HOT3D

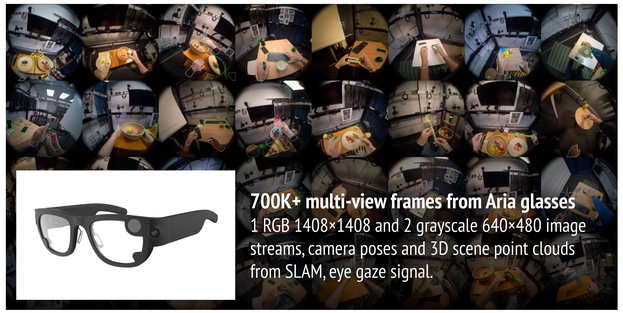
Hand and Object Tracking in 3D from Egocentric Multi-View Videos

Meta

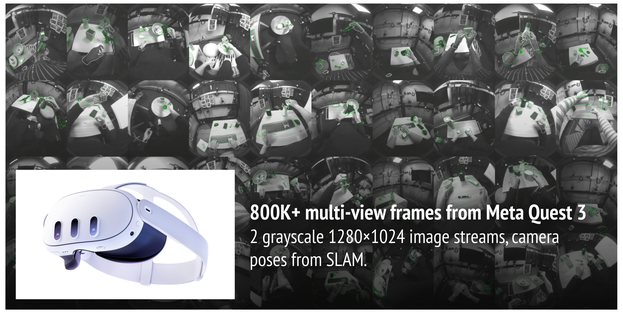
Prithviraj Banerjee, Sindi Shkodrani, Pierre Moulon, Shreyas Hampali, Shangchen Han, Fan Zhang, Linguang Zhang, Jade Fountain, Edward Miller, Selen Basol, Richard Newcombe, Robert Wang, Jakob Engel, Tomas Hodan

Contributions

1. **HOT3D dataset** for benchmarking egocentric 3D vision tasks on hands and objects (tracking, reconstruction, hand-object interaction understanding, etc.).
2. **Strong baselines for tasks enabled by HOT3D:** multi-view 6DoF object pose estimation and 3D lifting of unknown in-hand objects (DINOv2-based).
3. **Demonstrated effectiveness of multi-view egocentric data** for 3D hand tracking, 6DoF object pose estimation, and 3D lifting of in-hand objects.



700K+ multi-view frames from Aria glasses
1 RGB 1408×1408 and 2 grayscale 640×480 image streams, camera poses and 3D scene point clouds from SLAM, eye gaze signal.

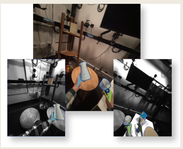


800K+ multi-view frames from Meta Quest 3
2 grayscale 1280×1024 image streams, camera poses from SLAM.

Example Aria frames with contours of hands and objects in GT poses shown in white and green.

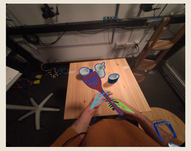
3D scene point cloud with 3D models in GT poses.

HOT3D dataset properties



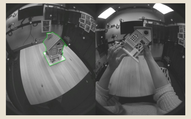
833 minutes of egocentric, multi-view, synchronized recordings
1.5M multi-view frames (3.7M+ images) recorded at 30 FPS with Project Aria, a research prototype of AI glasses, and Quest 3, a VR headset sold in millions of units.

19 participants, 4 everyday scenarios
In addition to simple pick-up/observe/put-down actions, recordings show scenarios resembling typical actions in a kitchen, office, and living room.



Accurate 3D GT annotations of hands & objects
Hands and objects are annotated with ground-truth 3D poses and shapes. Poses were obtained by a professional MoCap system using small (3mm) optical markers.

High-fidelity 3D object models
3D models of 33 rigid objects captured with high-res geometry and PBR materials, using an in-house 3D scanner. Objects of diverse appearance, size, and affordances.



Sequences for object onboarding
Two types of reference sequences (with static and hand-manipulated objects) to enable research on model-free 3D object tracking and 3D reconstruction.

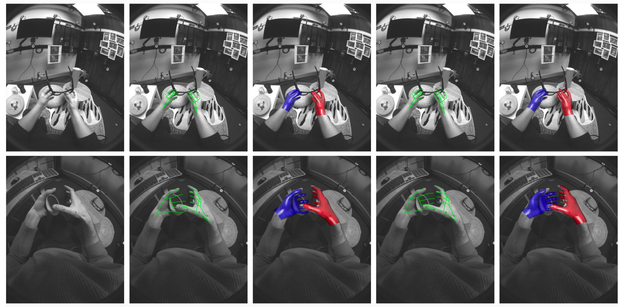
HOT3D-Clips

- **A set of curated sub-sequences** provided to enable straightforward comparison of various tracking and pose estimation methods.
- **Each of 3832 clips has 150 frames (5 sec)** with GT annotations for all modeled objects and hands and passing our visual inspection.
- **Already used in public challenges:** BOP 2024 and Hand Tracking 2024.

Experiment 1: 3D hand pose tracking

Task: Given a hand shape (3D hand skeleton in the canonical pose) and GT 2D bounding boxes of visible hands, estimate the hand poses (3D locations of skeleton joints) in every frame of the input sequence. Evaluated on HOT3D and UmeTrack datasets.

Methods: Single- and multi-view variants of the UmeTrack tracker (Han 2022).



Input | GT skeleton | GT mesh | Pred. skeleton | Pred. mesh

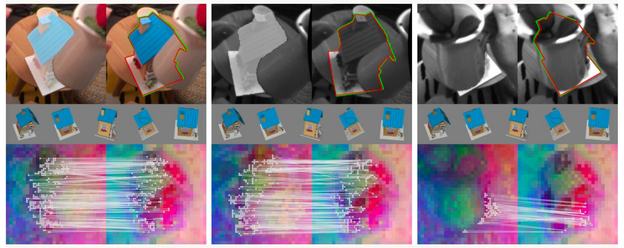
Training dataset	Views	MKPE on UmeTrack ↓	MKPE on HOT3D ↓
UmeTrack	1	13.6	24.2
UmeTrack	2	9.7	25.6
HOT3D-Quest3	1	23.7	18.0
HOT3D-Quest3	2	30.3	13.1
UmeTrack + HOT3D-Quest3	1	13.4	15.4
UmeTrack + HOT3D-Quest3	2	**9.5**	**10.9**

Hand tracking from two views is 41% more accurate than from a single view (9.5 vs 13.4 MKPE on UmeTrack and 10.9 vs 15.4 on HOT3D).

Experiment 2: CAD-based 6DoF object pose estimation

Task: Given a single frame and a CAD model of the target object, estimate the 6DoF pose of the object seen in the frame (without any object-specific training).

Methods: Single-view FoundPose (Örnek 2024) and our multi-view extension.



The pose (red contour) is estimated by gPnP-RANSAC from 2D-3D correspondences established between retrieved templates and multiple input views (3 in case of Aria).

Test dataset	Views	Recall [%] ↑ 5 cm, 5°	10 cm, 10°	20 cm, 20°
HOT3D-Aria	1	25.2	41.7	54.5
HOT3D-Aria	3	**33.8**	**52.9**	**66.2**
HOT3D-Quest3	1	28.9	46.6	58.9
HOT3D-Quest3	2	**36.9**	**55.9**	**66.4**

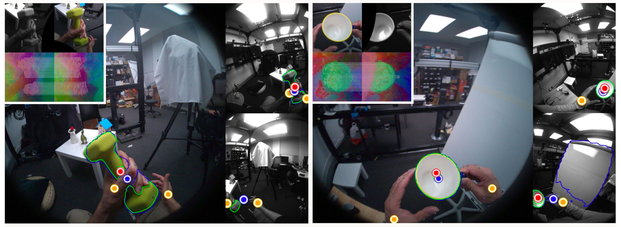
Our multi-view FoundPose extension is 13–34% more accurate than the original single-view variant.

Experiment 3: 3D lifting of unknown in-hand objects

Task: Given per-view 2D segmentation masks of an unknown in-hand object, predicted by EgoHOS (Zhang 2022) or our Mask R-CNN, the goal is to estimate the 3D object location. Useful for object indexing and long-term tracking.

Methods:

- **HandProxy:** Using GT 3D palm center as the 3D object location.
- **MonoDepth:** Similar to OSNOM (Plizzari 2024), relying on monocular depth prediction (using Depth Anything V2) aligned to a 3D point cloud from SLAM.
- **StereoMatch:** Stereo matching of DINOv2 patch features, converting the 2D-2D matches to a 3D point cloud via triangulation, and taking the robust mean.

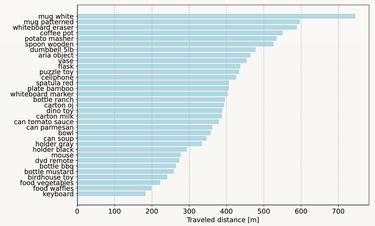


GT 3D object locations are shown in green, predictions from StereoMatch in red, MonoDepth in blue, and HandProxy in orange. DINOv2 stereo matching is in the top left.

Method	Test dataset	Views	Recall [%] ↑ 5 cm	10 cm	20 cm	30 cm
HandProxy	HOT3D-Aria	–	0.5	13.5	90.6	98.4
Using ground-truth 2D segmentation masks:						
MonoDepth	HOT3D-Aria	1	14.3	30.2	53.6	69.9
StereoMatch	HOT3D-Aria	3	**64.4**	**86.2**	**95.5**	**96.9**
StereoMatch	HOT3D-Quest3	2	76.4	96.8	99.1	99.2
Using 2D segmentation masks predicted by MRCNN-DA:						
MonoDepth	HOT3D-Aria	1	11.1	23.3	43.7	58.2
StereoMatch	HOT3D-Aria	3	**42.6**	**56.4**	**63.6**	**66.0**
StereoMatch	HOT3D-Quest3	2	59.1	75.3	80.4	81.3

StereoMatch significantly outperforms the single-view MonoDepth approach, especially at stricter thresholds of correctness. HandProxy is competitive only at relaxed thresholds (e.g. 30cm used in OSNOM).

Bonus: Distances traveled by HOT3D objects



Participants moved the 33 objects over 13 km. While objects like the keyboard and waffles were mostly resting, the white mug is a true explorer!

Download HOT3D dataset and toolkit:
facebookresearch.github.io/hot3d