HOT3D: Hand and Object Tracking in 3D from Egocentric Multi-View Videos (Supplementary Material)

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 Julian Engel, Tomas Hodan

Meta Reality Labs facebookresearch.github.io/hot3d

In this supplement, we provide details about the Aria glasses (App. A) and the Quest 3 headset (App. B) which were used to record the HOT3D dataset. We also describe our procedure for ground-truth annotation (App. C), provide additional data statistics (App. D) and quantitative results (App. E).

A. Aria glasses

Project Aria [1] is an egocentric recording device in glasses form-factor created by Meta. It is designed as a *research tool* for egocentric machine perception and contextualized AI research, and available to researchers across the world via projectaria.com.

A.1. Device and sensors

Project Aria is built to emulate future AR/smart glasses catering to machine perception and egocentric AI rather than human consumption. Aria is designed to be wearable for long periods of time without obstructing or impeding the wearer, allowing for natural motion even when performing highly dynamic activities, such as playing soccer or dancing. Its total weight is 75g (a single GoPro camera has over 150g) and fits just like a pair of glasses.

Further, the device integrates a rich sensor suite that is tightly calibrated and time-synchronized, capturing a broad range of modalities. For HOT3D, *recording profile 15* is used, which uses the following sensor configuration (all streams come with metadata such as precise timestamps and per-frame exposure times):

- **One rolling-shutter RGB camera** recording at 30 fps and 1408×1408 px. The camera is fitted with an F-Theta fisheye lens with a field of view (FOV) of $110°$.
- **Two global-shutter monochrome cameras** recording at 30 fps and 640×480 px. These cameras provide additional peripheral vision and are fitted with F-Theta fisheye lenses with $150°$ FOV.
- **Two monochrome eye-tracking cameras** recording at 10 fps and 320×240 px resolution.
- **Two IMUs** (800 Hz and 1000 Hz respectively), **a barometer** (50 fps) and **a magnetometer** (10 fps).
- **GNSS and WiFi** scanning was disabled for HOT3D.
- **Audio** recording was disabled for HOT3D for privacy reasons.

Figure 1. **Project Aria research glasses.**

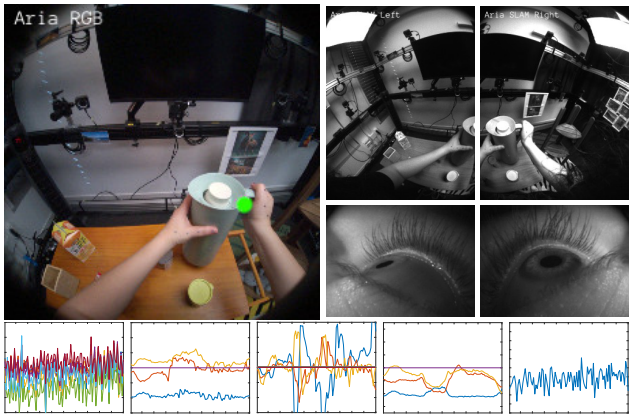

Figure 2. **Sensor streams recorded by the Project Aria device.** Top: RGB camera, left and right monochrome and eye cameras. Bottom: 10-second extracts from microphones, accelerometer, gyroscope, magnetometer and barometer respectively.

A.2. Machine Perception Services (MPS)

Project Aria's machine perception service (MPS) provides software building blocks that simplify leveraging the different modalities recorded. These functionalities are likely to be available as real-time, on-device capabilities in future AR- or smart-glasses. We use the following core functionalities currently offered by Project Aria, and include their raw output as part of the dataset. See [1] and the technical documentation[1] for more details.

[1] https://facebookresearch.github.io/projectaria_tools/docs/intro

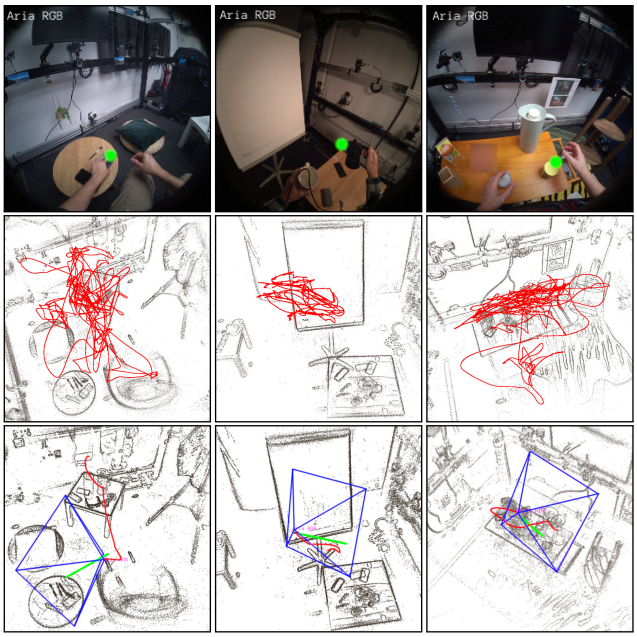

Figure 3. **Aria MPS output.** Shown is output for three recordings in a living room, office and kitchen scenario respectively (left to right). Top: RGB view and gaze (green dot). Middle: Point cloud and estimated egocentric camera trajectory for the full recording. Bottom: 3D view of a specific point in time, showing the RGB camera frustum (blue), gaze vector (green) and trajectory from the previous second (red).

Calibration. All sensors are intrinsically and extrinsically calibrated, and tiny deformations due to temperature changes or stress applied to the glasses frame are further corrected by time-varying online calibration from MPS.

Aria 6 DoF localization. Every recording is localized precisely and robustly in a common, metric, gravity-aligned coordinate frame, using a state-of-the-art VIO and SLAM algorithm. This provides millimeter-accurate 6 DoF poses for every captured frame and high-frequency (1 kHz) motion in-between frames.

Eye gaze. The gaze direction of the user is estimated as two outward-facing rays anchored approximately at the wearer's eyes, allowing to approximately estimate not only the direction the user is looking in, but also the depth their eyes are focused on. We use an optional eye gaze calibration procedure, where the mobile companion app directs the wearer to gaze at a pattern on the phone screen while performing specific head movements. This information was then used to generate a more accurate eye gaze direction, personalized to the particular wearer.

Point clouds. A 3D point cloud of static scene elements is triangulated from the moving Aria device, using photometric stereo over consecutive frames or left/right SLAM camera. Points are added causally over time, and will include points on any object that is observed while static for several seconds. The output contains both the 3D point clouds as well as the raw 2D observations of every point in the camera images it was triangulated from.

A.3. Processing summary

All Aria recordings are anonymized in a very first step, using the public EgoBlur [5] model and following Project Aria's responsible innovation principles.

Then, the MPS pipeline is invoked for each full Aria recording, which are typically about 2 minutes long and include many instances of hand-object interactions with different objects. Next, we 7DoF-align the MPS output with the OptiTrack coordinate frame (App. C). In total, we have processed 199 Aria recordings with a total length of 391 minutes.

A.4. Tools and ecosystem

Documentation and open-source tooling for Aria recordings and MPS output is available on GitHub[2] and includes Python and C++ tools to convert, load, and visualize data, as well as sample code for common computer vision tasks.

B. Quest 3 headset

Quest 3 [4], shown in Fig. 4, is the latest production headset from Meta for virtual- and mixed-reality experiences. For the HOT3D data collection we used a developer version of the Quest 3 headset. This version has four global-shutter monochrome cameras with fisheye lenses, 1280x1024 px image resolution, 18 PPD (Pixels Per Degree), and records at 30 fps. Two of the cameras are on the front side of the headset, roughly aligned with eyes, and two on the sides. HOT3D only includes images from the two front cameras as those capture the relevant scene part (the two side cameras are useful for applications like SLAM). Example images are in Fig. 5. Data from other sensors present in the consumer version of Quest 3, including a gyroscope and an accelerometer, were not recorded. The intrinsic and extrinsic parameters of the headset cameras were calibrated with a ChArUco board. Both the headset and the board were attached a set of optical markers and tracked by the motion-capture system described in App. C, which allowed to estimate camera-to-headset transformations. At recording time, the headset pose was still tracked by the motion-capture system and used to calculate per-frame camera-to-world transformations.

C. Marker-based motion capture

The poses of hands and objects were tracked using optical markers attached on their surface. For both hands and objects we used 3 mm markers with an adhesive layer at their bottom. Such markers are small enough not to influence hand-object interactions. Each hand was attached 19 markers and each object around 10. The marker locations were then semi-automatically registered to 3D models of hands and objects obtained by custom 3D scanners.

At recording time, the optical markers were tracked by multiple infrared OptiTrack cameras attached on a rig shown in Fig. 5 of the main paper. The intrinsic and extrinsic parameters of the infrared cameras were calibrated before every capturing session. Hand poses were calculated by fitting the participant's UmeTrack

[2] https://github.com/facebookresearch/projectaria_tools

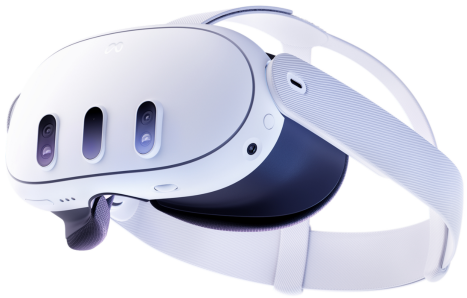

Figure 4. **Meta Quest 3 headset for virtual and mixed reality.**

Figure 5. **Sample images from Quest 3.** Shown are synchronized images from the two front Quest 3 cameras used for the HOT3D collection.

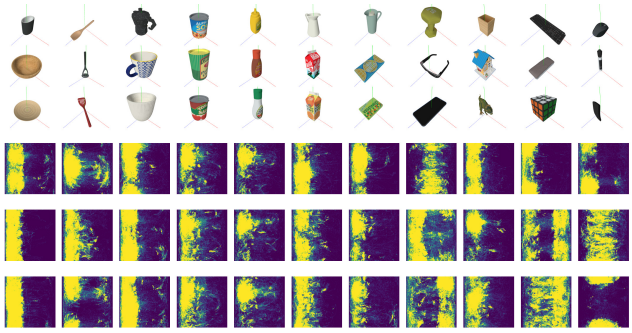

Figure 6. **Object orientation statistics.** Top: 3D object models in their canonical poses. Bottom: Distribution of azimuth and elevation angles under which the objects are observed across the dataset. The vertical axis is the azimuth angle $[0°, 360°]$ (angle along the green axis), and the horizontal axis is the elevation angle $[-90°, 90°]$ (angle w.r.t. the plane defined by the red and blue axes).

hand model [3] to the tracked optical markers, as in [2]. Object poses were estimated by aligning the tracked markers to their registered 3D locations in the model coordinate frame. To achieve reliable tracking, it was important to ensure that the marker constellation on each object is sufficiently distinct. Data frames from different sources were synchronized with SMPTE timecode.

D. Object orientation statistics

When recording, we asked subjects to naturally interact with the objects. Consequently, orientation distributions of the observed objects (Fig. 6) reveal clear object-specific pose biases, which may be useful as prior information at inference (we see that the bowl tends to be seen upright, the birdhouse from the front and upright, *etc.*).

E. Additional quantitative results

The results of 2D segmentation and 3D lifting of in-hand objects presented in Tables 4 and 5 of the main paper were obtained by evaluating methods on clips from both training and test splits. To allow the community to compare their results against our results on these two tasks, in Tables 1 and 2 we additionally provide results obtained on clips from the training split for which the ground-truth annotations are publicly available. Evaluating on the training split is possible as both of these tasks are training-free and therefore do not require any training split.

References

[1] Jakob Engel, Kiran Somasundaram, Michael Goesele, Albert Sun, Alexander Gamino, Andrew Turner, Arjang Talattof, Arnie Yuan, Bilal Souti, Brighid Meredith, Cheng Peng, Chris Sweeney, Cole Wilson, Dan Barnes, Daniel DeTone, David Caruso, Derek Valleroy, Dinesh Ginjupalli, Duncan Frost, Edward Miller, Elias Mueggler, Evgeniy Oleinik, Fan Zhang, Guruprasad Somasundaram, Gustavo Solaira, Harry Lanaras, Henry Howard-Jenkins, Huixuan Tang, Hyo Jin Kim, Jaime Rivera, Ji Luo, Jing Dong, Julian Straub, Kevin Bailey, Kevin Eckenhoff, Lingni Ma, Luis Pesqueira, Mark Schwesinger, Maurizio Monge, Nan Yang, Nick Charron, Nikhil Raina, Omkar Parkhi, Peter Borschowa, Pierre Moulon, Prince Gupta, Raul Mur-Artal, Robbie Pennington, Sachin Kulkarni, Sagar Miglani, Santosh Gondi, Saransh Solanki, Sean Diener, Shangyi Cheng, Simon Green, Steve Saarinen, Suvam Patra, Tassos Mourikis, Thomas Whelan, Tripti Singh, Vasileios Balntas, Vijay Baiyya, Wilson Dreewes, Xiaqing Pan, Yang Lou, Yipu Zhao, Yusuf Mansour, Yuyang Zou, Zhaoyang Lv, Zijian Wang, Mingfei Yan, Carl Ren, Renzo De Nardi, and Richard Newcombe. Project Aria: A new tool for egocentric multi-modal AI research, 2023. 1

[2] Shangchen Han, Beibei Liu, Robert Wang, Yuting Ye, Christopher D. Twigg, and Kenrick Kin. Online optical marker-based hand tracking with deep labels. *ACM Trans. Graph.*, 37(4):166:1–166:10, July 2018. 3

[3] Shangchen Han, Po-chen Wu, Yubo Zhang, Beibei Liu, Linguang Zhang, Zheng Wang, Weiguang Si, Peizhao Zhang, Yujun Cai, Tomas Hodan, et al. UmeTrack: Unified multi-view end-to-end hand tracking for VR. In *SIGGRAPH Asia 2022*, 2022. 3

[4] Meta. Quest 3. https://www.meta.com/quest/quest-3/, 2023. 2

[5] Nikhil Raina, Guruprasad Somasundaram, Kang Zheng, Sagar Miglani, Steve Saarinen, Jeff Meissner, Mark Schwesinger, Luis Pesqueira, Ishita Prasad, Edward Miller, Prince Gupta, Mingfei Yan, Richard Newcombe, Carl Ren, and Omkar M Parkhi. EgoBlur: Responsible innovation in Aria, 2023. 2

[6] Lingzhi Zhang, Shenghao Zhou, Simon Stent, and Jianbo Shi. Fine-grained egocentric hand-object segmentation: Dataset, model, and applications. In *ECCV*, 2022. 4

Method	Test dataset	Object in hand (mIoU ↑) for training + test / training / test split			
		Either	Left	Right	Both
EgoHOS [6]	EgoHOS	–	62.2	44.4	52.8
EgoHOS [6]	HOT3D-Aria	42.6 / 43.5 / 40.1	21.0 / 21.0 / 21.3	26.3 / 25.8 / 28.2	32.5 / 33.0 / 30.5
MRCNN	HOT3D-Aria	47.1 / 48.1 / 43.7	–	–	–
MRCNN-DA	HOT3D-Aria	**55.2 / 56.3 / 51.6**	–	–	–
EgoHOS [6]	HOT3D-Quest3	33.1 / 33.8 / 31.4	13.5 / 12.9 / 15.0	14.4 / 13.9 / 15.6	24.8 / 25.6 / 22.9
MRCNN	HOT3D-Quest3	37.8 / 38.2 / 36.9	–	–	–
MRCNN-DA	HOT3D-Quest3	**54.7 / 54.7 / 54.8**	–	–	–

Table 1. **2D segmentation of in-hand objects.** Each cell shows the mIoU score achieved on the training + test, training, and test split, respectively.

Method	Test dataset	Views	Recall [%] ↑ for training + test / training / test split			
			5 cm	10 cm	20 cm	30 cm
HandProxy	HOT3D-Aria	–	0.5 / 0.5 / 0.6	13.5 / 11.6 / 20.2	90.6 / 89.9 / 93.3	98.4 / 98.0 / 99.3
Using ground-truth 2D segmentation masks:						
MonoDepth	HOT3D-Aria	1	14.3 / 13.4 / 17.5	30.2 / 28.8 / 34.8	53.6 / 51.7 / 60.4	69.9 / 68.2 / 76.0
StereoMatch	HOT3D-Aria	3	**64.4 / 65.0 / 62.6**	**86.2 / 86.3 / 86.0**	**95.5 / 95.1 / 96.8**	**96.9 / 96.6 / 98.3**
StereoMatch	HOT3D-Quest3	2	76.4 / 78.0 / 72.8	96.8 / 96.9 / 96.5	99.1 / 99.2 / 99.1	99.2 / 99.2 / 99.1
Using 2D segmentation masks predicted by MRCNN-DA:						
MonoDepth	HOT3D-Aria	1	11.1 / 10.6 / 12.7	23.3 / 22.4 / 26.5	43.7 / 42.6 / 47.5	58.2 / 57.5 / 60.8
StereoMatch	HOT3D-Aria	3	**42.6 / 43.9 / 38.4**	**56.4/ 57.6 / 52.2**	**63.6 / 64.9 / 59.1**	**66.0 / 67.4 / 61.2**
StereoMatch	HOT3D-Quest3	2	59.1 / 60.1 / 56.9	75.3 / 75.6 / 74.6	80.4 / 80.6 / 79.9	81.3 / 81.6 / 80.7

Table 2. **3D lifting of in-hand objects.** Each cell shows the recall rate achieved on the training + test, training, and test split, respectively.