric videos poses significant challenges, especially when the hands are not visible within the camera frame or face severe occlusions. To address this, we enhance our proposed 3D hand motion reconstruction framework with a novel motion infilling module that estimates the missing and occluded hands. To reconstruct the camera trajectory in the world-coordinate system, we follow a hybrid method that adapts the estimated camera trajectory derived from monocular DROID-SLAM [36] to the world space using a metric foundational model [44]. Nevertheless, directly using the DROID-SLAM method and the estimated world-scale from metric networks to adjust the camera trajectory leads to faulty camera trajectories that do not accurately represent the true scale of the environment. We effectively overcome this by proposing an adaptive version of DROID-SLAM that excludes the hand regions from the bundle adjustment state and achieves accurate and robust camera trajectories from egocentric videos. Similarly, we propose a normalization of the metric space to achieve accurate world scales.

To sum up, in this paper, we present HaWoR, a robust method for 3D hand motion estimation in world coordinates from single, in-the-wild video. Specifically:

- We propose the first, to the best of our knowledge, 3D hand motion estimation method in the world-coordinate system. In contrast with previous methods that tackle 3D hand pose estimation in the camera space, we model 3D human hands in the global space, making a significant step towards real-world 3D hand motion reconstruction.
- The proposed hand motion reconstruction method leverages a novel infiller network and is able to capture high-fidelity hand motions even from videos with missing frames and severe occlusions.
- Finally, we propose a robust single-shot camera trajectory estimation pipeline tailored to egocentric videos, which achieves state-of-the-art performance compared to greedy optimization-based methods.

2. Related Work

3D Hand Pose Estimation Hand pose estimation has been widely studied for over than a decade, where early methods utilized depth cameras to reconstruct the 3D hand [13, 25, 35]. In the pioneering work of Boukhayma *et al.* [8], the authors proposed the first single-image 3D hand reconstruction method trained to estimate the hand parameters of MANO model [32]. Several methods have followed [8] by regressing MANO parameters [2, 47] or directly predicting the 3D hand vertices [20, 22, 23]. Recently, the importance of data and model scaling has been extensively highlighted, with large-scale transformer models being introduced to enhance reconstruction quality [21, 49]. In particular, Pavlakos *et al.* [28] demonstrated that by utilizing a pretrained large-scale Vision Transformer (ViT) and scaling the data can effectively improve the performance. Potamias *et al.* [30] introduced a refinement mechanism to progressively deforms the estimated hand pose resulting in state-of-the-art 3D hand pose estimations with accurate image-alignment.

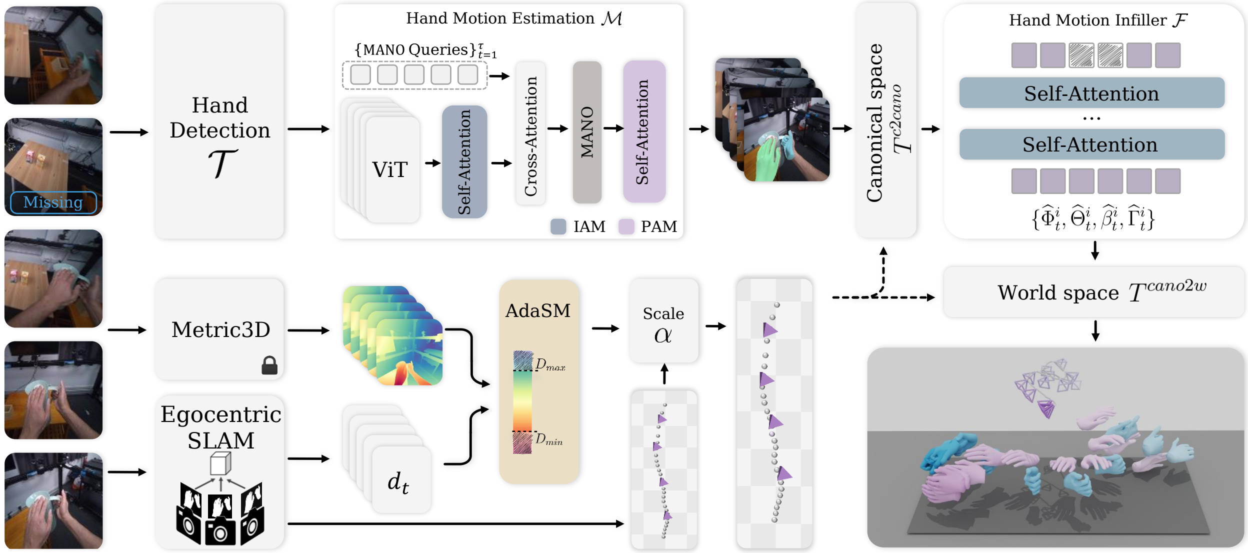
Missing
Hand Detection $\mathcal{T}$
Hand Motion Estimation $\mathcal{M}$
$\{\text{MANO Queries}\}_{t=1}^{T}$
ViT
Self-Attention
Cross-Attention
MANO
Self-Attention
IAM
PAM
Canonical space $\mathcal{T}^{c2cano}$
Hand Motion Infiller $\mathcal{F}$
Self-Attention
...
Self-Attention
$\{\widehat{\Phi}_t^i, \widehat{\Theta}_t^i, \widehat{\beta}_t^i, \widehat{\Gamma}_t^i\}$
World space $\mathcal{T}^{cano2w}$
Metric3D
Egocentric SLAM
d_t
AdaSM
D_{max}
D_{min}
Scale α

Figure 2. **Overview of our method.** Given an egocentric video $\mathbf{V}$ with a set of detected hands from an off-the-shelf detector [30], we utilize a large-scale transformer-based module with two levels of data-driven motion priors to reconstruct the 3D hand motions in the camera frame. To reconstruct hand movements beyond the view frustum, we introduce a novel hand motion infiller network designed to complete the missing frames in the hand motion sequence. We estimate world-space camera trajectories using an adaptive egocentric SLAM module that is accompanied by a foundation metric model [44] to accurately align the SLAM reconstructions to the world-coordinates.

3D Body and Camera Reconstruction. Estimation of body and camera trajectory in world coordinate system was initially approached using multi-camera setup [15] or additional wearable devices (e.g., IMU [40] or electromagnetic sensors [16]). GLAMR [46] introduced the first method for estimating global human trajectories from monocular videos with dynamic cameras, using a global trajectory regressor to infer the overall human trajectories from localized body movements. Several methods [19, 43] proposed to decouple camera and human motion by optimizing together a SLAM camera trajectory and the human motion, utilizing motion priors to constrain the optimization. Differently, WHAM [33] proposed training a regression network that given an input video and camera estimation directly predicts the global human trajectory. Recently, some studies [41, 48] combined SLAM estimations with a metric depth network to further enhance the metric scale of camera trajectories. However, these methods are primarily designed for third-person, full-body motion, which presents challenges that are markedly different from those encountered in egocentric hand motion. We propose a high-fidelity world-space 3D hand motion estimation approach to address these challenges effectively.

3. Method

Given an egocentric video $\mathbf{V} \in \mathbb{R}^{T\times H\times W\times 3}$, we aim to accurately reconstruct the complete 3D motion of hand i represented with MANO [32] pose $\{\Theta_t^i \in \mathbb{R}^{15\times 3}\}_{t=0}^{T}$ and shape parameters $\{\beta_t^i \in \mathbb{R}^{10}\}_{t=0}^{T}$ along with a global orientation $\{\Phi_t^i \in \mathbb{R}^{3}\}_{t=0}^{T}$ and root translation $\{\Gamma_t^i \in \mathbb{R}^{3}\}_{t=0}^{T}$ expressed in the world-coordinate system. The proposed method is composed of three main modules: i) the hand motion estimation network $\mathcal{M}$ that reconstructs robust hand motions in the camera-frame ii) the camera-trajectory estimation module that effectively predicts the camera pose in the world-coordinates and iii) the motion infiller network $\mathcal{F}$ that restores non-visible and occluded hands and reinforces the temporal coherence of the reconstructed 3D hand motion. An overview of the proposed framework is visualized in Fig. 2.

3.1. Hand Motion Estimation

Predicting hand motion from egocentric videos presents significant challenges due to the prevalence of severe occlusions, motion blur and perspective distortions. Despite advancements in single-image hand pose estimation [28, 30], directly extending these methods to hand motion estimation presents three key challenges that limit accurate and robust reconstructions. Firstly, such methods lack temporal coherence since they are trained on individual images, resulting in unpleasant jitter artifacts when applied to video reconstruction. Secondly, hands in egocentric videos often encounter a boundary truncation problem due to the limited field of view, leading to partial or incomplete hand visibility, which significantly deteriorates the performance of hand pose estimation frameworks. Thirdly, the lack of motion priors in hand pose estimation methods, combined with severe occlusions and motion blur in egocentric videos, can further reduce the realism of reconstructed hand motions. To effectively mitigate the aforementioned challenges, we propose a hand motion estimation network $\mathcal{M}$, which extends state-of-the-art single-image hand pose estimation methods [30] by learning spatio-temporal motion priors.

In particular, given an input video $\mathbf{V} \in \mathbb{R}^{T \times H \times W \times 3}$, we first use multi-hand detection [30] and tracking [1] methods to obtain the bounding box sequence of each hand i. We utilize the pre-trained ViT backbone of the state-of-the-art 3D hand reconstruction method WiLoR [30] to extract robust image-aligned features $\mathbf{f}_t^i$ for each frame t of the video. To mitigate truncated hands and thereby enhance the temporal consistency of the extracted image-aligned features, we introduce a temporal Image Attention Module (IAM) that updates the feature tokens $\hat{\mathbf{f}}_t^i$ with temporal information. Using temporal self-attention, appearance features are fused across adjacent frames, enhancing the robustness of the backbone features at boundary regions. Following [30], we utilize an additional token to regress MANO pose $\widetilde{\Theta}_t^i$ and shape $\widetilde{\beta}_t^i$ parameters along with the hand orientation $\widetilde{\Phi}_t^{c_t,i}$ and camera-space hand translation $\widetilde{\Gamma}_t^{c_t,i}$.

Nevertheless, despite IAM layer significantly enhancing the image features on truncated and occluded regions, the features extracted from ViT backbone still suffer from baked appearance and background elements and fail to capture expressive hand motion cues. To tackle this, we introduce an additional Pose Attention Module (PAM), which applies temporal self-attention directly to the MANO [32] pose parameters. Effectively, PAM learns hand motion priors to constrain the 3D reconstructions and improve the temporal coherence of motion.

Loss function. To train the hand motion estimation module $\mathcal{M}$, we utilize a set of loss functions, including 3D and 2D hand joint losses, along with direct MANO parameters supervision. The overall loss function is formulated as:

$$\begin{aligned}
&\mathcal{L}_{\mathcal{M}} = \sum_{t=1}^{T} (\lambda_1 \mathcal{L}_{\text{3D}}^t + \lambda_2 \mathcal{L}_{\text{2D}}^t + \lambda_3 \mathcal{L}_{\text{MANO}}^t), \\
&\mathcal{L}_{\text{3D}}^t = ||\mathbf{J}_{\text{3D}}^t - \widetilde{\mathbf{J}}_{\text{3D}}^t||_1, \\
&\mathcal{L}_{\text{2D}}^t = ||\mathbf{J}_{\text{2D}}^t - \widetilde{\mathbf{J}}_{\text{2D}}^t||_1, \\
&\mathcal{L}_{\text{MANO}}^t = ||\Theta_t - \widetilde{\Theta}_t||_2^2 + ||\beta_t - \widetilde{\beta}_t||_2^2,
\end{aligned} \tag{1}$$

where each λ_i is a weighting factor that balances the influence of the respective loss terms, $\mathbf{J}_{\text{2D}}$ and $\mathbf{J}_{\text{3D}}$ are 2D and 3D joints, respectively.

3.2. Camera Trajectory Estimation

Estimating the camera motion in the world frame from an egocentric video can be viewed as a camera localization problem. However, despite the success of SLAM methods in addressing camera localization, two major challenges prevent their direct application to egocentric hand videos: Firstly, in egocentric videos, hands occupy a substantial portion of the field of view, which can highly influence the feature-matching step of structure-from-motion methods, leading to imprecise camera motions. Secondly, SLAM methods estimate camera translation up to an arbitrary scale, which does not reflect real-world translations. To tackle the aforementioned challenges, we propose a hybrid approach that leverages an adaptive SLAM method tailored to egocentric videos coupled with a foundational metric depth model to achieve robust camera pose estimation.

Adaptive Egocentric SLAM. Despite advancements in SLAM methods, such as DROID-SLAM [36], which demonstrate robustness against subtle dynamic objects, large hand movements in egocentric views can severely impact the reconstruction accuracy of SLAM approaches. Following [41], we utilize a dual-masking strategy to exclude the hand motion from the reconstructed camera trajectory. In particular, we project the reconstructed 3D hand motions in the image space to define a hand mask $\mathbf{M}_t$. We then mask the hand regions in both the input images and the predicted confidence maps of DROID-SLAM [36].

$$\hat{I}_t = (1 - \mathbf{M}_t) \cdot I_t, \hat{w}_t = (1 - \mathbf{M}_t) \cdot w_t, \tag{2}$$

where I_t is the input image at timestamp t and w_t is the confidence map of DROID-SLAM. This step eliminates the dynamic hand regions from both the feature extraction and the dense bundle adjustment steps of DROID-SLAM, making the camera trajectory estimation robust to dynamic hands. Specifically, masking the confidence map w_t effectively excludes the corresponding coordinates from the re-projection error calculation, ensuring that only background pixels contribute to camera motion estimation in the Dense Bundle Adjustment (DBA) process and enhances robustness against hand motion.

Metric Scale Estimation. Given that monocular SLAM methods lack absolute depth information, they can only estimate the camera trajectory up to an arbitrary scale factor without a fixed world-scale reference. Hence, SLAM methods can only estimate relative depths $\mathbf{d}_t$ in arbitrary units that do not correspond to a fixed scale. To reconstruct high-fidelity camera translation scale α in real-world coordinates, we propose a robust scale estimation approach that integrates a metric network with a dynamic sampling. Specifically, we utilize Metric3D [44], a foundational model trained on large-scale datasets that can reliably predict metric-scale depth from a single image, ensuring generalization to in-the-wild data. For each keyframe of DROID-SLAM, we use Metric3D [44] to predict a scene depth $\mathbf{D}_t$ in meters. Furthermore, given that current metic networks are less accurate in regions too close and too far from the camera, we propose a dynamic sampling strategy to effectively increase the robustness of the scale estimation. Specifically, we mask out both the hand regions as well as points that are either near or far away from the camera and restrict the estimation of the scaling factor to reliable points within an intermediate range and outside the hand region. The optimal min-max depth interval priors are derived by optimizing the scale accuracy on the egocentric training dataset. Given the obtained hand masks and the thresholds for distance, the adaptive sampling module (AdaSM) selects a point set S_t that satisfies:

$$S_t = \{p \mid p \notin \mathbf{M}_t, \mathbf{D}_{\text{min}} < \mathbf{D}_t(p) < \mathbf{D}_{\text{max}}\}. \tag{3}$$

Following [41], we estimate the final scale α by optimizing the alignment between SLAM and Metric3D depth estimations on the sampled set as:

$$E(\alpha) = \sum_{p \in S_t} \mathcal{L}_{\text{GM}}(\mathbf{D}_t(p) - \alpha \cdot \mathbf{d}_t(p)), \tag{4}$$

where $\mathcal{L}_{\text{GM}}$ is the German-McClure loss function [6]. This approach optimizes the scale α estimation by focusing on regions where depth prediction is more reliable, thereby mitigating the influence of outliers such as moving hands or extreme depth values, achieving highly precise and robust scale estimation.

3.3. Hand Motion Infiller

Due to the limited field of view in egocentric videos, hands are often outside of the visible frame, leading to distorted and incomplete 3D reconstruction of the hand motion. To address this issue, we introduce a novel hand motion infiller network $\mathcal{F}$, which is tailored to complete the out-of-bounds hands and reconstruct the full 3D hand motion sequence. In particular, given an incomplete T-frame sequence of hand i MANO parameters of $\{\widetilde{\Theta}_t^i, \widetilde{\beta}_t^i, \widetilde{\Phi}_t^{c_t,i}, \widetilde{\Gamma}_t^{c_t,i}\}$ predicted from the hand motion estimation network in each camera frame c_t (with missing frames set to zero), the motion infiller network predicts a complete motion $\{\widehat{\Theta}_t^i, \widehat{\beta}_t^i, \widehat{\Phi}_t^i, \widehat{\Gamma}_t^i\}$ that accurately fills the missing frames.

Canonical space transformation. As a first step, we transform the input sequence from camera space to canonical space, which decouples the hand motion from the dynamic camera and aligns the sequence start state to zero translation and zero rotation. This operation can standardize the input sequence and facilitate training. Specifically, we first compute the camera-to-canonical transformation $T^{c_i 2cano,i} = [R^{c_i 2cano,i} | t^{c_i 2cano,i}]$ that aligns the first frame's hand rotation and translation to zero. Subsequently, the hand rotations and translations are transformed into canonical space:

$$\begin{aligned} \Phi_t^{cano,i} &= R^{c_t 2cano,i} \times \Phi_t^{c_t,i}, \\ \Gamma_t^{cano,i} &= R^{c_t 2cano,i} \times \Gamma_t^{c_t,i} + t^{c_t 2cano,i}. \end{aligned} \tag{5}$$

Infiller Network. Predicting the hand pose of missing frames can be considered a motion-in-between task. To this end, we follow [17] and build our motion infiller network using a transformer-encoder architecture trained to predict the missing pose tokens. Specifically, we initially project the input MANO sequences to D-dimension latent vectors and then feed them to a set of stacked multi-head self-attention layers. Given that transformer encoder does not explicitly capture the auto-regressive nature of the motion, we incorporate positional embeddings [39] to encode the temporal information of each frame. The output tokens are passed to a simple fully-connected decoder that regresses the MANO sequence in canonical space. Finally, we convert the MANO sequence to the world space by computing the canonical-to-world transformation $T^{cano2w,i} = [R^{cano2w,i} | t^{cano2w,i}]$.

Training. To train the motion infiller network, we use HOT3D [4] since it provides both egocentric and third-person views of the hands, enabling us to easily identify and label the frames of each video where the hands are out of the egocentric camera frustum. To augment the training data, we sample additional video sequences and randomly mask frame segments while retaining the start and the end frames to serve as context for the infiller network. To facilitate the training process, we initialize the MANO parameters of the missing frames using a pose interpolation scheme. Specifically, for a given motion sequence, translations and shape parameters are linearly interpolated, while global rotations and pose parameters are interpolated with spherical linear interpolation (SLERP). This can reduce the workload of the infiller network and enable more robust reconstructions.

Loss Functions. We train the motion infiller network using a combination of loss functions to penalize the world translation and orientation along with the hand pose and shape. The overall loss function is formulated as:

$$\begin{aligned} \mathcal{L}_{\mathcal{F}} &= \sum_{t=1}^{T} (\gamma_1 \mathcal{L}_\Gamma^t + \gamma_2 \mathcal{L}_\Phi^t + \gamma_3 \mathcal{L}_\Theta^t + \gamma_4 \mathcal{L}_\beta^t), \\ \mathcal{L}_\Gamma^t &= ||\Gamma_t - \hat{\Gamma}_t||_1, \mathcal{L}_\Phi^t = ||\Phi_t - \hat{\Phi}_t||_1, \\ \mathcal{L}_\Theta^t &= ||\Theta_t - \hat{\Theta}_t||_1, \mathcal{L}_\beta^t = ||\beta_t - \hat{\beta}_t||_1, \end{aligned} \tag{6}$$

where each γ_i is a weighting factor that balances the influence of the respective loss terms.

4. Experiments

Datasets. To assess the camera-frame hand motion reconstruction performance of HaWoR and baseline models we utilize DexYCB [9] dataset, which comprises videos capturing hand-object interactions from a set of static cameras including conditions of severe occlusion. To evaluate the reconstructed world-space camera and hand trajectories along with the infiller reconstructions, we use

HOT3D dataset [4], that contains egocentric videos captured from dynamic cameras accompanied with ground-truth camera trajectories along with MANO annotations in world-coordinates.

Evaluation Metrics. To evaluate the 3D hand pose in the camera-frame, we use Procrustes-Aligned Mean Per Joint Position Error (PA-MPJPE) and the Area Under the Curve (AUC) of correctly localized keypoints. Following [43], we assess the hand estimation in the world-frame using World MPJPE (W-MPJPE) and World Aligned MPJPE (WA-MPJPE). In addition, we evaluate the error of the entire trajectory with root translation error (RTE) and compute acceleration error (Accel) to evaluate the smoothness of motion. Frechet Inception Distance (FID) is used to measure the motion filling quality. To quantify the quality of the camera trajectory, we compute the Average Trajectory Error (ATE) that aligns the scale of GT and ATS-S that uses the estimated scale, as described in [41].

4.1. Camera-frame 3D Hand Motion

To achieve accurate world-space hand motion reconstruction it is essential to achieve robust and high fidelity hand motion estimation in the camera-frame. Given that egocentric videos often face sever occlusions, we follow [12] and utilize DexYCB dataset that provides explicit annotations regarding the occluded frames within a video. Specifically, in Tab. 1 we compare HaWoR against image- and video-based methods for camera-frame 3D hand motion reconstruction under different occlusion ratio levels. As can be observed, HaWoR archives robust performance across different occlusion rates. In contrast, WiLoR [30] and Deformer [12] that serve as state-of-the-art methods for 3D hand pose estimation from single-image and video, respectively, face a huge performance degradation on videos with increased occlusion rates. It is important to note that HaWoR performance on sever occlusion rate (75%-100%) shows a more significant improvement than state-of-the-art methods (*i.e.*, WiLoR [30] 5.68 vs HaWoR 5.07).

4.2. World-frame 3D Hand Motion

In this section, we quantitatively and qualitatively evaluate HaWoR in hand motion reconstruction in the world-space.

Baselines. Given that HaWoR is currently the first, to the best of our knowledge, framework that tackles world-space hand motion reconstruction, we implement a set of strong baseline methods that follow the literature of world-grounded human body motion estimation [33, 43, 46]. In particular, we use state-of-the-art performing methods for hand pose estimation, namely HaMeR [28], WiLoR [30] and HandDGP [38], coupled with DROID-SLAM [36], to recover the world-space hand and camera motion. Additionally, we implement an optimization-based method that closely follows SLAHMR [43] by combining DROID-SLAM with the powerful hand motion prior (HMP) [11] to align the hand pose estimations with the camera trajectories. Additionally, we compare the proposed world-frame camera trajectory with different metric depth estimation methods, including ZoeDepth [7], DepthAnythingV2 [42] and Metric3DV2 [44]. To facilitate understanding between the contribution of each network component, we divide the evaluation into two steps to assess both the global camera trajectory and the reconstructed hand motion in world-coordinates.

	Methods	All		50%-75%		75%-100%	
		MPJPE	AUC	MPJPE	AUC	MPJPE	AUC
image-based	Spurr *et al.* [34]	6.83	86.4	8.00	84.0	10.65	78.8
	MeshGraphormer [23]	6.41	87.2	7.22	85.6	7.76	84.5
	SemiHandObj [24]	6.33	87.4	7.17	85.7	8.96	82.1
	HandOccNet [27]	5.80	88.4	6.43	87.2	7.37	85.3
	WiLoR [30]	5.01	90.0	5.42	89.2	5.68	88.7
temporal	S^2HAND(V) [37]	7.27	85.5	7.71	84.6	7.87	84.3
	VIBE [18]	6.43	87.1	6.84	86.4	7.06	85.8
	TCMR [10]	6.28	87.5	6.58	86.8	6.95	86.1
	Deformer [12]	5.22	89.6	5.70	88.6	6.34	87.3
	Proposed	**4.76**	**90.5**	**5.03**	**89.9**	**5.07**	**89.9**

Table 1. Quantitative camera-frame comparison of state-of-the-art hand pose estimation methods on the **DexYCB** test dataset. We compare PA-MPJPE and AUC results, especially the split under large occlusion proportion (50%-75% and 75%-100%), which highlights our robustness in challenging visibility conditions.

Methods	ATE↓	ATE-S↓			
	All	Short	Med	Long	All
DROID [36]	3.80	-	-	-	-
DROID + ZoeDepth [7]	3.80	25.03	39.39	75.95	43.58
DROID + DepthAnyV2 [42]	3.80	18.14	25.50	43.60	27.49
DROID + Metric3DV2 [44]	3.80	14.28	21.56	29.10	21.07
Proposed w/o Scale	3.36	-	-	-	-
Proposed w. ZoeDepth [7]	3.36	11.91	25.34	36.05	23.67
Proposed w. DepthAnyV2 [42]	3.36	14.63	20.49	25.54	19.85
Proposed w/o AdaSM	3.36	14.03	22.38	27.49	20.97
Proposed	**3.36**	**9.31**	**15.86**	**19.26**	**14.61**

Table 2. Evaluation of camera estimation with aligned scale (**ATE**) and estimated scale (**ATE-S**). We also report the split results of short ($< 5m$), medium ($3m - 5m$) and long ($> 5m$) displacement. ATE and ATE-S is in mm.

Global Camera Trajectory. In Tab. 2 we evaluate the camera trajectory estimation of the proposed adaptive egocentric SLAM method compared to baseline approaches that naively combine DROID-SLAM [36] with metric networks.

As can be easily observed, although directly apply-

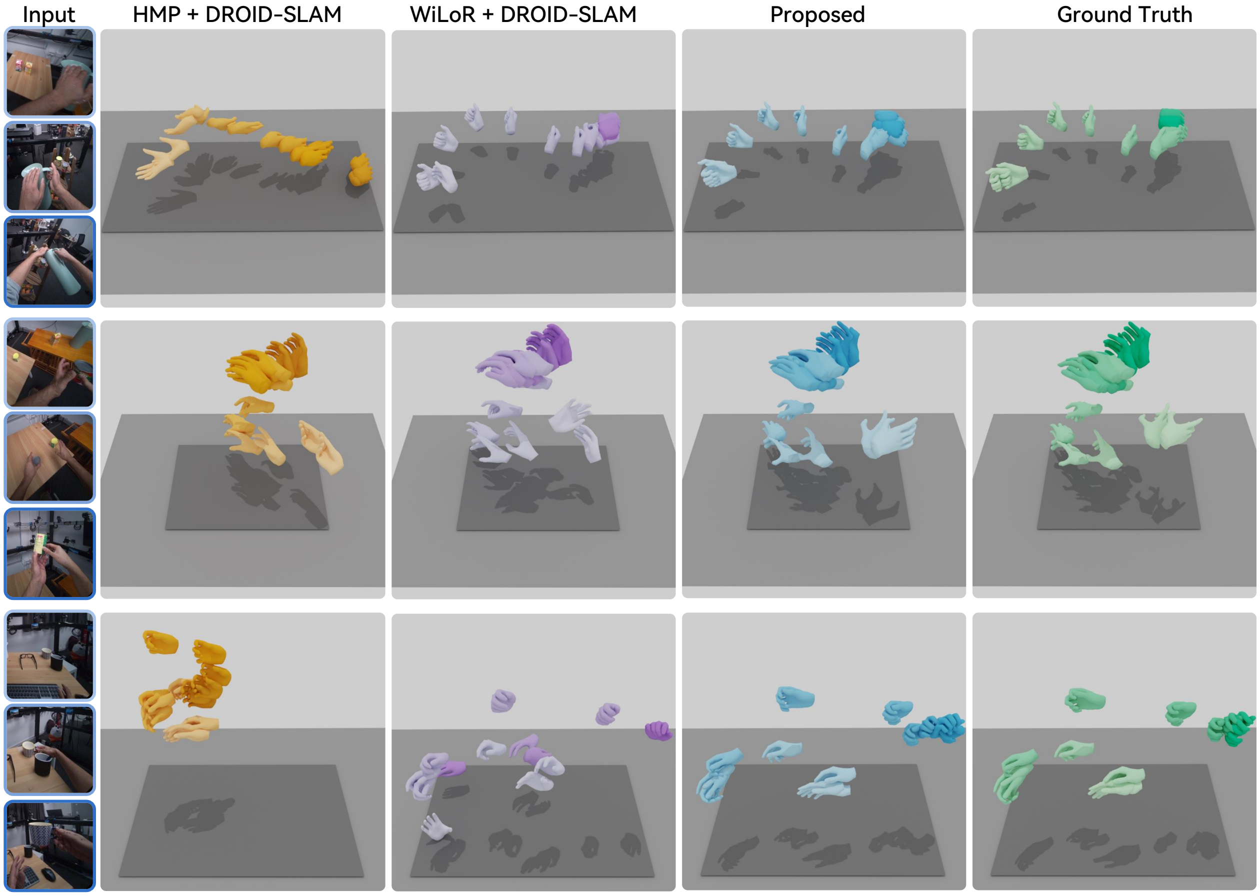

Figure 3. **Visualization** of right-hand estimated trajectories on challenging cases of **HOT3D**. The first example depicts *someone picking up a kettle, turning around, and pouring water*. The second example depicts *the subject placing a tin on the table and then picking up another*. The third video depicts *the subject using a mouse keyboard and then reaching for a cup to drink water*. In contrast to the baseline methods, HaWoR achieves robust hand trajectories, especially in challenging scenarios with large hand movements and truncated hands.

Method	PA-MPJPE	W-MPJPE	WA-MPJPE	RTE	Accel
HaMeR-SLAM [28]	9.39	156.03	43.37	4.77	19.25
HandDGP-SLAM [38]	17.88	154.30	42.93	3.18	20.17
WiLoR-SLAM [30]	6.00	151.67	39.49	2.99	8.02
HMP-SLAM [11]	10.51	119.41	39.46	2.79	5.50
Proposed	**4.79**	**33.20**	**11.27**	**0.78**	**5.41**

Table 3. Quantitative evaluation in world-space coordinates on **HOT3D** dataset.

ing metric models to SLAM-estimated camera trajectories may be sufficient for third-person body motion reconstruction [33, 43, 45, 46], it falls short in accurately reconstructing camera trajectories in egocentric scenarios, which are characterized by significant occlusions and hands occupying a large portion of the frame. In contrast, the proposed approach effectively mitigates this issue by providing accurate visual cues during the SLAM bundle adjustment step that facilitate the reconstruction performance in egocentric scenarios. It is also important to note that the trajectory error is further exacerbated when it comes to estimating the actual world-scale estimations (ATE-S). The effect of the adaptive SLAM proposed in HaWoR can be further validated in Fig. 4, where the camera trajectories match the ground truth camera motion, addressing the limitations of DROID-SLAM approach in egocentric views.

Hand motion estimation in world-coordinates. In Tab. 3 we report the performance of HaWoR and the baseline methods in hand motion reconstruction in the world-coordinates. HaWoR significantly outperforms both regression and optimization-based baselines by a large margin under both camera (PA-MPJPE) and world-space reconstructions (W-MPJPE). Furthermore, HaWoR produces more stable and robust motion reconstructions that are unaffected by occlusions, as indicated by the RTE metric. Besides, compared to the baseline methods, HaWoR achieves significantly lower acceleration error, validating the smoothness

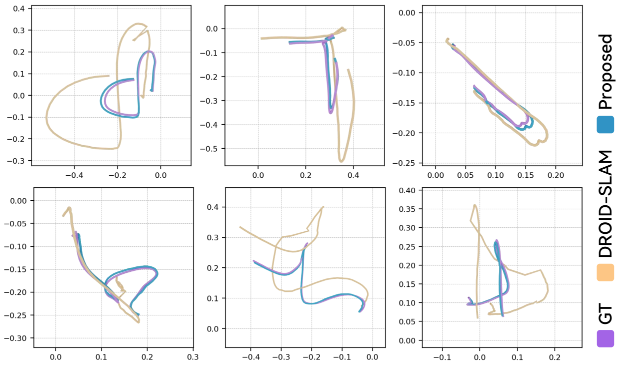

Figure 4. **Camera global trajectory**. The proposed adaptive SLAM approach demonstrates precise camera trajectory estimation while recovering accurate real-world scale, outperforming DROID-SLAM, which struggles with both trajectory accuracy and scale consistency.

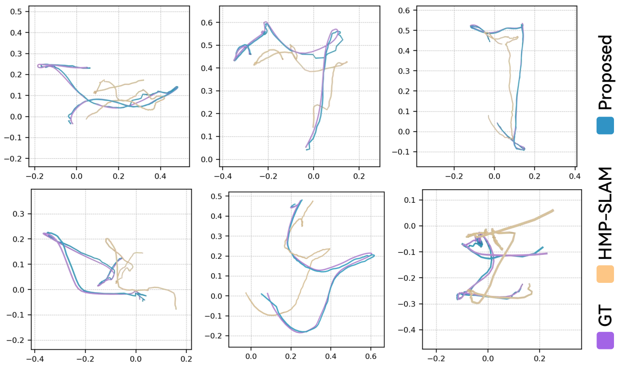

Figure 5. **Hand global trajectory** for the right hand on HOT3D. Compared to HMP-SLAM, HaWoR produces accurate trajectories even for complex and long-range hand movements.

of the reconstructed motions across frames.

The accuracy of the proposed hand motion can be further validated in both Fig. 3 and Fig. 5, where we compare the hand trajectories estimated from HaWoR and an optimization-based approach that utilizes hand motion priors to guide the motion (*HMP-SLAM*). HaWoR achieves accurate hand trajectories that follow the ground truth even on complex motions that the baseline methods fail. It is also important to note that apart from the superior performance in camera trajectory estimation, HaWoR requires only a single forward pass of 40 ms per frame, significantly reducing inference runtime by 75% compared to optimization-based method of HMP-SLAM, that requires 160 ms for per frame.

4.3. Ablation

We perform an ablation study to assess the effect of key components in our framework. In particular, we initially report the effect of the image and pose motion priors that compose the proposed hand motion estimation network. As can be seen in Tab. 4(a), both IAM and PAM modules contribute to the performance of HaWoR, improving the robustness of the reconstructions. Furthermore, we evaluate the contribution of the motion infiller network and its generalization performance on HOT3D [4] datasets. LERP is using frame linear interpolation, where root translation and shape are linearly interpolated, and joint rotations are spherically linear interpolated. We also compare with replicating the last visible pose. As can be observed from Tab. 4(b), the proposed motion infiller network can significantly outperform naive motion completion methods.

Method	PA-MPJPE	W-MPJPE	WA-MPJPE	RTE	Accel
w/o Pretrained ViT	7.59	86.80	19.46	1.26	9.09
w/o IAM & PAM	5.07	44.60	13.85	0.93	8.42
w/o PAM	4.80	36.32	12.40	0.88	6.03
Proposed	**4.79**	**33.20**	**11.27**	**0.78**	**5.41**

(a) **Hand motion components**. Here is the ablation results without IAM (Image Attention Module), PAM (Pose Attention Module) or pretrained ViT.

Method	FID	PA-MPJPE	W-MPJPE	WA-MPJPE	RTE
Last Pose	1.52	7.83	116.79	78.78	13.04
LERP	1.42	6.33	75.01	49.16	9.39
Proposed	**0.57**	**6.22**	**66.25**	**37.22**	**7.41**

(b) **Motion Infiller**. We experiment on the invisible sequences of HOT3D [4] validation dataset.

Table 4. Ablations study on the key modules of our method.

5. Conclusion and Limitations

In this work we present HaWoR, a high-fidelity 3D hand motion reconstruction method in the world-space. HaWoR is founded on a powerful camera-frame transformer-based hand motion reconstruction module and a robust infiller network to estimate and fill the motion-in-between missing frames. To align the camera-frame hand motions in the world-coordinate system we propose an adaptive egocentric SLAM module that facilitates global camera trajectory estimation under challenging and occluded egocentric views. Through extensive experimental results we demonstrate that HaWoR outperforms previous methods and achieves state-of-the-art performance under different benchmark datasets. However, while HaWoR significantly accelerates hand motion reconstruction compared to previous approaches, the runtime performance is still far from real-time. In the future, we could explore foundational models to directly estimate world-space camera trajectories to make a step towards real-time world-frame hand motion estimation.

Acknowledgements This work was supported in part by the National Natural Science Foundation of China (62322113, 62376156), and Shanghai Municipal Science and Technology Major Project (2021SHZDZX0102).

References

[1] Nir Aharon, Roy Orfaig, and Ben-Zion Bobrovsky. Bot-sort: Robust associations multi-pedestrian tracking. *arXiv preprint arXiv:2206.14651*, 2022. 4

[2] Seungryul Baek, Kwang In Kim, and Tae-Kyun Kim. Pushing the envelope for rgb-based dense 3d hand pose estimation via neural rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1067–1076, 2019. 2

[3] Vasileios Baltatzis, Rolandos Alexandros Potamias, Evangelos Ververas, Guanxiong Sun, Jiankang Deng, and Stefanos Zafeiriou. Neural sign actors: A diffusion model for 3d sign language production from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1985–1995, 2024. 1

[4] 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, 5, 6, 8

[5] Siddhant Bansal, Michael Wray, and Dima Damen. Hoi-ref: Hand-object interaction referral in egocentric vision. *arXiv preprint arXiv:2404.09933*, 2024. 1

[6] Jonathan T Barron. A general and adaptive robust loss function. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4331–4339, 2019. 5

[7] Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*, 2023. 6

[8] 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. 2

[9] 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. 5

[10] Hongsuk Choi, Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. Beyond static features for temporally consistent 3d human pose and shape from a video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1964–1973, 2021. 6

[11] Enes Duran, Muhammed Kocabas, Vasileios Choutas, Zicong Fan, and Michael J Black. Hmp: Hand motion priors for pose and shape estimation from video. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6353–6363, 2024. 6, 7

[12] Qichen Fu, Xingyu Liu, Ran Xu, Juan Carlos Niebles, and Kris M Kitani. Deformer: Dynamic fusion transformer for robust hand pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23600–23611, 2023. 6

[13] Liuhao Ge, Hui Liang, Junsong Yuan, and Daniel Thalmann. Robust 3d hand pose estimation in single depth images: from single-view cnn to multi-view cnns. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3593–3601, 2016. 2

[14] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11090–11100, 2022. 1

[15] Hanbyul Joo, Hao Liu, Lei Tan, Lin Gui, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social motion capture. In *Proceedings of the IEEE international conference on computer vision*, pages 3334–3342, 2015. 3

[16] Manuel Kaufmann, Jie Song, Chen Guo, Kaiyue Shen, Tianjian Jiang, Chengcheng Tang, Juan José Zárate, and Otmar Hilliges. Emdb: The electromagnetic database of global 3d human pose and shape in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14632–14643, 2023. 3

[17] Jihoon Kim, Taehyun Byun, Seungyoun Shin, Jungdam Won, and Sungjoon Choi. Conditional motion inbetweening. *Pattern Recognition*, 132:108894, 2022. 5

[18] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. Vibe: Video inference for human body pose and shape estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5253–5263, 2020. 6

[19] Muhammed Kocabas, Ye Yuan, Pavlo Molchanov, Yunrong Guo, Michael J Black, Otmar Hilliges, Jan Kautz, and Umar Iqbal. Pace: Human and camera motion estimation from in-the-wild videos. In *2024 International Conference on 3D Vision (3DV)*, pages 397–408. IEEE, 2024. 3

[20] Dominik Kulon, Haoyang Wang, Riza Alp Güler, Michael M. Bronstein, and Stefanos Zafeiriou. Single image 3d hand reconstruction with mesh convolutions. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2019. 2

[21] Han Li, Bowen Shi, Wenrui Dai, Hongwei Zheng, Botao Wang, Yu Sun, Min Guo, Chenglin Li, Junni Zou, and Hongkai Xiong. Pose-oriented transformer with uncertainty-guided refinement for 2d-to-3d human pose estimation. In *Proceedings of the AAAI conference on artificial intelligence*, pages 1296–1304, 2023. 2

[22] Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-end human pose and mesh reconstruction with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1954–1963, 2021. 1, 2

[23] Kevin Lin, Lijuan Wang, and Zicheng Liu. Mesh graphormer. In *Proceedings of the IEEE/CVF international*

conference on computer vision, pages 12939–12948, 2021. 1, 2, 6

[24] Shaowei Liu, Hanwen Jiang, Jiarui Xu, Sifei Liu, and Xiaolong Wang. Semi-supervised 3d hand-object poses estimation with interactions in time. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 6

[25] Iason Oikonomidis, Nikolaos Kyriazis, Antonis A Argyros, et al. Efficient model-based 3d tracking of hand articulations using kinect. In *BmVC*, page 3, 2011. 2

[26] Hyun Soo Park, Takaaki Shiratori, Iain Matthews, and Yaser Sheikh. 3d trajectory reconstruction under perspective projection. *International Journal of Computer Vision*, 115:115–135, 2015. 2

[27] JoonKyu Park, Yeonguk Oh, Gyeongsik Moon, Hongsuk Choi, and Kyoung Mu Lee. Handoccnet: Occlusion-robust 3d hand mesh estimation network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1496–1505, 2022. 6

[28] 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, 2, 3, 6, 7

[29] Rolandos Alexandros Potamias, Stylianos Ploumpis, Stylianos Moschoglou, Vasileios Triantafyllou, and Stefanos Zafeiriou. Handy: Towards a high fidelity 3d hand shape and appearance model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4670–4680, 2023.

[30] 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, 2, 3, 4, 6, 7

[31] Jing Qi, Li Ma, Zhenchao Cui, and Yushu Yu. Computer vision-based hand gesture recognition for human-robot interaction: a review. *Complex & Intelligent Systems*, 10(1): 1581–1606, 2024. 1

[32] Javier Romero, Dimitrios Tzionas, and Michael J Black. Embodied hands: modeling and capturing hands and bodies together. *ACM Transactions on Graphics (TOG)*, 36(6):1–17, 2017. 2, 3, 4

[33] Soyong Shin, Juyong Kim, Eni Halilaj, and Michael J Black. Wham: Reconstructing world-grounded humans with accurate 3d motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2070–2080, 2024. 2, 3, 6, 7

[34] Adrian Spurr, Umar Iqbal, Pavlo Molchanov, Otmar Hilliges, and Jan Kautz. Weakly supervised 3d hand pose estimation via biomechanical constraints. In *European conference on computer vision*, pages 211–228. Springer, 2020. 6

[35] Andrea Tagliasacchi, Matthias Schröder, Anastasia Tkach, Sofien Bouaziz, Mario Botsch, and Mark Pauly. Robust articulated-icp for real-time hand tracking. In *Computer graphics forum*, pages 101–114. Wiley Online Library, 2015. 2

[36] Zachary Teed and Jia Deng. Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras. *Advances in neural information processing systems*, 34:16558–16569, 2021. 2, 4, 6

[37] Zhigang Tu, Zhisheng Huang, Yujin Chen, Di Kang, Linchao Bao, Bisheng Yang, and Junsong Yuan. Consistent 3d hand reconstruction in video via self-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8):9469–9485, 2023. 6

[38] Eugene Valassakis and Guillermo Garcia-Hernando. Handdgp: Camera-space hand mesh prediction with differentiable global positioning. In *European Conference on Computer Vision*, pages 479–496. Springer, 2025. 6, 7

[39] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 5

[40] Timo Von Marcard, Roberto Henschel, Michael J Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *Proceedings of the European conference on computer vision (ECCV)*, pages 601–617, 2018. 2, 3

[41] Yufu Wang, Ziyun Wang, Lingjie Liu, and Kostas Daniilidis. Tram: Global trajectory and motion of 3d humans from in-the-wild videos. In *European Conference on Computer Vision*, pages 467–487. Springer, 2025. 2, 3, 4, 5, 6

[42] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10371–10381, 2024. 6

[43] Vickie Ye, Georgios Pavlakos, Jitendra Malik, and Angjoo Kanazawa. Decoupling human and camera motion from videos in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 21222–21232, 2023. 2, 3, 6, 7

[44] Wei Yin, Chi Zhang, Hao Chen, Zhipeng Cai, Gang Yu, Kaixuan Wang, Xiaozhi Chen, and Chunhua Shen. Metric3d: Towards zero-shot metric 3d prediction from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9043–9053, 2023. 2, 3, 4, 6

[45] Wanqi Yin, Zhongang Cai, Ruisi Wang, Fanzhou Wang, Chen Wei, Haiyi Mei, Weiye Xiao, Zhitao Yang, Qingping Sun, Atsushi Yamashita, et al. Whac: World-grounded humans and cameras. In *European Conference on Computer Vision*, pages 20–37. Springer, 2025. 7

[46] Ye Yuan, Umar Iqbal, Pavlo Molchanov, Kris Kitani, and Jan Kautz. Glamr: Global occlusion-aware human mesh recovery with dynamic cameras. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11038–11049, 2022. 2, 3, 6, 7

[47] Xiong Zhang, Qiang Li, Hong Mo, Wenbo Zhang, and Wen Zheng. End-to-end hand mesh recovery from a monocular rgb image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2354–2364, 2019. 2

[48] Yizhou Zhao, Tuanfeng Yang Wang, Bhiksha Raj, Min Xu, Jimei Yang, and Chun-Hao Paul Huang. Synergistic global-

space camera and human reconstruction from videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1216–1226, 2024. 3

[49] Hongwei Zheng, Han Li, Wenrui Dai, Ziyang Zheng, Chenglin Li, Junni Zou, and Hongkai Xiong. Hipart: Hierarchical pose autoregressive transformer for occluded 3d human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. 2

[50] Ronglai Zuo, Rolandos Alexandros Potamias, Evangelos Ververas, Jiankang Deng, and Stefanos Zafeiriou. Signs as tokens: An autoregressive multilingual sign language generator. *arXiv preprint arXiv:2411.17799*, 2024. 1