arXiv:2308.13561v3 [cs.HC] 1 Oct 2023

Project Aria: A New Tool for Egocentric Multi-Modal AI Research

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, Richard Newcombe

Meta Reality Labs Research

Abstract

Egocentric, multi-modal data as available on future augmented reality (AR) devices provides unique challenges and opportunities for machine perception. These future devices will need to be all-day wearable in a socially acceptable form-factor to support always available, context-aware and personalized AI applications. Our team at Meta Reality Labs Research built the Project Aria device: An egocentric, multi-modal data recording and streaming device with the goal to foster and accelerate research in this area. In this paper, we describe the Project Aria device hardware including its sensor configuration, the corresponding software tools, and the available machine perception functionalities that make it the ideal tool for egocentric machine perception and contextual AI research.

1. Introduction

All-day wearable AR glasses promise to be the next big paradigm shift in computing: They can lift interaction with the digital world from 2D screens into the 3D world around us; blending seamlessly into our lives as opposed to diverting our attention into small 2D rectangles held in our hands. Yet, in order to be more than mere 3D versions of 2D screens, they also require a new compute- and interaction-paradigm that is context-aware, highly personalized and natural to interact with. Creating this new paradigm is a significant challenge that still requires a broad range of research to be solved. Fortunately, the recent breakthrough of internet-trained Large Language Models such as GPT4 [15] and llama2 [7] promise to solve a major part of this challenge, making it a lot more tractable: They demonstrate that modern Transformer architectures combined with sufficient training data enable both long-range reasoning and information retrieval while also adopting seamlessly to new tasks using a local context window and prompt engineering. Taken together, this enables human-level interaction with digital agents that blend into how we humans interact naturally.

Figure 1. The Project Aria device.

1.1. Project Aria

However, we believe that this paradigm shift in personalized computing requires another crucial ingredient: the ability for AI Agents to adapt to the unique, personal context and preferences of the *individual user*. Some of this can be achieved by leveraging a persons digital footprint – the

aggregate of interactions with the internet through phones and laptops. It is only a matter of time for such personalized AI Assistants to emerge. Yet, our digital footprint only represents a small fraction of the experiences that matter to us, and rarely contain the most important ones, which play out in the real world, are highly personal and often take place in unpredictable places and at unpredictable times.

Importantly, this applies not only on an individual level, but also in aggregate: The wealth of digital data accumulated over 25 years in the digital realm only represents a small – and often severely biased – fraction of the sum total of human experience. For example, the overwhelming majority of images available on the web and used to train AI models were consciously captured with handheld devices and curated before upload. Any outtakes, any data that is not deemed interesting is typically deleted or edited although it arguably represents the majority of situations we encounter in our daily life.

A direct consequence of this is that many state of the art methods in the space of Machine Perception and AI (DALL·E2 [19], GPT-4 [15], DreamFusion [17], SAM [11], MaskRCNN [10], CLIP [18], DINOv2 [16]) excel when applied to allocentric 2D images and viewpoints, but fare comparatively poorly at tasks that involve egocentric data or require structured reasoning and understanding in 3D/4D space. While this is partially due to the increase in data/compute requirements, it is also a direct result of the aforementioned shortcomings of data typically available through web platforms.

In order to address this gap and to enable the next big paradigm shift towards context-aware, personalized, and human-oriented AI, we have created *Project Aria*: At its core, the Project Aria device is a data-capture system in glasses form factor that is sufficiently light and unobtrusive to be worn for long time-spans without inhibiting natural activities and behavior (see Figure 1), allowing to capture *ecologically valid* data. The device features a rich multimodal sensor suite that approximates what can be expected in future AR glasses for the purpose of environment- and user-understanding. The onboard battery allows the device to record 1-2 hours of data (with the nominal recording profile). Much longer recordings are possible with an external power bank.

In this technical report, we introduce the Project Aria device as well as the software and Machine Perception Services that come with it. We make these available to research institutions around the world to foster advancements in the field of egocentric perception towards personalized AI. We also summarize the principles and standards that we have established and adopted to protect the privacy of both wearers and bystanders in accordance with Meta's Responsible Innovation Principles [12].

Since its launch in 2020, Project Aria devices have been used by research groups in the USA, UK, Switzerland, India, Canada, Singapore, Colombia and Japan. In 2022, we released the Aria Pilot Dataset [2]. To find out more about Project Aria and how to become a research partner, please visit our website [4].

This report is organized as follows: Section 2 introduces the Project Aria device and its capabilities. Section 3 introduces the software and tools required for recording and using recordings for research. Section 4 enumerates the set of first-level machine perception capabilities we offer through a web service in order to accelerate and simplify use of Project Aria data. Section 5 details the privacy and responsible innovation principles we have established, and how the Project Aria device implements them. Section 6 gives a set of example research applications that are enabled by Project Aria. We end this paper with a conclusion in Section 7.

2. Device

2.1. Sensor Suite

We built the Project Aria device to emulate future wearable devices catering to machine perception rather than human consumption. To mimic the sensor stack needed for machine perception capabilities on these future devices, we integrate a rich suite of sensors that record egocentric multimodal data (see Figures 2 – 5).

Sensor streams are tightly calibrated and time-aligned to make some problems fundamentally easier to solve. This is aligned to future expectations of wearable device hardware, but introduces new challenges - such as the lack of Optical Image Stabilization (OIS) or Auto-Focus (AF).

To satisfy the strict requirements for capturing representative data in a wearable form-factor constraints, we built the Project Aria device as a data-collection and streaming device. Specifically, it is not designed to handle on-device computation workloads.

2.2. Form Factor and Fit

The Project Aria device is designed to contain a rich set of sensors while being light (around 75g) and socially acceptable. The devices provide a good fit for a broad range of the population with two device sizes. They come with adjustable nose pads and temples (i.e. the temple tips can be bent inwards or outwards to improve fit). The glasses temple have a lot more flexibility than a conventional pair of glasses, which means the small size can stretch up to what many glasses might call a medium or large size.

The sensors on the Project Aria device have been chosen to (1) approximate what we expect to be available on future all-day-wearable AR glasses, and (2) fit into the strict size, weight and power envelope required by the target form-factor to obtain ecologically valid data. We selected a broad range of sensors beyond just cameras as we believe multi-

Figure 2. Example images from the Project Aria device cameras. Left to right: left Mono Scene camera, POV (RGB) camera, right Mono Scene camera, two Eye-tracking cameras. Output of POV (RGB) and Mono Scene cameras are rotated for visualization.

modal egocentric data to be key to solve various Machine Perception and AI tasks – see Section 6 for some examples. The following gives an overview of what sensors are available on a Project Aria device (see also Figure 6):

- **Mono Scene Cameras**: Two monochrome, global-shutter cameras on the left and right side of the glasses. They have a horizontal field of view (HFOV) of 150° and are angled outwards to maximize peripheral vision while allowing for some stereo overlap. These cameras are used for supporting machine perception capabilities such as Visual SLAM [6, 13, 14]. They have a resolution of 640x480 pixels and use Fisheye (F-Theta) lenses.
- **Point of View (POV) RGB Camera**: A single high-resolution, rolling-shutter RGB camera on the left side of the glasses with a HFOV of 110°. The RGB camera also uses a F-Theta Fisheye lens, has a maximum resolution of 2880x2880 pixels and faces forward with an approximately 4° bias towards the ground.
- **Eye Tracking Cameras**: Two monochrome, global-shutter, inward-facing cameras for eye tracking with a diagonal field of view (DFOV) of 80°. They typically operate at a 320x240 pixels resolution.
- **IMUs**: Two inertial measurement units (IMU), one on each side of the glasses. The left IMU is configured to sample at 800 Hz with saturation limits of 4g (accelerometer) and 500°/s (gyroscope). The right IMU samples at 1000 Hz with saturation limits of 8g and 1000°/s. We intentionally chose different IMU models so that their higher-order error behaviors are more likely to be uncorrelated.
- **Microphones**: A microphone array comprised of 7 microphones distributed around the glasses (5 front, 1 on each side), allowing to capture spatial audio at 24 bits with a configurable sample rate of up to 48 kHz.

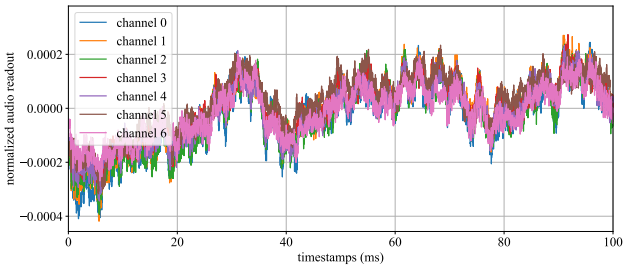

Figure 3. Example time-series data from the multi-channel microphone array on the Project Aria device. Audio data is saved in 32-bit format and normalized in the range of $[-1, 1]$.

- **Magnetometer**: A magnetometer located on the rim of the glasses to minimize electromagnetic interference, measuring the ambient magnetic field (3-axis) with a resolution of 0.1 μT and a sample rate of 10 Hz.
- **Barometer & Thermometer**: A barometer sensor capturing local air pressure and temperature at a resolution of 0.66 Pa and 0.005° C, respectively, with a sample rate of 50 Hz.
- **GNSS receiver**: A global navigation satellite system receiver supporting GPS and Galileo constellations, providing pseudo-range measurements as well as lat/long/height solutions with a sample rate of 1 Hz.
- **Wi-Fi & Bluetooth transceiver**: A Wi-Fi and Bluetooth radio with the ability to regularly scan and record received signal strengths (RSSI) from Wi-Fi beacon frames (both 2.4G and 5G) and from Bluetooth beacons. Scanning is nominally performed at 0.1 Hz.

Many of these sensor settings can be configured through *recording profiles* that are selected at the start of a recording and allow to modify camera frame-rate and resolution, as well as to enable or disable different sensor streams. This

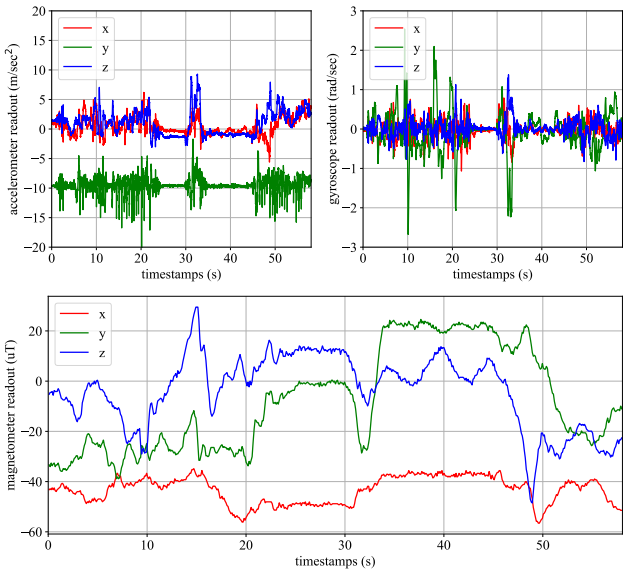

Figure 4. Example data from various motion sensor and location signal data. Top to bottom: accelerometer and gyroscope data provided by the IMUs, and magnetic field measurements provided by the magnetometer.

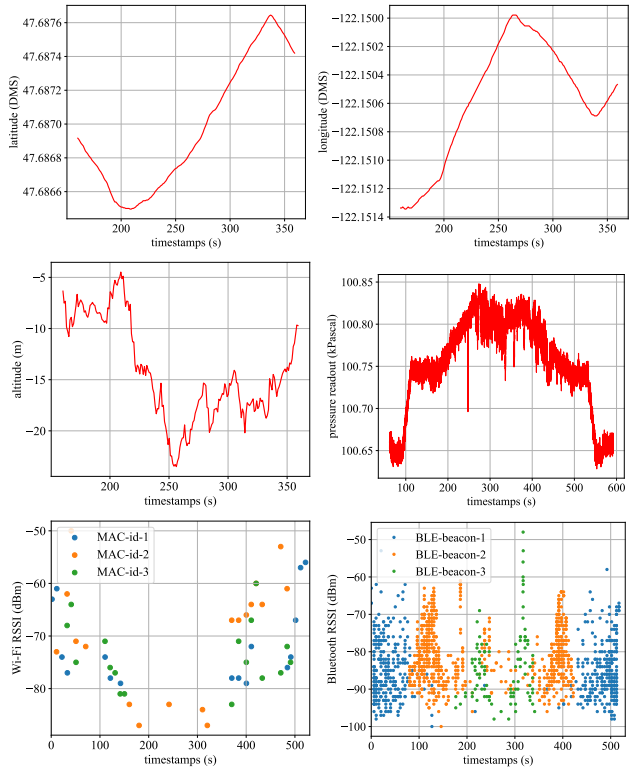

Figure 5. Illustration of the GNSS and Wi-Fi/Bluetooth sensor data. Top to bottom: GNSS signal (individual plots of latitude, longitude, altitude), pressure measurements provided by the barometer, signal strengths from different sources as recorded by the Wi-Fi and Bluetooth receivers.

is important as recording all sensors at maximum resolution and rate is not feasible due to power and bandwidth limitations. Furthermore, disabling sensor streams or reducing their rate and/or resolution can be desirable from a privacy perspective, and is an effective strategy to prolong maximum recording time. Available recording profiles are listed in the Project Aria documentation site [1].

2.3. Mounting and Rigidity

Precise 6DoF alignment – that is, relative positioning and orientation between all sensors – is important for many basic machine perception algorithms, and the Project Aria device has been designed to facilitate this. All sensors are mounted onto a magnesium frame spanning the front of the glasses. With the most non-rigid portion being the nose-bridge, there are two primary sensor clusters on the left and right side of the glasses. Sensors within the same cluster have a strong rigid connection.

We provide extrinsic and intrinsic calibration parameters for each sensor computed at manufacturing time (factory calibration). Through our Machine Perception Services (MPS, see Section 4) we additionally make more accurate online-calibration parameters available that account for the small deformations/changes that might occur while wearing the glasses. Please refer to [1] for more details.

2.4. Time and Time-alignment

The Project Aria device is designed to allow accurate timestamping of all sensor data with respect to a local on-device time source. This is essential to combine different modalities in downstream machine perception tasks.

In addition, Project Aria devices provide the ability to align and convert local timestamps to a time domain shared across multiple devices. Accurately translating to a common time domain is critical when combining or comparing data from different sources and devices. In our sample datasets, we use SMPTE LTC timecode [20] to provide a common accurate time domain across multiple Project Aria devices (see Aria Pilot Dataset [2]). For situations where lower accuracy is acceptable we have also implemented a methodology for sharing a common time domain over Wi-Fi leveraging the TicSync timing protocol [9]. For both methodologies, the inner working of the time sharing mechanism is handled at the device level. This means, from the perspective of a user of the data, every sensor reading simply includes a timestamp in the aligned time domain in addition to a timestamp in the local device time. Please refer to the Project Aria documentation [1] for more details on the exact definition and conventions for timestamps of the different sensor modalities.

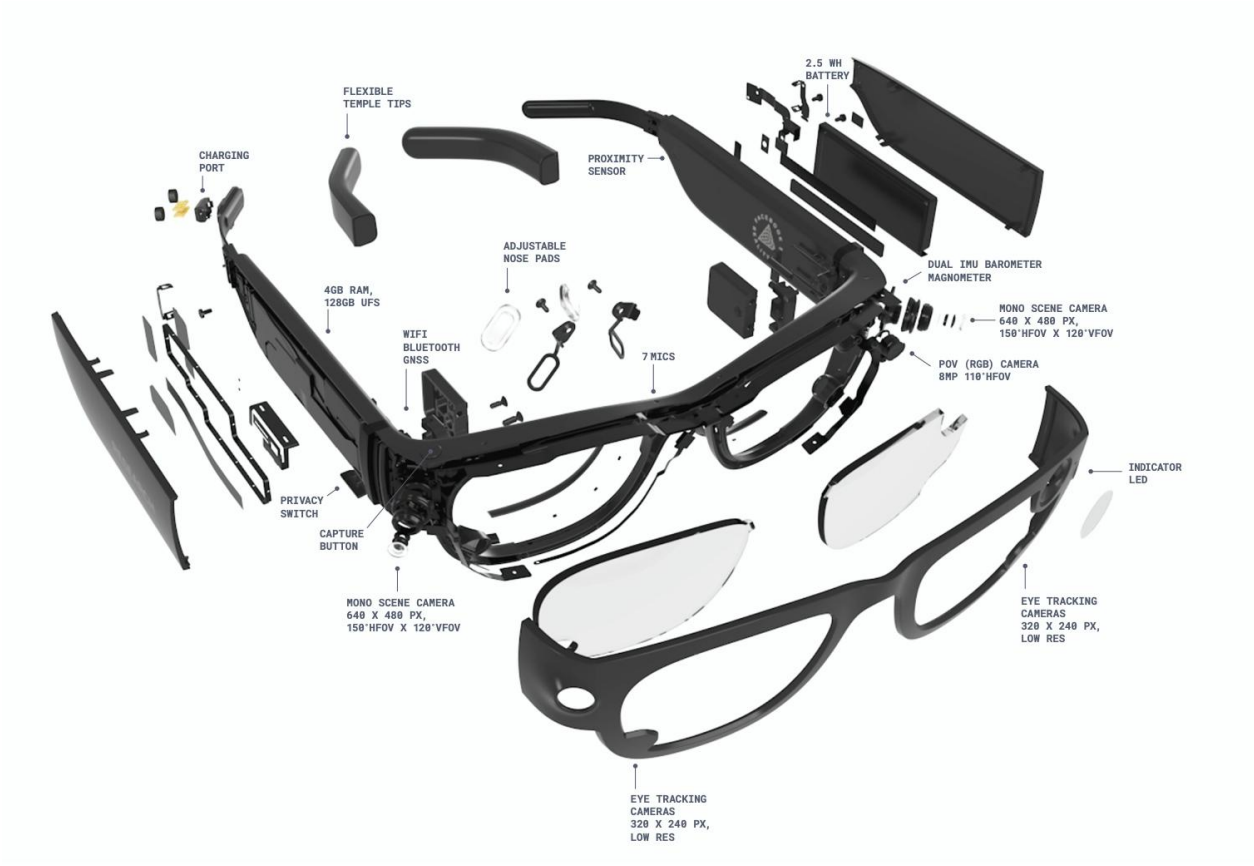

Figure 6. Project Aria device hardware overview of the components, the various sensors, switches, LEDs, battery, etc.

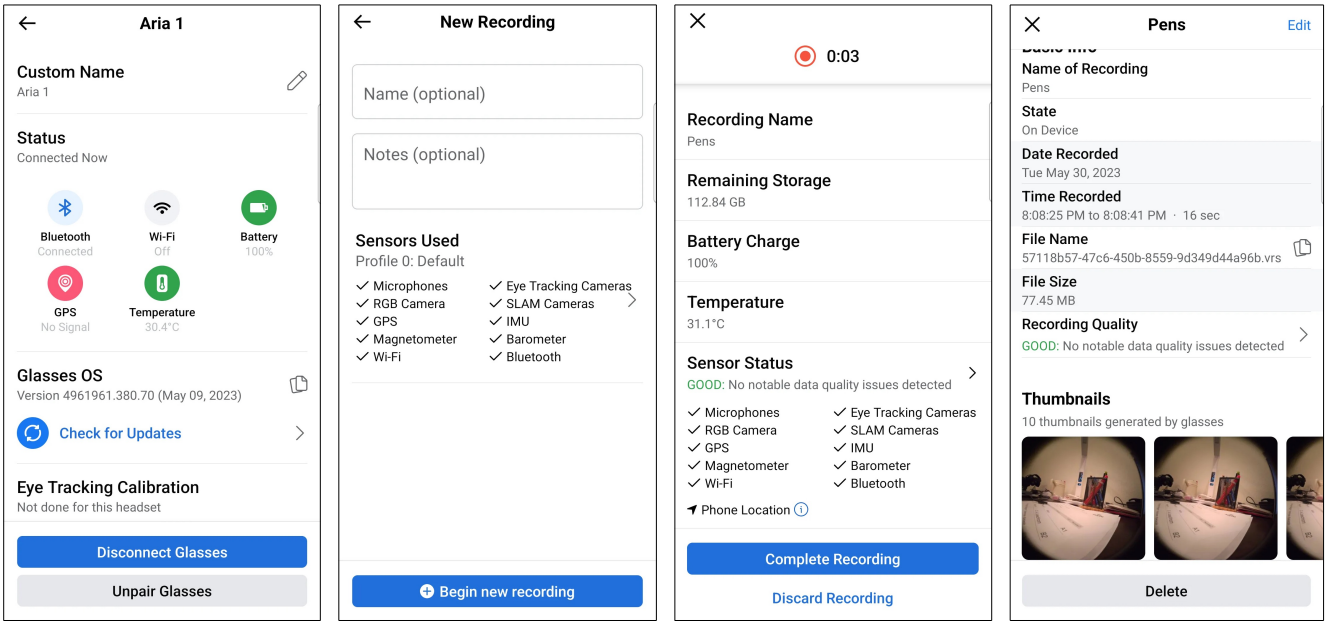

Figure 7. Basic functionalities of the mobile phone companion app. From left to right, the screenshots illustrate a) the device status information, b) the recording profile that configures the sensors before recording, c) status of the on-going recording, d) thumbnail preview of the recordings visible after a recording finishes

3. Recording Tools

The primary interface to interact with Project Aria devices is a mobile phone companion app. Recordings can be initiated and stopped via the app or the device's capture button. The app is also used to set the sensor configuration by selecting a recording profile. The profile controls which sensors record, at what frequency and resolution, as well as the output format (e.g., storing images in RAW or JPEG format). Multiple recording profiles are available to tailor for different research use cases. Additional profiles are being added as necessary. Once a recording is made on the device, thumbnail previews are available on the companion app for convenient review of the captured data. These functionalities are illustrated in Figure 7.

Once the recording is complete, the user can download the recorded data from a Project Aria device via an USB connection to a local machine for further processing. A user can optionally upload their recordings to our Machine Perception Services (MPS), which apply state-of-the-art processing to recover device trajectories, online calibration, a semi-dense point cloud and eye gaze information (see Section 4 for details).

All device sensors are recorded in VRS file format [22]. We selected VRS as the data container because it is an open file format designed to record and playback streams of AR sensor data and because it supports very large file sizes. The VRS files contain streams of time-sorted records generated for each sensor, with one set of sensors per stream.

In order to easily visualize and interact with data, we provide Project Aria tools as part of an open source repository [3]. This is a set of tools and libraries for accessing, visualizing and manipulating recordings from Aria. The C++/Python toolkit includes VRS Data Provider and Viewer interfaces, enabling researchers to read and visualize Project Aria sequences, to retrieve and interact with device calibration data, and to read and process the output of the MPS. More details about these tools are available from documentation website [1]. The Project Aria tools are available to install from PyPI via `pip install projectaria_tools` (see [3] for more details).

4. Machine Perception Services

We provide a range of foundational machine perception capabilities upon which research partners can build their projects. We expect these capabilities to be provided in a similar form on any future AR device.

These capabilities are exposed as Machine Perception Services (MPS) enabled by a set of proprietary algorithms that are designed for Project Aria devices and provide superior accuracy and robustness on the recorded data compared to current off-the-shelf open source solutions.

MPS is provided by post-processing VRS recordings on Meta's backend servers. To use the service, Project Aria research partners upload the recording, and, later on, can download the results. Furthermore, we include MPS output in public Aria-based datasets making it available to the broader community.

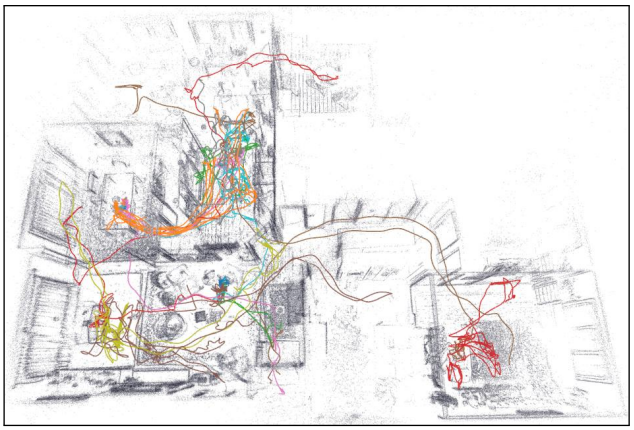

Figure 8. Closed-loop trajectories and semi-dense point clouds of 18 recordings of Aria Pilot Dataset [2] collected in the same home space.

Please refer to the Project Aria documentation site [1] for more information about Aria MPS, the output format and their specifications, and an overview over the tooling available to visualize and make use of the data.

4.1. Trajectories

A highly accurate 6-DoF device trajectory is the foundation to understand the geometric relation of the device and its wearer to the environment. Device trajectories are generated by a state-of-the-art VIO and SLAM system – similar to what can be expected on future AR and VR HMD's – followed by offline post-processing and refinement. We use multiple of the available sensors (including cameras, IMUs, GNSS, Wi-Fi, and barometer) to improve accuracy and robustness, and further take advantage of precise knowledge of the sensor models, timing, and rigidity of Project Aria devices. This allows us to robustly localize the device even under the often challenging conditions that occur with real-world data - such as fast motion, low or highly dynamic lighting, partial or temporary occlusion of the cameras, as well as a wide range of static and dynamic environments.

We provide two types of trajectories as output, open loop and closed loop. The *open loop trajectory* is a high frequency (1 kHz) odometry estimation, computed strictly causally with a real-time-compatible method. The accumulated translation drift of this open loop trajectory is no more than 0.4% of the distance traveled, and usually significantly less. The *closed loop trajectory* is a 1 kHz trajectory estimated in post-processing. It is fully optimized and provides poses in a single frame of reference. We also provide

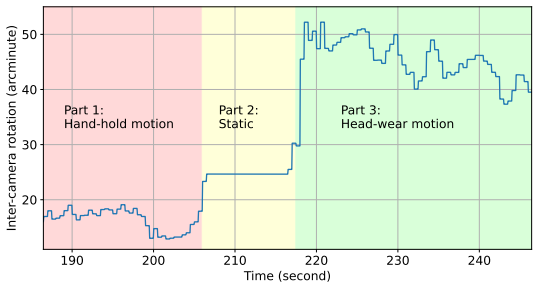

Figure 9. Example of rotational deformation between left and right Mono Scene cameras, estimated by MPS. The recording used for this figure contains 3 sections: hand-held motion, no motion, and head-worn motion. The effect of external force applied to the glasses when worn on the head is clearly visible.

the ability to jointly process multiple recordings, placing them into a common coordinate frame, as shown in Figure 8. Closed loop trajectories have a typical global RMSE translation error of no more than 1.5 cm in room-scale scenarios.

4.2. Online Calibration

Accurate device calibration is essential to enable high geometric accuracy for downstream 3D perception tasks. Even though Project Aria devices are built to be as rigid as possible, device calibration parameters are not perfectly constant over time due to temperature changes, aging, and external forces applied to the device. Figure 9 shows an example where taking off the device after wearing causes around 25 arcmin instantaneous rotational deformation between the left and right Mono Scene cameras (corresponding to roughly 1.5 pixel shift in the images). To account for this, we estimate the time-varying intrinsic and extrinsic calibrations of cameras and IMUs as part of MPS, and make the result available to researchers.

4.3. Semi-Dense Point Cloud

To provide an intuitive understanding of the environment a recording was taken in, we compute semi-dense tracks and point clouds as part of MPS; see Figure 8 for an example. Similar to the odometry trajectory, these tracks are computed causally and provide an accurate – though partial – reconstruction of the static portion of the environment. We also provide the sets of all 2D observations that were used to triangulate each 3D point. Tracks are obtained by continuously spawning new points in images-regions with high gradient, and tracking these over time and across the left/right Mono Scene camera using affine-invariant photo-consistency of local patches. Finally, the 3D point clouds are post-processed and placed into the global frame of reference that is defined by the closed loop trajectories.

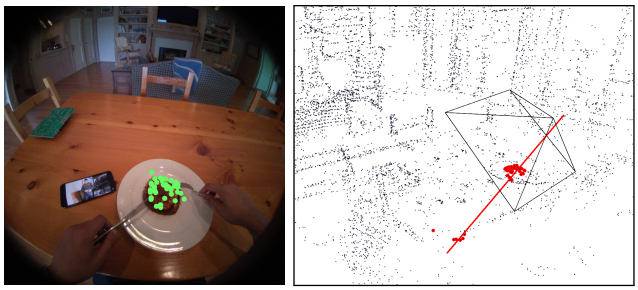

Figure 10. Eye gaze computed on a recording in which the user is looking at the food on their plate. Left: The RGB image with semi-dense points that are close to the gaze ray projected as green dots. Right: The Project Aria device pose (shown as RGB camera frustum), semi-dense points, and eye gaze ray. Semi-dense points close to the gaze ray are highlighted in red.

4.4. Eye Gaze Tracking

Gaze direction is an important indicator of a wearer's attention, and will likely be one of the crucial inputs to context-aware, personalized AI agents. We compute and provide eye gaze from the Project Aria device eye tracking cameras, estimating a single per-frame 3D ray anchored to the central pupil frame, also called a cyclopean eye frame [1]. We also provide confidence intervals, as eye tracking accuracy can vary by situation and user.

Furthermore, the Aria companion app described in Section 3 implements the option to capture personalized eye calibration - this allows to improve the accuracy of eye gaze tracking by compensating for user-specific biases. With the current model, we observe a median gaze ray error of 1.5° after applying the personalized calibration.

5. Privacy Considerations

AR glasses and in general egocentric recording devices such as Project Aria devices promise to make technology more accessible, but also pose novel and unique challenges for security and privacy: the more they succeed in being unobtrusive, the more important it becomes to preserve and respect the privacy not only of the wearer, but also that of bystanders and individuals the wearer interacts with.

A stated goal of Project Aria is to pioneer responsible innovation for research leveraging egocentric data and devices, both by establishing guidelines and principles to preserve privacy of wearers and bystanders as well as by building privacy-facilitating features directly into the device where possible.

Throughout the development of Project Aria we followed Meta's Responsible Innovation Principles [12], which express our commitment to building inclusive,

[1] The central pupil frame origin is defined at the midpoint between the left and right pupils.

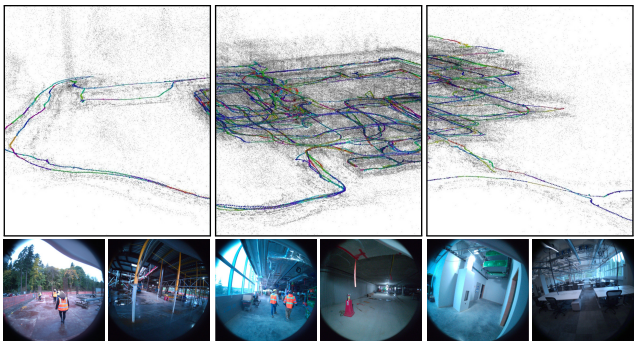

Figure 11. Top: A 3D map created from 275 Aria recordings (175 hours of data), captured over 15 months in a construction site, showing – from left to right – the state of the map after one, six, and fifteen months. The images at the bottom are samples from the respective points in time, depicting the transformation from an empty lot to an office building.

privacy-centric products.

In concert with our principles we have designed and made available privacy-centric hardware and software features to research partners. The Project Aria device has an LED indicator that signals to bystanders when the device is recording raw data. Furthermore, Project Aria devices have a privacy switch: When activated during a recording session, the device immediately stops and deletes the current recording. This allows a wearer to immediately and easily fulfill a request of a bystander to delete any audio or video recording that might have been captured of them.

We require all Project Aria partners to follow the Project Aria community guidance [5] and to practice responsible research, protecting the privacy of those who wear the research devices, and most importantly, those who do not.

6. Example Research Applications

This section provides a brief overview of research tasks that leverage Project Aria's unique features, from low-level machine perception functions to high-level user- and environment-understanding. Project Aria is designed to enable and connect research across this spectrum: While the former benefits from well-calibrated and understood sensors and access to raw data, the latter can leverage the multiple modalities available or build upon the machine perception functions provided by Aria Machine Perception Services (MPS). Note that this is neither an exhaustive review of the respective fields nor a complete list of tasks that are enabled by Project Aria. It is meant to provide examples how Project Aria devices or Aria data can be leveraged, with an emphasis on Aria's unique combination of form-factor, sensors and Machine Perception Services.

Figure 12. Two NeRF reconstructions obtained from Aria recordings using NerfStudio [21]. The left shows the result from a carefully curated, hand-held recording that covers the space well and avoids rapid motion. The right shows the result from an egocentric recording during a natural activity. The bottom figures visualize the raw pointcloud and trajectory from MPS. The resulting difference in quality is clearly visible.

6.1. Life-long Mapping and Re-localization

Precise 6DoF localization through SLAM or SfM is a common first step across many applications, as well as a base requirement for AR/VR world-locked rendering. It is a comparatively mature field, and Project Aria provides metric 6DoF trajectories as part of Aria MPS (see Sec. 4.1). However, many challenges remain: One of these is to reliably re-localize across strong environment changes that occur in natural environments (typically due to lighting, weather, or human activity), as well as updating maps with such changes over time. Figure 11 shows a map created from 275 Aria recordings, captured over 15 months, in a building that's being built: Using Aria as convenient recording device and building upon the per-recording trajectories from Aria MPS allows to focus on the core problem of long-range re-localization and map-updating under such strong environmental changes.

6.2. Egocentric Scene Reconstruction and Understanding

Reconstructing the surrounding scene and semantically identifying objects in it is a key problem for AR/VR applications – from creating photo-realistic, virtual memories to identifying affordances of objects in the surroundings as part of a context-aware AI assistant. This becomes particularly challenging when the input data is not neatly curated or intentionally taken for this purpose, but rather stems from a form-factor and power-constrained wearable device undergoing natural, unconstrained human motion.

Figure 12 shows the results obtained by state-of-the-art methods for NeRF reconstruction [21] on Aria data, comparing careful "scanning" motion with a natural activity.

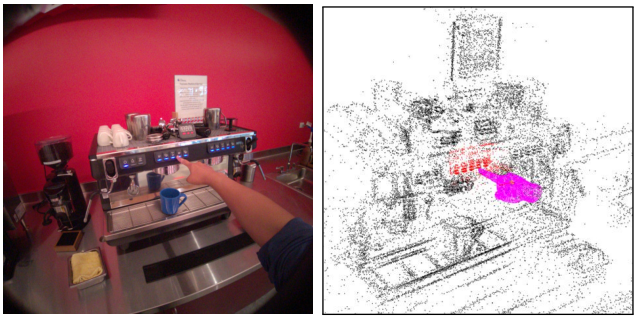

Figure 13. We use UmeTrack [8] to track the articulated 3D hand-pose of the wearer in an Aria recording. Combined with the point-cloud from MPS, this allows to identify when the wearer touches a static object. The figures visualize this by coloring points within 8 cm of the hand in red.

6.3. Object Interaction and Manipulation

Recognizing or tracking objects the user is interacting with, or identifying how the user is interacting with them, is another core egocentric machine perception task. It combines hand-tracking with object tracking, recognition, or scene understanding to connect *things* with user intent or actions. Figure 13 shows an example application that identifies when the user's hand is near an object in the scene. The approach uses all 3 cameras as well as the trajectory and pointcloud provided by MPS – which can help in particular to resolve the otherwise common scale/depth ambiguity. Figure 10 showed a similar approach using eye gaze to identify what the wearer is looking at.

6.4. Activity Recognition and Attention

Identifying what the wearer is doing or paying attention to is another likely component of any contextual AI assistant. While much information can be derived from egocentric images or videos alone, significant additional signal can be derived from other modalities, including spatial audio, motion, or eye gaze. Figure 14 shows two example situations where these additional signals allow to disambiguate what the wearer is doing in otherwise ambiguous egocentric views.

6.5. Summarization and Question Answering

Summarization and Question Answering goes a step further than activity recognition, aiming to summarize relevant events and activities that occur over longer time periods and allowing to answer questions about them. Note that "relevance" in this context is highly subjective and personal, making signals such as eye gaze or spatial audio key to select the most relevant information for the user.

Furthermore, "longer time periods" can vary from a few minutes to hours, days, and years – with longer time-spans becoming increasingly important towards personalized AI assistants, but requiring datasets that do not currently exist. We believe that Project Aria's unique combination of form-factor and machine perception capabilities enables more research in this field towards egocentric, longitudinal summarization and question answering.

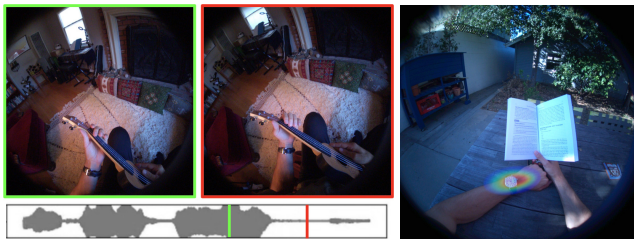

Figure 14. Left: two egocentric views of the wearer interacting with a guitar. The audio stream visualized below allows to disambiguate whether the wearer is actively playing or just holding the guitar. Right: the eye gaze (visualized as a heatmap) allows to distinguish whether the wearer is looking at the time or reading a book.

7. Conclusion

With Project Aria, we introduce a new tool for the research community that can capture ecologically valid data as we expect it to be captured by future egocentric devices. We also make available a set of spatial AI machine perception technologies as a foundational building block for higher-level contextualized AI applications. The data, and derived machine perception results, can provide the foundation for building novel compute and interaction paradigms needed in order to make AR successful and the rich integrated sensor suite provides unique opportunities to explore novel research, applications and use cases in a wide range of areas towards always-on contextualized AI.

Acknowledgements

Project Aria was made possible by the contributions of the Project Aria team from Meta Reality Labs Research. We are indebted to the complete team and all partners of Project Aria who enabled its inception and continue to develop the platform.

References

[1] Project Aria Documentation. https://facebookresearch.github.io/projectaria_tools/. 4, 6

[2] Project Aria Pilot Dataset. https://facebookresearch.github.io/projectaria_tools/docs/open_datasets/pilot_dataset. 2, 4, 6

[3] Project Aria Tools on GitHub. https://github.com/facebookresearch/projectaria_tools. 6

[4] Project Aria Website. https://www.projectaria.com/. 2

[5] Project Aria Community Guidelines. https://about.meta.com/realitylabs/projectaria/community-guidelines/. 8

[6] Jakob Engel, Thomas Schöps, and Daniel Cremers. LSD-SLAM: Large-scale direct monocular SLAM. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part II 13*, pages 834–849. Springer, 2014. 3

[7] Hugo Touvron et.al. Llama 2: Open foundation and fine-tuned chat models, 2023. 1

[8] 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 *SIGGRAPH Asia 2022 Conference Papers, SA 2022, Daegu, Republic of Korea, December 6-9, 2022*, 2022. 9

[9] Alastair Harrison and Paul Newman. TICSync: Knowing when things happened. In *2011 IEEE International Conference on Robotics and Automation*, pages 356–363, 2011. 4

[10] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2961–2969, 2017. 2

[11] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 2

[12] Meta Responsible Innovation Principles. https://about.meta.com/metaverse/responsible-innovation/. 2, 7

[13] Anastasios I Mourikis and Stergios I Roumeliotis. A multi-state constraint Kalman filter for vision-aided inertial navigation. In *Proceedings 2007 IEEE International Conference on Robotics and Automation*, pages 3565–3572. IEEE, 2007. 3

[14] Raul Mur-Artal and Juan D Tardós. ORB-SLAM2: An open-source slam system for monocular, stereo, and RBG-D cameras. *IEEE Transactions on Robotics*, 33(5):1255–1262, 2017. 3

[15] OpenAI. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 1, 2

[16] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 2

[17] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. DreamFusion: Text-to-3D using 2D diffusion. *arXiv*, 2022. 2

[18] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 2

[19] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 2

[20] Linear Timecode. https://en.wikipedia.org/wiki/Linear_timecode. 4

[21] Matthew Tancik, Ethan Weber, Evonne Ng, Ruilong Li, Brent Yi, Justin Kerr, Terrance Wang, Alexander Kristoffersen, Jake Austin, Kamyar Salahi, Abhik Ahuja, David McAllister, and Angjoo Kanazawa. Nerfstudio: A modular framework for neural radiance field development. In *ACM SIGGRAPH 2023 Conference Proceedings*, SIGGRAPH '23, 2023. 8

[22] VRS Documentation. https://facebookresearch.github.io/vrs/docs/Overview. 6