STREAMHAR: STREAMING WORLD-GROUNDED HAND RECONSTRUCTION FROM EGOCENTRIC VIDEOS

Anonymous authors
Paper under double-blind review

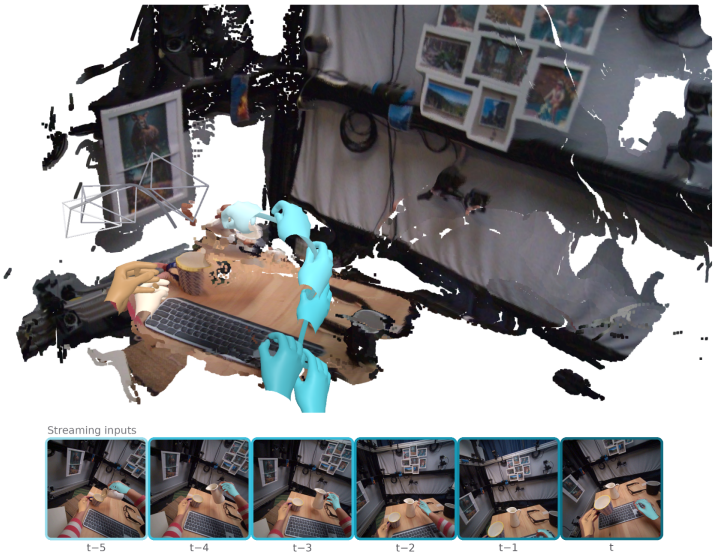


Figure 1: **Online world-space hand reconstruction.** StreamHaR estimates online both hands and the head camera in a metric world frame (top) from an egocentric video stream. At the bottom, the estimated hands projected into the video stream are shown.

ABSTRACT

Augmented reality, teleoperation and robot learning from demonstration require the wearer's hands reconstructed in a metric world frame as the egocentric video arrives. Existing world-space methods are offline, relying on future frames and global optimization, limiting their interactive applicability. To address this, we present StreamHaR, a fully online method that emits a metric, world-space hand at every frame with a state of constant size, built to be causal, faithful, general and efficient. StreamHaR splits the world-space hand into three problems with three estimators: the wrist is measured by a per-frame Perspective-n-Point solve on projection-consistent keypoints, the articulation is recognized by an encoder, and the camera comes from a causal odometer that anchors its metric scale from monocular depth, so the trajectory is metric the moment it is emitted, with no bundle adjustment and no calibration. With zero look-ahead and its own detector, StreamHaR reaches the world-space accuracy of the offline state of the art at a lower articulation error and transfers zero-shot to unseen headsets.

1 INTRODUCTION

People act on the world primarily through their hands, and egocentric video from head-worn cameras captures their actions from the wearer's point of view. Reconstructing the wearer's hands in 3D from such video drives hand tracking for Virtual Reality (VR) and Augmented Reality (AR) (Han et al., 2022; Banerjee et al., 2025) and sign-language avatars (Yu et al., 2024; Zuo et al., 2025). The hand motion is required for teleoperation (Cheng et al., 2024) and robot learning from human demonstration (Hoque et al., 2026; Zhang et al., 2026; Li et al., 2025), where the hand is retarget to a robot hand.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

With the parametric MANO model (Romero et al., 2017), there has been rapid progress on monocular hand mesh recovery, driven by large annotated datasets (Zimmermann et al., 2019; Moon et al., 2020; Jin et al., 2020; Chao et al., 2021) and approaches that recover an accurate mesh from a single crop (Lin et al., 2021a;b; Pavlakos et al., 2024; Potamias et al., 2025), over short clips (Lyu et al., 2026) and for two interacting hands (Li et al., 2022; Yu et al., 2023; Lee et al., 2023; Ren et al., 2023; Moon, 2023). These methods estimate the hand relative to the camera, not in the world: they recover *what the hand does*, not *where it is*. For a head-worn camera, this difference matters, because the camera moves with the head and the observed motion combines hand and camera movement (Yu et al., 2025). Following the same shift in world-grounded human mesh recovery (Yuan et al., 2022; Ye et al., 2023; Kocabas et al., 2024; Shin et al., 2024; Wang et al., 2024), recent work reconstructs the hand in the world space (Ye et al., 2025; Yu et al., 2025; Zhang et al., 2025; Ye et al., 2026), where the hand is expressed in a fixed, metric frame; hence, its trajectory reflects its true motion and stays consistent over time.

The world-space hand reconstruction methods (Zhang et al., 2025; Yu et al., 2025) available today are offline. They read every frame before they emit the first pose, revise earlier frames based on the later ones, and answer seconds to minutes after the clip has ended (Table 1). The applications provided above, however, consume the hand as the video arrives: an AR system attaches content to the hand *now*, a teleoperated hand follows the operator within a few frames, and a robot in a feedback loop cannot wait for the recording to end. These *streaming* applications require online reconstruction, where frames arrive one at a time, and each estimate depends only on the current frame and the ones before it. Online processing enables the low latency these applications need, at a compute and memory cost that does not grow with the video (Zhao et al., 2026). For full bodies, recent work has formulated the requirements of online world-grounded reconstruction (Zhao et al., 2026; Chen et al., 2026), and the same applies to hands. First, the whole pipeline must be *causal*, including the camera estimate. Second, it must be *faithful*, placing and posing the hand as accurately as an offline method. Third, it must be *efficient*, updating once per frame with constant memory. Egocentric capture adds a fourth requirement: the method must be *general*, because headsets, optics and frame rates differ and no dataset covers them all. So far, no method meets all four for hands.

Hands make the causal requirement harder than bodies do. A hand covers a few thousand pixels, moves faster than the body it belongs to, is occluded by the object it holds and leaves the field of view, all while the camera itself is moving. Above all, the depth of the wrist, which fixes where the hand is in the world, cannot be read from a single frame the way a body's scale can (Zhang et al., 2025). Offline methods resolve it with a one-shot scale estimate or a clip-level optimization; an online method has to resolve it from the frames it has, and has to estimate the camera position without bundle adjustment at the same time.

We present StreamHaR, an online method that estimates the 3D hand pose in a metric world space at every frame of an egocentric stream. Each estimate uses only the frames received so far and a fixed-size state, with zero look-ahead. Our key motivation is to split the problem into three components. The hand's articulation is a recognition problem, which we learn from data. The wrist position is a measurement. Once, we know where the wrist appears in the image and the shape of the hand, its depth follows from projective geometry. The camera motion is independent of the hand and can be estimated separately. StreamHaR gives each its own estimator. Two hand encoders with separate roles: a keypoint encoder trained so that its image and 3D keypoints agree under projection, from which a closed-form per-frame Perspective-n-Point (PnP) solver estimates the wrist position, and a pose encoder that feeds a constant-memory articulation refiner. The camera is estimated by causal visual odometry whose metric scale is anchored repeatedly inside its own state from monocular depth. The approach does not require bundle adjustment, an one-shot scale estimate and any calibration. Figure 2 gives an overview. Our contributions are as follows:

- We present the first *online* world-space hand reconstruction method that is causal with zero look-ahead and has constant memory.
- We decompose the problem into three components that include estimating the wrist potions in closed form, the hand finger pose and shapes as part by recognition modules and the external camera position by incremental odometry without bundle adjustment.
- StreamHaR achieves state-of-the-performance using world-space metrics on Ego-Exo4D and transfers zero-shot to a HoloLens headset it was never trained on, where it surpasses the offline systems.

2 Related Work

Hand Pose and Shape from Images and Videos Early 3D hand pose estimation relied on depth sensors (Ge et al., 2016; Oikonomidis et al., 2011; Tagliasacchi et al., 2015). Monocular RGB methods then progressed rapidly, driven by the MANO parametric model (Romero et al., 2017) and large annotated datasets (Zimmermann et al., 2019; Moon et al., 2020; Jin et al., 2020). Scaling transformer backbones has since produced strong single-frame models, such as HaMeR (Pavlakos et al., 2024) and WiLoR (Potamias et al., 2025), which regress accurate MANO parameters from a single image crop. Beyond a single hand, interacting hands have been addressed with graph-based mesh regression (Li et al., 2022), cross-hand attention (Yu et al., 2023), learned interaction priors (Ren et al., 2023; Lee et al., 2023), and strategies that bridge the gap to in-the-wild images (Moon, 2023). Beyond a single frame, ReViV (Lyu et al., 2026) regresses hand joints over a short window of frames, which helps temporal stabilization. However, all of these methods predict the hand relative to the camera or to its own root joint, under a weak-perspective camera. They recover *what the hand does*, but not *where it is* in a fixed, metric world frame. StreamHaR builds on this line of work and adds what it lacks: a causal, metric, world-space hand trajectory.

World-space Motion Modelling. Recovering motion in world coordinates was first studied for the full body, using multi-camera rigs (Joo et al., 2015) or wearable sensors (Von Marcard et al., 2018; Kaufmann et al., 2023), and later from monocular video by separating camera motion from body motion. GLAMR (Yuan et al., 2022) infers the global trajectory from local body motion. Later works jointly optimize the body trajectory with camera poses from Simultaneous Localization and Mapping (SLAM) (Ye et al., 2023; Kocabas et al., 2024), WHAM (Shin et al., 2024) regresses the global trajectory directly, and TRAM (Wang et al., 2024) adds a metric-depth network to fix the SLAM scale. For *hands*, the same recipe has appeared only recently (Ye et al., 2025; Yu et al., 2025; Zhang et al., 2025; Ye et al., 2026): a pretrained single-frame model (Pavlakos et al., 2024; Potamias et al., 2025) predicts hands in the camera frame, and a camera estimate over the whole clip lifts them to world space. HaWoR (Zhang et al., 2025), the strongest egocentric method, combines DROID-SLAM (Teed & Deng, 2021) with global bundle adjustment over the full clip, Metric3Dv2 (Hu et al., 2024) for scale correction, and a bidirectional infiller. Dyn-HaMR (Yu et al., 2025) instead takes an optimization-based route, fitting two interacting hands and the camera jointly over the entire clip with a multi-stage objective in the style of SLAHMR (Ye et al., 2023); however, takes tens of minutes per sequence. Both methods are offline by design, since every frame must be available before any output is produced. In contrast, StreamHaR is, to our knowledge, the first streaming method for egocentric world-space hand reconstruction, and it surpasses the offline state of the art on world-space accuracy.

Streaming Inference. A streaming system produces an estimate for every frame from current and past inputs only, in constant memory, and does not revisit a frame once it has passed. Recent 3D perception methods achieve this with causal attention over cached frame tokens (Zhuo et al., 2026), a persistent recurrent state (Wang et al., 2025b), or incremental visual odometry (Teed et al., 2023; Murai et al., 2025). OnlineHMR (Zhao et al., 2026) and Human3R (Chen et al., 2026) bring world-grounded reconstruction to the streaming setting, but they target the full body rather than hands. World-space hand methods (Yu et al., 2025; Zhang et al., 2025), in turn, track the camera with DROID-SLAM (Teed & Deng, 2021) and global bundle adjustment, which rules out streaming. StreamHaR combines world-space hand reconstruction with causal visual odometry, making it suitable for streaming applications.

3 Method

Given an egocentric video that arrives one frame at a time, our goal is to recover the camera wearer's 3D hands in metric world coordinates at every frame. We work in the streaming setting, where the estimate for frame t uses only frames up to t. Sec. 3.1 introduces the parametric hand model, coordinate frames and notation. Sec. 3.2 recovers the hand's articulation and placement in the camera frame at every frame. Sec. 3.3 estimates the camera with a causal visual odometry system whose metric scale is anchored within its own state, so the trajectory is metric as soon as it is produced. Sec. 3.4 describes the training objectives. Finally, Appendix D describes a variant that accepts a one-second delay in exchange for better wrist depth, which we evaluate alongside the online system.

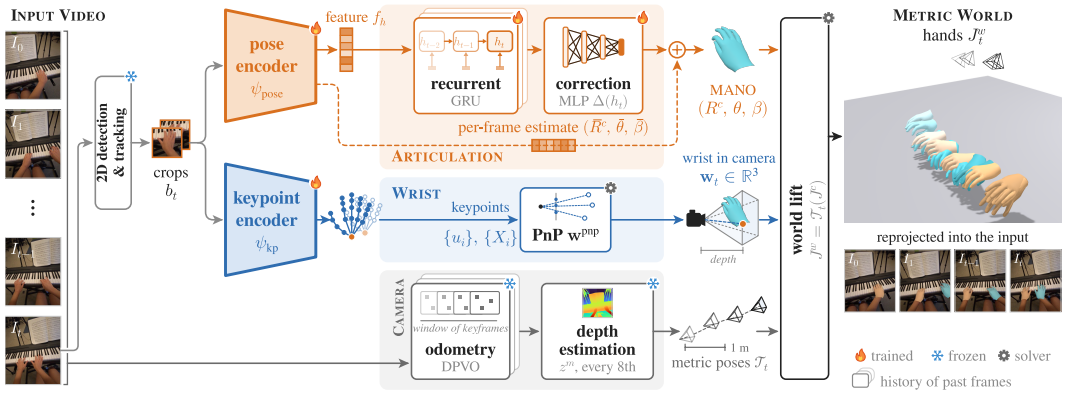


Figure 2: **StreamHaR.** Given an egocentric video, we use an off-the-shelf detector (Potamias et al., 2025) and vote-based tracking (Appendix A). The *keypoint encoder* ψ_{kp} estimates 2D and 3D wrist-centred keypoints from which the wrist is *measured* by a per-frame PnP solver. The *pose encoder* ψ_{pose} estimates the 3D hand parameters per frame, and an articulation refiner carries them across frames. An approach for causal odometry, metric-anchored from monocular depth, lifts the hand to the world frame. The estimated hand parameters, wrist position, and camera pose are then combined to provide an estimate of the hands in a metric world space. Trained, frozen and solver components are marked.

3.1 Preliminaries: Hand Model and Coordinate Frames

We represent each hand with the MANO (Romero et al., 2017) parametric hand model. MANO maps finger poses $\theta_t \in \mathbb{R}^{15\times 6}$, given as rotations in the 6D representation of Zhou et al. (2019), and shape $\beta_t \in \mathbb{R}^{10}$ to a mesh and 21 joints $J_t \in \mathbb{R}^{21\times 3}$ in a hand-centered frame, where joint 0 is the wrist.

Camera frame. Two quantities place this hand in the camera frame: its orientation $R^c_t \in SO(3)$ and the position $\mathbf{w}_t \in \mathbb{R}^3$ of its wrist joint,

$$J^c_{t,i} = R^c_t \left(J_{t,i} - J_{t,0}\right) + \mathbf{w}_t, \qquad i = 0, \ldots, 20, \tag{1}$$

so that $J^c_{t,0} = \mathbf{w}_t$ exactly. We use the wrist joint instead of MANO's root translation as reference, since we estimate the wrist position explicitly.

World frame. The camera pose at frame t consists of a rotation R_t and a metric position $\mathbf{p}_t$ in a world frame that stays fixed for the whole video. It maps the hand to the world coordinates,

$$J^w_{t,i} = R_t\, J^c_{t,i} + \mathbf{p}_t, \qquad R^w_t = R_t\, R^c_t, \tag{2}$$

and we write $\mathcal{T}_t(\mathbf{x}) = R_t\mathbf{x} + \mathbf{p}_t$ for this camera-to-world map.

3.2 Online Hand Recovery in the Camera Frame

An off-the-shelf detector provides a box b_t for each hand in frame I_t, and a causal vote keeps each hand's left / right identity consistent across frames (Appendix A). The following estimates are computed per hand and per frame, from the current frame and a recurrent state.

Wrist Measurement. The keypoint encoder ψ_{kp} is a pretrained ViT that takes the crop of I_t inside b_t and returns the 21 keypoints twice, as image points $u_i \in \mathbb{R}^2$ and as wrist-centred 3D points $X_i \in \mathbb{R}^3$:

$$\left(\{u_i\}, \{X_i\}\right) = \psi_{\text{kp}}(I_t,\, b_t). \tag{3}$$

Here, the wrist is the translation that brings the 3D points onto the image points,

$$\mathbf{w}^{\text{pnp}} = \arg\min_{\mathbf{w}\in\mathbb{R}^3} \sum_i \left\| \pi\left(X_i + \mathbf{w}\right) - u_i \right\|^2. \tag{4}$$

Since the wrist-centred points X_i already carry the hand's orientation, this is a PnP problem over the translation alone. Multiplying out the projection π makes it linear in $\mathbf{w}$, so we solve it in closed form

by least squares. The streaming system uses this solution directly as its wrist estimate, $\mathbf{w}_t = \mathbf{w}_t^{\text{pnp}}$, which depends on the current frame only.

Keypoint consistency. Eq. 4 returns the correct depth only if the two keypoint sets agree with each other under π. The 3D points, placed at the true wrist, must project onto the 2D points, and the metric size of $\{X_i\}$ must be right, because that size is what fixes the depth. A network trained for pose accuracy alone guarantees neither. We therefore train the keypoint encoder with an anchor loss that ties both keypoint sets to the same ground-truth camera-frame joints $J_i^\star$,

$$\mathcal{L}_{\text{anc}} = \sum_i \left\| u_i - \pi(J_i^\star) \right\|_1 + \sum_i \left\| X_i - (J_i^\star - J_0^\star) \right\|_1, \tag{5}$$

so that the two sets are consistent under π by construction.

Articulation Refinement. The pose encoder ψ_{pose} reads the same crop of I_t inside b_t, and returns an image feature f_h and a per-frame estimate of orientation, finger pose and shape:

$$\left(f_h,\ \bar{R}^c, \bar{\theta}, \bar{\beta}\right) = \psi_{\text{pose}}(I_t,\ b_t). \tag{6}$$

The pose encoder sees one crop at a time. Therefore, an *articulation refiner* carries its feature across frames: a Gated Recurrent Unit (GRU) (Cho et al., 2014) with state h_t, followed by an MLP head Δ whose last layer is initialized to zero and which predicts a residual on the per-frame estimate,

$$h_t = \text{GRU}\left(f_h,\ h_{t-1}\right), \qquad \left(R^c, \theta, \beta\right) = \left(\bar{R}^c, \bar{\theta}, \bar{\beta}\right) \oplus \Delta(h_t), \tag{7}$$

where $\oplus$ composes the orientation and finger rotations in 6D and adds the shape residual. Because Δ starts at zero, the untrained system reproduces the per-frame estimate exactly, while the training only adds the corrections that temporal supervision justifies. The state h_t is the only memory of the online system. The camera-frame joints of the frame follow from Eq. 1 with (R^c, θ, β) and $\mathbf{w}_t$.

3.3 Metric Camera from Incremental Odometry

Incremental odometry. The camera poses $(R_t, \mathbf{p}_t)$ are estimated by visual odometry using DPVO (Teed et al., 2023) on fisheye-rectified frames. DPVO is an online method that keeps a sliding window of recent keyframes, each carrying a few sparse patches with an inverse depth. The window is updated once per frame. Since DPVO estimates the camera trajectory only up to a global scale, we fix the scale repeatedly every eighth frame a, using everything DPVO has accumulated. Every patch j in the current window is re-projected into frame a, which gives its depth z_j^{vo} in the odometer's scale, and we obtain the metric depth z_j^{m} at the same pixel using the metric-depth network (Wang et al., 2025c). Patches that fall on the hand are excluded, since the hand is not part of the static scene. The scale of the window is the robust fit

$$g_a = \arg\min_g \sum_j \rho\left(g\, z_j^{\text{vo}} - z_j^{\text{m}}\right), \qquad \rho(r) = \frac{r^2}{(0.5\,\text{m})^2 + r^2}, \tag{8}$$

where ρ saturates for residuals beyond half a meter, such that a few wrong depth estimates do not confound the estimate. It is initialized at the median ratio and solved by iteratively reweighted least squares. The first anchor a rescales the entire trajectory so far. Each later anchor rescales only the current window by the damped factor $\sqrt{g_a}$. Camera centers, patch depths and the stored relative poses all transform exactly under a change of scale, so new patches emerge in metric, while already estimated poses are never rescaled. The entire processing is online and provides a camera trajectory in metric space.

3.4 Training

Both encoders are initialised by the pretrained hand Vision Transformer (ViT) (Potamias et al., 2025). While we finetune the keypoint encoder with the anchor loss of Eq. 5, the pose encoder and the articulation refiner are trained with the closure and articulation objectives, which we describe in Eqs. 12 and 9. During training, we use T=32 consecutive frames as window with known camera poses. Before computing the loss, we map each ground-truth joint into the corresponding camera frame using $\mathcal{T}_t^{-1}$. The visual odometry system and the metric-depth network stay frozen.

Articulation objective. Let J_i^t be the predicted joints of frame t and c_i^t their visibility. The articulation objective is

$$\mathcal{L} = \lambda_j \mathcal{L}_j + \lambda_\theta \mathcal{L}_\theta + \lambda_r \mathcal{L}_r + \lambda_\Delta \mathcal{L}_\Delta + \lambda_W \mathcal{L}_W, \tag{9}$$

where $\mathcal{L}_j$ is the visibility-weighted ℓ_1 error of the joints in 3D and in the image, $\mathcal{L}_\theta = \sum_t \arccos\big(\frac{1}{2}(\mathrm{tr}(R_t^{c\top} R_t^{c\star}) - 1)\big)$ is the geodesic error of the orientation, and $\mathcal{L}_r$ is a regularizer the includes a bone-length consistency term (Spurr et al., 2020), smoothness terms, and a shape prior $\|\beta\|^2$.

The last two terms are specific to the reconstruction in the world frame,

$$\mathcal{L}_\Delta = \textstyle\sum_{t,i} c_i^t \big\| \dot{J}_i^t - \dot{J}_i^{\star t} \big\|^2, \qquad \dot{J}^t := J^t - J^{t-1}, \tag{10}$$

$$\mathcal{L}_W = \min_{\mathrm{sim}} \textstyle\sum_{t,i} \big\| \mathrm{sim}\big(J_i^{w,t}\big) - J_i^{w\star,t} \big\|_1, \tag{11}$$

where $\dot{J}$ denotes the frame-to-frame displacement of the joints, J^w the joints lifted to the world frame with the known camera parameters, and sim a single similarity transform per window, over which we minimize. The *motion-matching* term $\mathcal{L}_\Delta$ is zero when the velocity of the joints is the same as the ground truth velocity. It therefore encourages temporal consistency without introducing a bias towards a non-moving hand. The *windowed alignment* term $\mathcal{L}_W$ measures the error of the lifted joints in the world frame but up to a similarity transform within each window. While this term ensures consistency within a window, the global location of the wrist is not considered as it is estimated separately. The loss weights are listed in Appendix A.

Orientation closure. Every term in $\mathcal{L}$ except $\mathcal{L}_\theta$ is invariant to a rotation of the whole hand about the wrist. Since our goal is to obtain accurate hand poses in a metric world coordinate system, we add an additional term to $\mathcal{L}_\theta$ that uses the ground-truth rotation $R_t^{c\star}$ as well. We use MANO forward kinematics FK to obtain the 3D joints given the estimated finger poses θ_t and shape parameters β_t, and compare it to the ground-truth joints:

$$\mathcal{L}_{\text{close}} = \sum_{t,i} c_i^t \left\| \mathrm{FK}_i\big(R_t^{c\star}, \theta_t, \beta_t\big) - J_i^{\star t} \right\|_1. \tag{12}$$

Training procedure. We do not train all components from the beginning but add them step by step to the training. First, we train the articulation refiner with both encoders frozen. Second, the pose encoder is additionally finetuned, with a learning rate $100\times$ lower than for the the articulation refiner. Third, starting from these weights, the keypoint encoder is finetuned with the *keypoint objective* $\mathcal{L}_{\text{anc}} + \mathcal{L}$, and the pose encoder and articulation refiner with the *pose objective* $\mathcal{L}_{\text{close}} + \mathcal{L}$.

4 Experiments

We evaluate StreamHaR quantitatively and qualitatively against camera-space and offline world-space baselines (Sec. 4.1). Implementation details are given in Appendix A and additional results in Appendix C; the supplementary material provides a video demonstration.

4.1 Setup

Datasets. Our five benchmarks cover both world-space trajectories under head motion and hand articulation across capture settings:

- *HOT3D-Aria* (Banerjee et al., 2025): Aria recordings of hand-object interaction, with MANO ground truth and a metric device trajectory.
- *Ego-Exo4D* (Grauman et al., 2024): Aria video from eight sites, with manual 3D hand pose annotations.
- *HOI4D* (Liu et al., 2022): long sequences from a head-mounted Kinect at 15 fps, with camera poses.
- *HO-Cap* (Wang et al., 2025a): HoloLens recordings, of which we use sequences from four subjects, including the two official test subjects. None of the methods is trained on HoloLens data, so this benchmark tests generalization to an unseen headset.
- *DexYCB* (Chao et al., 2021): static multi-camera video, evaluated with ground-truth hand crops to isolate articulation accuracy (Appendix B).

Metrics. We evaluate our method on three aspects. *Camera frame*: Mean Per-Joint Position Error (MPJPE) and Procrustes-Aligned MPJPE (PA-MPJPE) in mm, and the Area Under the Curve (AUC) of correctly localized keypoints. *World frame*: the frame-weighted joint error after one similarity alignment per clip, World and World Aligned (W-/WA-) MPJPE over 100-frame segments aligned on the first frame and on the whole segment, the median root-translation error (RTE) and the acceleration error (Accel), Global and Global Aligned (G-/GA-) MPJPE over 128-frame segments following Yu et al. (2025). *Latency*: the look-ahead, i.e. the number of future frames an output depends on, the throughput in Frames per Second (FPS), and the average delay between a frame's arrival and its output as in Zhao et al. (2026). Appendix B details every protocol.

Baselines. We compare with the per-frame estimators WiLoR (Potamias et al., 2025) and HaMeR (Pavlakos et al., 2024), the video model ReViV (Lyu et al., 2026) with a two-second window, and the offline world-space systems HaWoR (Zhang et al., 2025) and Dyn-HaMR (Yu et al., 2025). All are re-run from their released models and scored with the same code. Every system, ours included, runs the detector and tracker it ships. Both per-frame models read the boxes of the WiLoR detector, which is also the detector used by HaWoR and StreamHaR. Where a dataset supplies camera poses, every method is lifted with them.

4.2 Results

Table 1: **World-space hand reconstruction on HOT3D-Aria.** Errors in mm; FPS and average delay in sec. on one H200 GPU, for the world-space methods only (per-frame methods greyed). StreamHaR is the only world-space method with zero look-ahead.

Method	Online	Look-ahead	FPS ↑	Delay (s) ↓	World ↓	PA ↓	WA ↓	W ↓	RTE % ↓	Accel ↓
per-frame, camera space										
WiLoR (Potamias et al., 2025)	✓	0	—	—	—	4.99	—	—	—	—
HaMeR (Pavlakos et al., 2024)	✓	0	—	—	—	9.71	—	—	—	—
ReViV (Lyu et al., 2026)	✗	2 s	—	—	—	11.16	—	—	—	—
offline, world space										
HaWoR (Zhang et al., 2025)	✗	clip	4.55	16.39	33.00	5.76	**22.92**	**44.50**	**5.94**	**10.55**
Dyn-HaMR (Yu et al., 2025)	✗	clip	0.30	245.40	68.45	11.75	48.86	137.23	19.95	11.67
online, world space										
StreamHaR (Ours)	✓	**0**	**15.83**	**0.06**	**30.69**	**4.81**	25.33	56.92	9.64	12.07

World-space accuracy with zero look-ahead. Table 1 compares all methods on HOT3D-Aria, and Figure 3 shows the trajectories for three of its clips. StreamHaR matches the world-space accuracy of the offline state of the art. The World error is within noise of HaWoR's (30.69 vs. 33.00 mm) and 55% lower than that of Dyn-HaMR, which optimizes each clip separately. StreamHaR also has the lowest articulation error of all methods, including the per-frame ones. Its PA-MPJPE is 16% lower than HaWoR's and slightly below that of WiLoR (4.81 vs. 4.99 mm). HaWoR retains an advantage on the segment-level and smoothness metrics (W-MPJPE, WA-MPJPE, RTE and Accel), which its SLAM and motion infiller obtain by using future frames.

Online and efficient. Accuracy is only half of the comparison; Table 1 also reports what each answer costs. StreamHaR updates a constant-size state once per frame and runs at 15.83 FPS, so each frame's hand is available 0.06 s after the frame arrives. The offline systems must first see the whole clip: HaWoR runs at 4.55 FPS and answers 16.39 s after a frame on average, Dyn-HaMR at 0.30 FPS after 245.40 s. StreamHaR is thus 3.5× faster than HaWoR and answers more than 250× sooner; Table 7 breaks its frame time down by stage.

Fast head motion and long sequences. On Ego-Exo4D (Table 2), the offline systems reach a G-MPJPE of 67.79 mm (Dyn-HaMR) and 69.77 mm (HaWoR), against 40.35 mm for StreamHaR, which also leads on PA-MPJPE. Only HaMeR, which places the hand with its own camera model, has a lower absolute MPJPE; Dyn-HaMR's is dominated by four records on which its optimization diverges. On HOI4D, StreamHaR has the best MPJPE, while WiLoR and the two offline systems keep lower G- and GA-MPJPE.

Zero-shot transfer. On HO-Cap (Table 3), StreamHaR reaches 17.57 mm MPJPE and 6.23 mm PA-MPJPE, against 24.31 mm and 8.68 mm for HaWoR, the stronger offline system. Away from the

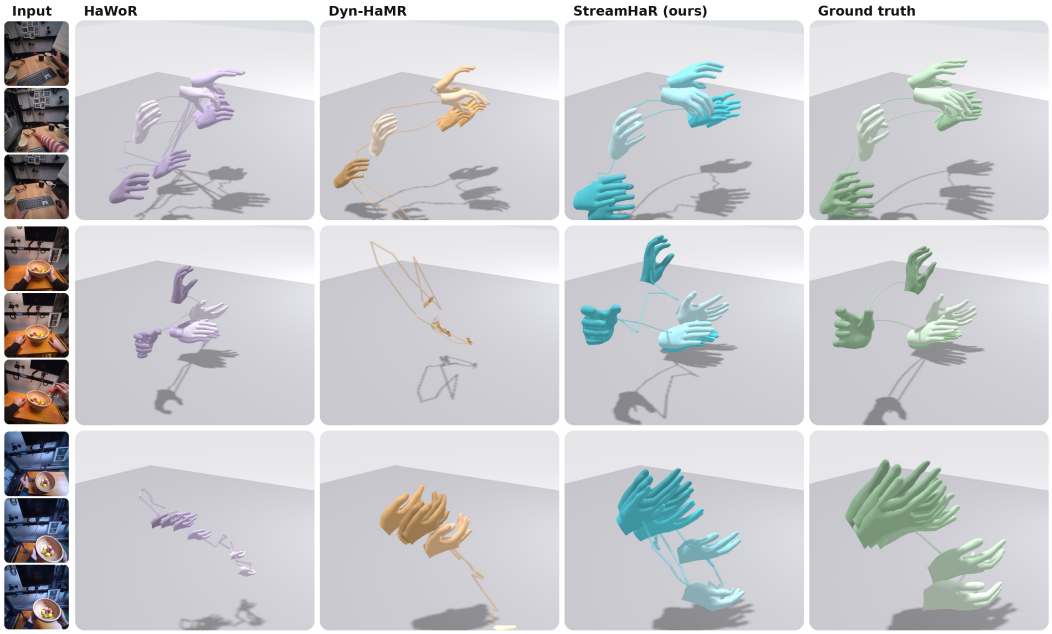


Figure 3: **Right-hand trajectories on challenging HOT3D-Aria clips.** Input frames, HaWoR, Dyn-HaMR, StreamHaR and ground truth, aligned as in Table 1. From top to bottom, the subjects pick up a mug and return to the keyboard, hold a bowl and reach to a shelf for a bottle, and lift a bowl of fruit off the table in a dark room. Where the baselines drift, lose scale or collapse, StreamHaR stays close to the ground truth.

Table 2: **Egocentric video with fast head motion (Ego-Exo4D) and long sequences (HOI4D).** Errors in mm; MPJPE is absolute on Ego-Exo4D and root-relative on HOI4D. StreamHaR leads all world-space metrics on Ego-Exo4D.

			Ego-Exo4D					HOI4D			
Method	Online	World	MPJPE ↓	PA ↓	G ↓	GA ↓	WA ↓	MPJPE ↓	PA ↓	G ↓	GA ↓
WiLoR (Potamias et al., 2025)	✓	✗	174.34	11.96	44.19	23.84	21.83	24.91	**9.42**	77.68	**28.11**
HaMeR (Pavlakos et al., 2024)	✓	✗	**52.06**	12.70	48.28	23.73	21.80	31.31	9.68	94.41	33.24
Dyn-HaMR (Yu et al., 2025)	✗	✓	800.75	15.89	67.79	29.61	27.27	25.51	11.12	84.90	28.54
HaWoR (Zhang et al., 2025)	✗	✓	93.76	13.76	69.77	28.59	26.15	26.24	10.36	**76.08**	29.12
StreamHaR (Ours)	✓	✓	83.89	**11.92**	**40.35**	**21.89**	**20.44**	**24.76**	9.43	89.04	33.09

shared training data, StreamHaR's articulation lead over HaWoR widens from 0.95 mm PA-MPJPE on HOT3D-Aria to 2.45 mm.

In-the-wild video. The same model also runs unchanged on real-world head-mounted video. Figure 4 compares it with HaWoR in the camera view on factory footage from Egocentric-100K (Build AI, 2025). Appendix C.2 adds the world-space view, both hands and the head path, on the uncalibrated kitchen clips released with HaWoR, for which we assume a focal length.

Table 3: **Zero-shot transfer to an unseen device (HO-Cap, HoloLens).** Camera-frame, root-relative errors in mm. StreamHaR lowers both errors by about a quarter.

Method	Online	World	MPJPE ↓	PA ↓
Dyn-HaMR (Yu et al., 2025)	✗	✓	24.88	9.18
HaWoR (Zhang et al., 2025)	✗	✓	24.31	8.68
StreamHaR (Ours)	✓	✓	**17.57**	**6.23**

Ablation studies. *Articulation path* (Table 4) shows that a single shared encoder cannot match two specialized roles. Trained with both objectives, it lowers the World error to 28.15 mm, compared with 30.69 mm for the two specialists. However, its articulation error rises on all three benchmarks, to a PA-MPJPE of 5.29, 6.47 and 5.02 mm on HOT3D-Aria, HO-Cap and DexYCB, respectively. Trained with the keypoint objective alone, it is worse on every metric. The two encoders therefore trade 2.5 mm of World error on HOT3D-Aria for better articulation on every benchmark and better

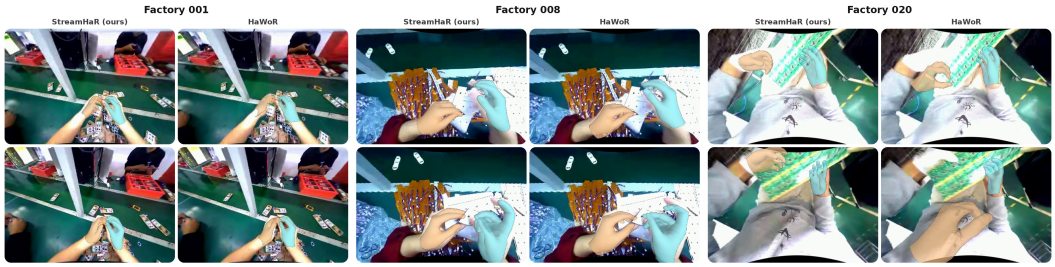


Figure 4: **Qualitative comparison with HaWoR on in-the-wild egocentric factory video.** Egocentric-100K (Build AI, 2025) workers at three sites, where each pair shows both methods on the same frame. StreamHaR keeps both hands through gloves, motion blur and the image border, where HaWoR drops a hand or places it in empty space.

Table 4: **Ablation of the articulation path.** Each row removes one component of StreamHaR: *one encoder* replaces the two specialists with a single encoder trained with the keypoint objective or with both, and *no articulation refiner* keeps the pose encoder's per-frame output.

	HOT3D-Aria		HO-Cap		DexYCB	
Variant	World ↓	PA ↓	MPJPE ↓	PA ↓	PA ↓	AUC ↑
one encoder, both	**28.15**	5.29	18.17	6.47	5.02	90.0
one encoder, keypoint	31.40	5.93	17.91	6.60	5.22	89.6
no articulation refiner	30.68	4.83	**17.55**	6.27	4.97	90.1
StreamHaR	30.69	**4.81**	17.57	**6.23**	**4.88**	**90.2**

Table 5: **Ablation of the placement path.** Each row removes one component of StreamHaR on HOT3D-Aria; none changes the articulation, so PA-MPJPE is the same in every row. *No anchor loss*: PnP on the pose encoder's keypoints.

Variant	World ↓	WA ↓	W ↓	RTE % ↓
no anchor loss	37.99	33.52	73.76	12.06
no metric anchor	40.72	35.56	86.93	13.78
one anchor at start	34.05	28.07	67.36	11.59
StreamHaR	**30.69**	**25.33**	**56.92**	**9.64**

transfer to HO-Cap. Finally, the articulation refiner lowers PA-MPJPE on each benchmark, most on DexYCB (4.97 to 4.88 mm). *Placement path* (Table 5) illustrates why every component is needed. Without the anchor loss, the World error rises by 7.3 mm, because PnP then reads keypoints that are not trained to agree under projection; without the metric anchor inside the odometer, it rises by 10.0 mm; anchoring only at the start costs 3.4 mm World error and 10.4 mm W-MPJPE.

5 Conclusion

We presented StreamHaR, an online method that reconstructs metric, world-space hands from egocentric video, emitting the estimate for each frame from the frames received so far and a state of constant size. Three design decisions drive the results. First, we split the world-space hand reconstruction into three subproblems, articulation, wrist placement and camera motion, and estimate each with its own estimator. Two specialized hand encoders handle the first two. Second, a per-frame PnP solve places the wrist using keypoints trained to be consistent under projection, so its depth follows from geometry rather than from a learned guess. Third, the camera comes from a causal visual odometry system whose metric scale is repeatedly re-anchored within its own state using monocular depth. The trajectory is therefore metric as soon as it is produced, without offline optimization or camera calibration. Without using any future frames, StreamHaR matches the world-space accuracy of the offline state of the art on HOT3D-Aria while achieving lower articulation error. It also transfers zero-shot to HoloLens video, and achieves the best world-space results on Ego-Exo4D, where head motion is fast.

6 AI Use Statement

In this work, we used generative AI tools to design and give feedback on experiments and ablations. We have not used generative AI tools to generate synthetic datasets, to develop theoretical models or conceptual frameworks, or to assist with translation. Formulating mathematical claims, proving them, writing proofs, and qualitative or thematic data analysis are also not applicable to this work. Additionally, we used generative AI tools to modify figures, to draft and edit parts of the paper for

readability. All AI-assisted work has been reviewed. Every reported number is produced by the evaluation scripts in our code base, and every figure was checked against the predictions it shows. We take responsibility for the final content of this work, including text, claims or artifacts produced.

REFERENCES

Prithviraj Banerjee, Sindi Shkodrani, Pierre Moulon, Shreyas Hampali, Shangchen Han, Fan Zhang, Linguang Zhang, Jade Fountain, Edward Miller, Selen Basol, et al. Hot3d: Hand and object tracking in 3d from egocentric multi-view videos. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2025.

Build AI. Egocentric-100k, 2025.

Yu-Wei Chao, Wei Yang, Yu Xiang, Pavlo Molchanov, Ankur Handa, Jonathan Tremblay, Yashraj S Narang, Karl Van Wyk, Umar Iqbal, Stan Birchfield, Jan Kautz, and Dieter Fox. DexYCB: A benchmark for capturing hand grasping of objects. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021.

Yue Chen, Xingyu Chen, Yuxuan Xue, Anpei Chen, Yuliang Xiu, and Gerard Pons-Moll. Human3r: Everyone everywhere all at once. In *Int. Conf. Learn. Represent.*, 2026.

Xuxin Cheng, Jialong Li, Shiqi Yang, Ge Yang, and Xiaolong Wang. Open-TeleVision: Teleoperation with immersive active visual feedback. In *Conf. Robot Learn.*, 2024.

Kyunghyun Cho, Bart Van Merriënboer, Çağlar Gulçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Conf. Empirical Methods Nat. Lang. Process.*, 2014.

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 *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023.

Qichen Fu, Xingyu Liu, Ran Xu, Juan Carlos Niebles, and Kris M Kitani. Deformer: Dynamic fusion transformer for robust hand pose estimation. In *Int. Conf. Comput. Vis.*, 2023.

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 *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016.

Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. Ego-Exo4D: Understanding skilled human activity from first- and third-person perspectives. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2024.

Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. Honnotate: A method for 3d annotation of hand and object poses. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020.

Shangchen Han, Po-chen Wu, Yubo Zhang, Beibei Liu, Linguang Zhang, Zheng Wang, Weiguang Si, Peizhao Zhang, Yujun Cai, Tomas Hodan, Randi Cabezas, Luan Tran, Muzaffer Akbay, Tsz-Ho Yu, Cem Keskin, and Robert Wang. UmeTrack: Unified multi-view end-to-end hand tracking for VR. In *ACM SIGGRAPH Asia Conf. Papers*, 2022.

Ryan Hoque, Peide Huang, David J. Yoon, Mouli Sivapurapu, and Jian Zhang. EgoDex: Learning dexterous manipulation from large-scale egocentric video. In *Int. Conf. Learn. Represent.*, 2026.

Mu Hu, Wei Yin, Chi Zhang, Zhipeng Cai, Xiaoxiao Long, Hao Chen, Kaixuan Wang, Gang Yu, Chunhua Shen, and Shaojie Shen. Metric3d v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2024.

Sheng Jin, Lumin Xu, Jin Xu, Can Wang, Wentao Liu, Chen Qian, Wanli Ouyang, and Ping Luo. Whole-body human pose estimation in the wild. In *Eur. Conf. Comput. Vis.*, 2020.

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 *Int. Conf. Comput. Vis.*, 2015.

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 *Int. Conf. Comput. Vis.*, 2023.

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 *Int. Conf. 3D Vis.* IEEE, 2024.

Jihyun Lee, Minhyuk Sung, Honggyu Choi, and Tae-Kyun Kim. Im2hands: Learning attentive implicit representation of interacting two-hand shapes. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023.

Kailin Li, Puhao Li, Tengyu Liu, Yuyang Li, and Siyuan Huang. ManipTrans: Efficient dexterous bimanual manipulation transfer via residual learning. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2025.

Mengcheng Li, Liang An, Hongwen Zhang, Lianpeng Wu, Feng Chen, Tao Yu, and Yebin Liu. Interacting attention graph for single image two-hand reconstruction. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022.

Kevin Lin, Lijuan Wang, and Zicheng Liu. Mesh graphormer. In *Int. Conf. Comput. Vis.*, 2021a.

Kevin Lin, Lijuan Wang, and Zicheng Liu. Mesh graphormer. In *Int. Conf. Comput. Vis.*, 2021b.

Yunze Liu, Yun Liu, Che Jiang, Kangbo Lyu, Weikang Wan, Hao Shen, Boqiang Liang, Zhoujie Fu, He Wang, and Li Yi. HOI4D: A 4d egocentric dataset for category-level human-object interaction. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022.

Xiaozhong Lyu, Gen Li, Zhiyin Qian, Xucong Zhang, Marc Pollefeys, and Siyu Tang. ReViV: Reconstructing the viewer and the view in 4D from monocular egocentric video. In *Eur. Conf. Comput. Vis.*, 2026.

Gyeongsik Moon. Bringing inputs to shared domains for 3d interacting hands recovery in the wild. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023.

Gyeongsik Moon, Shoou-I Yu, He Wen, Takaaki Shiratori, and Kyoung Mu Lee. Interhand2. 6m: A dataset and baseline for 3d interacting hand pose estimation from a single rgb image. In *Eur. Conf. Comput. Vis.*, 2020.

Riku Murai, Eric Dexheimer, and Andrew J. Davison. MASt3R-SLAM: Real-time dense SLAM with 3D reconstruction priors. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2025.

Iason Oikonomidis, Nikolaos Kyriazis, and Antonis Argyros. Efficient model-based 3d tracking of hand articulations using kinect. In *Brit. Mach. Vis. Conf.*, 2011.

JoonKyu Park, Yeonguk Oh, Gyeongsik Moon, Hongsuk Choi, and Kyoung Mu Lee. Handoccnet: Occlusion-robust 3d hand mesh estimation network. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022.

Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. Reconstructing hands in 3d with transformers. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2024.

Rolandos Alexandros Potamias, Jinglei Zhang, Jiankang Deng, and Stefanos Zafeiriou. Wilor: End-to-end 3d hand localization and reconstruction in-the-wild. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2025.

Pengfei Ren, Chao Wen, Xiaozheng Zheng, Zhou Xue, Haifeng Sun, Qi Qi, Jingyu Wang, and Jianxin Liao. Decoupled iterative refinement framework for interacting hands reconstruction from a single rgb image. In *Int. Conf. Comput. Vis.*, 2023.

Javier Romero, Dimitris Tzionas, and Michael J Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics*, 36(6), 2017.

Soyong Shin, Juyong Kim, Eni Halilaj, and Michael J Black. Wham: Reconstructing world-grounded humans with accurate 3d motion. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2024.

Adrian Spurr, Umar Iqbal, Pavlo Molchanov, Otmar Hilliges, and Jan Kautz. Weakly supervised 3d hand pose estimation via biomechanical constraints. In *Eur. Conf. Comput. Vis.*, 2020.

Andrea Tagliasacchi, Matthias Schröder, Anastasia Tkach, Sofien Bouaziz, Mario Botsch, and Mark Pauly. Robust articulated-icp for real-time hand tracking. *Comput. Graph. Forum*, 2015.

Zachary Teed and Jia Deng. Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras. *Adv. Neural Inform. Process. Syst.*, 2021.

Zachary Teed, Lahav Lipson, and Jia Deng. Deep patch visual odometry. In *Adv. Neural Inform. Process. Syst.*, 2023.

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 *Eur. Conf. Comput. Vis.*, 2018.

Jikai Wang, Qifan Zhang, Yu-Wei Chao, Bowen Wen, Xiaohu Guo, and Yu Xiang. HO-Cap: A capture system and dataset for 3d reconstruction and pose tracking of hand-object interaction. In *Adv. Neural Inform. Process. Syst.*, 2025a.

Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2025b.

Ruicheng Wang, Sicheng Xu, Yue Dong, Yu Deng, Jianfeng Xiang, Zelong Lv, Guangzhong Sun, Xin Tong, and Jiaolong Yang. MoGe-2: Accurate monocular geometry with metric scale and sharp details. In *Adv. Neural Inform. Process. Syst.*, 2025c.

Yufu Wang, Ziyun Wang, Lingjie Liu, and Kostas Daniilidis. Tram: Global trajectory and motion of 3d humans from in-the-wild videos. In *Eur. Conf. Comput. Vis.* Springer, 2024.

Vickie Ye, Georgios Pavlakos, Jitendra Malik, and Angjoo Kanazawa. Decoupling human and camera motion from videos in the wild. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023.

Yufei Ye, Yao Feng, Omid Taheri, Haiwen Feng, Shubham Tulsiani, and Michael J Black. Predicting 4d hand trajectory from monocular videos. *arXiv preprint arXiv:2501.08329*, 2025.

Yufei Ye, Jiaman Li, Ryan Rong, and C Karen Liu. Whole: World-grounded hand-object lifted from egocentric videos. *IEEE Conf. Comput. Vis. Pattern Recog. Findings*, 2026.

Zhengdi Yu, Shaoli Huang, Chen Fang, Toby P Breckon, and Jue Wang. Acr: Attention collaboration-based regressor for arbitrary two-hand reconstruction. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023.

Zhengdi Yu, Shaoli Huang, Yongkang Cheng, and Tolga Birdal. SignAvatars: A large-scale 3D sign language holistic motion dataset and benchmark. In *Eur. Conf. Comput. Vis.*, 2024.

Zhengdi Yu, Stefanos Zafeiriou, and Tolga Birdal. Dyn-HamR: Recovering 4d interacting hand motion from a dynamic camera. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2025.

Ye Yuan, Umar Iqbal, Pavlo Molchanov, Kris Kitani, and Jan Kautz. Glamr: Global occlusion-aware human mesh recovery with dynamic cameras. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022.

Gu Zhang, Qicheng Xu, Haozhe Zhang, Jianhan Ma, Long He, Yiming Bao, Zeyu Ping, Zhecheng Yuan, Chenhao Lu, Chengbo Yuan, et al. UniDex: A robot foundation suite for universal dexterous hand control from egocentric human videos. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2026.

Jinglei Zhang, Jiankang Deng, Chao Ma, and Rolandos Alexandros Potamias. Hawor: World-space hand motion reconstruction from egocentric videos. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2025.

Yiwen Zhao, Ce Zheng, Yufu Wang, Hsueh-Han Daniel Yang, Liting Wen, and László A. Jeni. Onlinehmr: Video-based online world-grounded human mesh recovery. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2026.

Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019.

Dong Zhuo, Wenzhao Zheng, Jiahe Guo, Yuqi Wu, Jie Zhou, and Jiwen Lu. Streaming visual geometry transformer. In *Int. Conf. Learn. Represent.*, 2026.

Christian Zimmermann, Duygu Ceylan, Jimei Yang, Bryan Russell, Max Argus, and Thomas Brox. Freihand: A dataset for markerless capture of hand pose and shape from single rgb images. In *Int. Conf. Comput. Vis.*, 2019.

Ronglai Zuo, Rolandos Alexandros Potamias, Evangelos Ververas, Jiankang Deng, and Stefanos Zafeiriou. Signs as tokens: A retrieval-enhanced multilingual sign language generator. In *Int. Conf. Comput. Vis.*, 2025.