On the Application of Egocentric Computer Vision to Industrial Inspection

Vivek Chavan*[1,2], Oliver Heimann[1], and Jörg Krüger[1,2]

[1] Fraunhofer IPK
[2] Technical University of Berlin, Germany

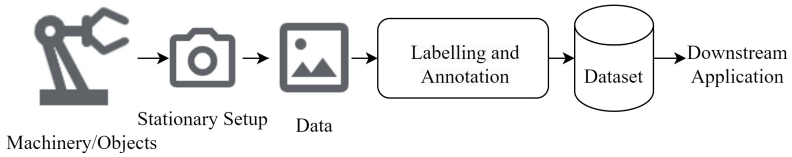
Abstract. Conventional Industrial Inspection is often carried out with dedicated digitisation and monitoring systems. This approach offers superior accuracy and reliability. However, it poses limitations w.r.t. setup costs and location requirements. Recently, Egocentric Vision has been getting increased attention from the community. We investigate the applicability of egocentric wearable devices for data collection and labelling for classification, detection, segmentation, and defect/anomaly detection use cases. We also explore scenarios where egocentric vision would be superior for data collection and inspection. Our approach involves a multimodal data collection pipeline, where the Subject-Matter Expert (SME) labels and annotates the data in natural language during digitisation, which is then processed to yield an annotated dataset for downstream applications. We also incorporate useful indicators such as user eye gaze (via wearable glasses) and hand tracking to understand and annotate the regions of interest in the collected data. Our investigation shows a domain gap when generalising the performance of Machine Learning (ML) models trained on egocentric data to allocentric/exocentric use cases. Further, we discuss the limitations and practical use cases of our approach, considering the hardware and power consumption requirements. Our code and collected data are available at: https://github.com/Vivek9Chavan/EgoVis24

Keywords: Egocentric Vision · Industrial Application · Data Collection · Multimodality

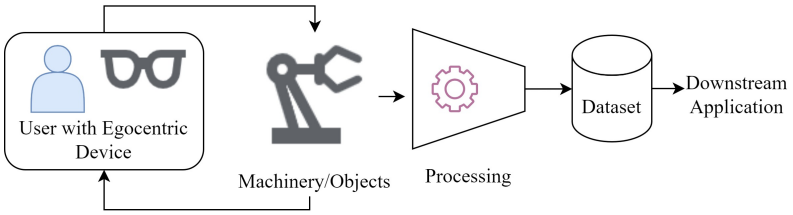
1 Introduction

Recent advancements in Computer Vision (CV) [4], Deep Learning (DL) [32] and related fields have translated to several real-world use cases [48]. Simultaneously, Industry 4.0 and digital transformation have led to data-driven improvements in production, manufacturing and industrial applications [37]. A key focus here is on efficiency improvement in manufacturing and production. Several CV

*Correspondence: vivek.chavan@ipk.fraunhofer.de



(a) Conventional data collection and labelling approach in industrial inspection scenarios, involving a fixed, stationary setup.



(b) Proposed approach for automated data collection and annotation using an egocentric wearable device, where the user describes their observation while interacting with the object. The data is then processed to obtain the labelled dataset.

Fig. 1: A comparison of the two approaches for data collection and processing for industrial inspection. Our work explores the latter case.

and DL approaches have been successfully implemented in the domain of industrial inspection, viz. classification, object detection and tracking, anomaly and defect detection, and robotics, among others. Industry 5.0 is projected to be the next big leap in industrial automation and technological advancement [15]. The focus would be on collaboration between humans and machines, to combine advanced technologies with human skills and creativity.

DL-based methods require large amounts of labelled data to train effectively. This is often a barrier to digital transformation and the adoption of Artificial Intelligence (AI). There have been algorithmic breakthroughs to reduce the amount of application-specific labelled data, such as pretrained backbones trained using self-supervised learning (SSL) [40, 49] and unsupervised anomaly detection [5]. However, in several other cases, data collection and labelling remain a bottleneck in real-world scenarios. Currently, dedicated digitisation workstations and stationary setups are required for collecting data and during real-world deployment (inference), as shown in Fig. 1a.

Egocentric Computer Vision has emerged as a prominent area of research, due to the commercial adoption of wearable devices such as the GoPro, Apple Vision Pro, Microsoft Hololens and Meta Ray Ban Glasses. A dominant feature of smart wearable devices in the future is expected to be multimodality and contextualisation.

In this paper, we envision varying scenarios in which egocentric wearable devices could be used for industrial scenarios and CV-based inspection tasks. We estimate that this opens up several new possibilities for development, but also introduces novel challenges and research gaps. We explore the use of egocentric vision technology to bridge the gap between SME know-how, current operating

procedures and the jump to digitisation/AI-assisted inspection. As shown in Fig. 1b, we explore the use of the current state-of-the-art (SOTA) methods and human expertise to interact with the machinery or objects for data collection and inspection. We also discuss new industrial inspection scenarios that are enabled by this approach.

2 Related Work

Industrial Inspection. This is a broad area of research and includes several application-specific niches [35]. Conventional applications such as Colour Inspection, Quality Control, Dimensional Measurement, Label and Packaging Inspection have been actively researched. DL approaches involve anomaly and defect detection, object identification and tracking, segmentation, and Optical Character Recognition (OCR), among others [35, 47, 48]. Industry-specific datasets have been published to drive the fields of DL/AI-based analysis and inspection forward [2, 29, 46]. Addressing a novel application scenario in industrial settings often requires substantial data collection and labelling. This, in turn, often requires custom inspection and digitisation hardware [21, 37].

Computer Vision and Deep Learning. Computer Vision tasks and benchmarks are heavily responsible for the significant breakthroughs in DL and AI [18, 30, 31]. In large part, this has been enabled by the availability of large amounts of data and computational resources [3, 9, 45]. The paradigm of foundational models that can be used for various downstream applications has been transforming various subdomains while requiring a low computational overhead [28, 38]. In recent years, the community has been actively focusing on 3D, Multiview and Multimodal applications of Computer Vision, with the goal of understanding and modelling the world efficaciously [11, 23, 26]. Particularly, language has emerged as a strong supplemental modality that yields generalisable and robust features [41, 52]. Developments in Natural Language Processing (NLP) have also been driving recent advancements in Vision, with architectures such as the Transformer [51] being successfully applied to Vision tasks [10, 34].

Egocentric Computer Vision. Smart wearables and egocentric devices are estimated to be the next computing paradigm after the desktop and mobile platform developments [1, 12, 16]. As a result, egocentric vision research has been getting broader recognition. The field has its foundations in human-centric understanding and computing [33, 36], which has also been accelerated due to the recent availability of large open-source datasets and benchmarks [8, 19, 20]. The data is often collected by the user wearing a portable camera device (e.g. head or chest-mounted GoPro) and performing various activities. The emphasis is often on procedural activities and video data. A significant portion of the published datasets focus on everyday activities and outdoor tasks [8, 19]. Some industry-specific tasks such as assembly-disassembly, goal step understanding and instance tracking have also been proposed [20, 46].

In this paper, we aim to explore the use of egocentric devices and related technologies as an aid to conventional industrial inspection setups. The goal is to

take the human-centric approach from egocentrism to unlock SME knowledge for AI-driven inspection. We investigate the use of egocentric devices for collecting and annotating data, and inference.

3 Methods

3.1 Egocentric Vision Setup

The simplest way to adopt egocentric devices in the industry would be using smartphones. They are easy to use, have good-quality cameras, and allow additional modalities such as audio, motion, and depth to be synchronously recorded. However, the device must be turned on and pointed at the object of interest to get the data. It can't be used hands-free without special equipment. Lightweight devices such as the GoPro can be easily mounted on the head or the chest and have been used for collecting egocentric datasets [33]. Another option we explored was AR/VR headsets, viz, Microsoft HoloLens and the Meta Quest. These devices allow good quality multimodal (video, audio, depth, motion) data to be captured and stored on the device. However, these devices are often heavy and uncomfortable to wear beyond a certain timeframe (in our estimate, 5-15 minutes when working on inspection tasks). Recently, the Project Aria device was proposed as a research tool for egocentric vision tasks [12]. The device mimics the sensor stack expected on future smart eyewear devices and offers additional resources via its open-source platform. The device also contains dedicated camera sensors for eye tracking and SLAM, which provide significant advantages over devices such as the GoPro. We use the Project Aria device and the related Machine Perception Services (MPS) for our study.

3.2 Data Collection and Annotation.

Language and Jargon. SMEs with extensive technical knowledge can explain their workflow and reasoning in natural language much more easily. This may involve occasional jargon and technical terms not known outside the organisation. To enable this use-case, we use Whisper from OpenAI [42] to transcribe the user's speech. The output is forwarded to Llama-3, the Large Language Model (LLM) from Meta AI [50] along with a small technical database for added context. This includes a small list of common objects and defect cases (e.g. discolouration, crack, dent, scratch, deformation), along with some regional language terms that the LLM may have trouble contextualising. Small databases may be forwarded to the LLM via a system-level prompt; larger databases can be managed via Retrieval Augmented Generation (RAG).

Pipeline. Fig. 2 shows an overview of the automated data collection process for industrial tasks. The data collected by the device is stored and used on the device for further use. The collected video stream can be optionally sampled to obtain an image dataset. We implement constant sampling and selection via a pretrained feature extractor [6]. The LLM processes the audio/speech data to

yield a structured data format (CSV or JSON). The labelled and processed data can be used for downstream analysis and training ML models. The SLAM camera sensors and the MPS enable the user to construct semi-dense 3D point clouds of the working environment. The Inertia Measurement Unit (IMU) sensors provide data on user movement, which can be projected into the 3D environment to capture a dynamic understanding of user interactions. Fig. 2 shows an example. The eye-tracking cameras provide data on the user's eye gaze as a directional vector and depth estimation via a custom-trained model.

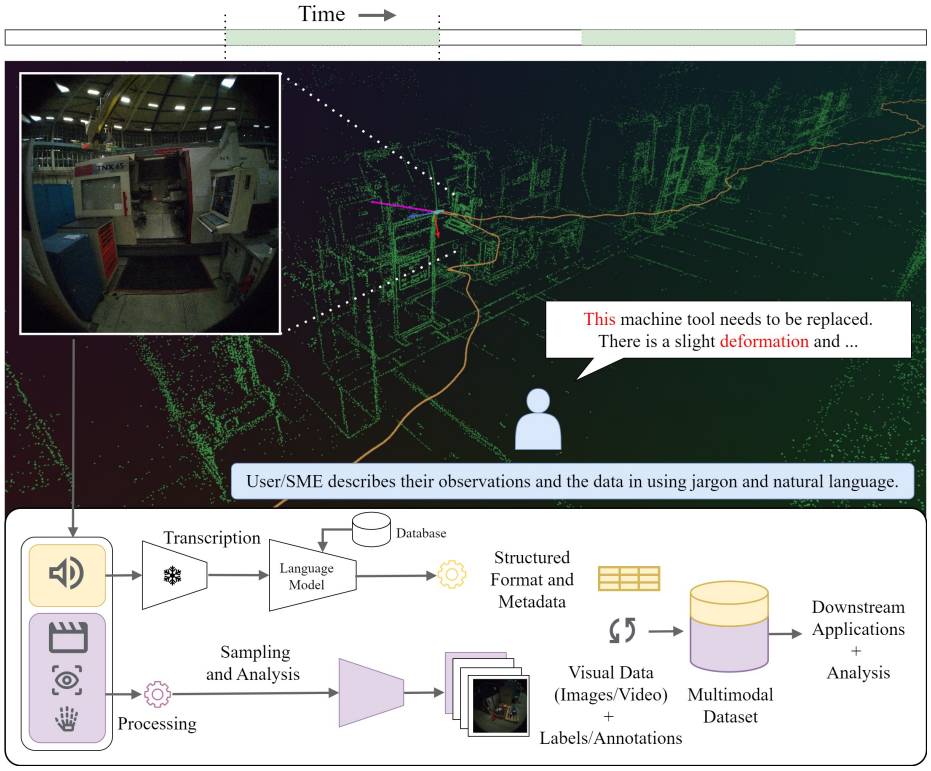


Fig. 2: A summary of the proposed pipeline. The User/SME wearing the egocentric device interacts with the object/machinery and documents their observation in natural language. The multimodal dataset is then processed to obtain image/video data, and the transcription, eye-gaze, and hand interaction provide the labels and annotations, along with metadata. **Top:** Point cloud reconstruction (green) example from a use case, along with the projected user trajectory (orange), and an RGB image of the user's Point of View. **Bottom:** A conceptualisation of the data processing method.

3.3 Inference and Testing

The Aria device does not have a direct output interface and does not allow on-device computation. The Software Development Kit (SDK) [17] enables the user to interface the stream from the device to be used in real-time for processing on a separate workstation. The connection may function via a wired USB connection or wireless over Wi-Fi. For this study, we use recorded data from the device for inference and testing, i.e. no real-time operations are performed. However, we estimate this does not impact the evaluation w.r.t. inspection tasks for industrial scenarios.

3.4 Anonymisation and Data Privacy

Egocentric devices offer conveniences such as always-on data capture and user-centric scene understanding. However, these features also pose challenges w.r.t. protecting sensitive business data (Intellectual Property, trade secrets etc.). User-specific metadata such as eye gaze and hand pose may be anonymised in certain scenarios. In other cases, it might be appropriate to identify the user for developing specific protocols and customising feedback. Additionally, the use of egocentric devices and wearable cameras often results in the recording of other humans and their private information without prior consent. We use the EgoBlur model [43] to anonymise faces and other private information in the collected data before use.

4 Experiments

Setup Details. For our study, we selected 12 participants with varying levels of expertise (1 SME, with over 2 years of experience with the traditional workflow and industrial inspection, 2 junior experts, with approx. 6 months of experience, and 9 interns with limited experience). The group consists of 11 males and 1 female participant, with 6 different nationalities and ethnic backgrounds, to obtain an impartial understanding of user experience. The data collected by the participants was processed on a dedicated Workstation. For our study, all participants used only English for demonstration and narration.

4.1 Understanding Existing Workflows

An essential first step w.r.t. Industry 4.0 and data-driven process improvement is understanding current practices and identifying bottlenecks and suitable tasks for automation. Egocentric devices, such as the Aria device, allow the non-intrusive collection of data for this objective with no added setup requirement.

Stationary Setups. Fig. 3 shows a scenario with an SME demonstrating the steps and considerations involved for sorting used metal parts based on visual features and dimension considerations. The intricacies involved in the process are captured with greater precision via the eye gaze prediction and hand

illustrations. The added context helps explain the workflow to a novice observer or third-party consultant.

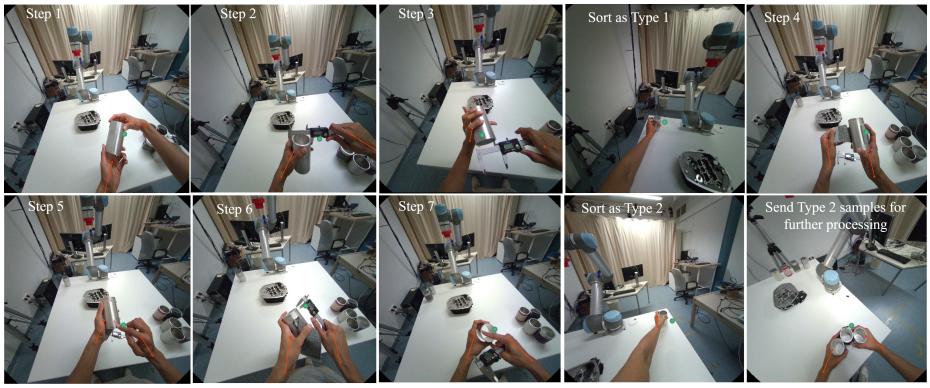


Fig. 3: Inspection workflows involving a single user and a stationary setup. The SME wearing the Aria glasses explains the steps and miscellaneous aspects of the operation using natural language. The goal (in the scenario shown) is to sort the parts as Type 1 and Type 2 based on surface features and dimension measurements. The eye gaze prediction (green) and the hand/wrist pose estimation (orange) can also be seen.

Non-stationary Workflows and Processes involving Multiple Users. We research a multistep use case where two users wearing the egocentric device carry work in tandem and move across different locations. In such circumstances, the semi-dense point cloud representation obtained using MPS processing of the SLAM cameras and user motion (and trajectory) provides relevant contextual information, as shown in Fig. 4. Inspection and review can also reveal fine-tuned details, such as the time/effort needed to do each step, obstacles, or bottlenecks.

4.2 Classification and Part Re-identification

Image classification is one of the most widely used applications of AI-based data processing [9, 30]. We collected video data for 100 commonly found industrial objects. The data consists of users interacting with the objects in their usual context (e.g. a drilling machine being switched on and operated). Fig. 5 shows some examples. The frames were randomly shuffled and split 80/20 into training and validation sets. The goal of the experiment is to assess whether the data collected using egocentric devices with a wide field of view can be used in industrial scenarios for training ML models. We fine-tune pretrained ResNet (RN) [22] models on the dataset and report a validation accuracy of 94.3% (RN50) to 96.1% (RN152) on distorted image data. The accuracy results on smaller objects (e.g. screws) were understandably worse. Undistorting the data improved the results marginally (<1%). We also digitise the same objects using

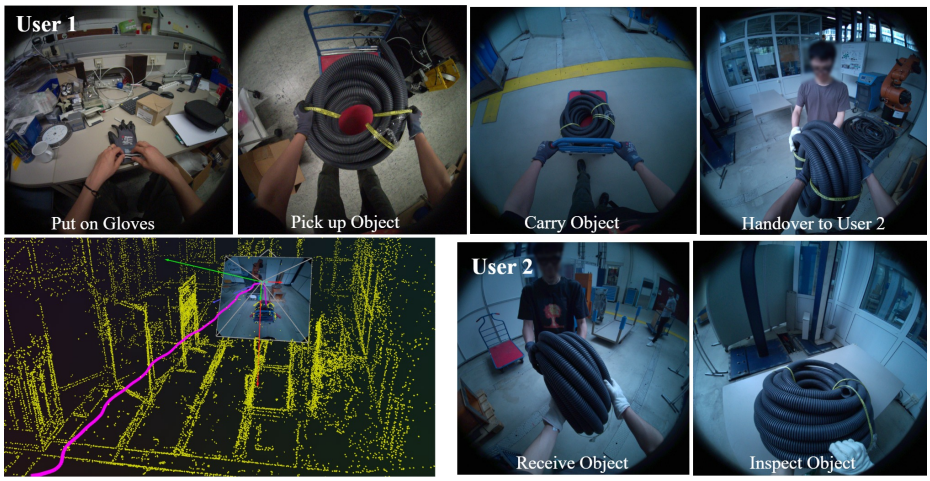


Fig. 4: Inspection of workflows involving multiple users. Both users wear the Aria glasses so that their motion and points of view can be captured. **Top:** Frames from the recording for User 1. During operation, the user comments on the actions. **Bottom Left:** A semi-dense point cloud representation of the working space (yellow) along with the trajectory for User 1 (pink), aria device coordinates, and the projected RGB frame. **Bottom Right:** Frames from the recording for User 2 along with commented actions.

a smartphone camera (Samsung Galaxy A51) and use it as a test set. We notice that the model trained on the undistorted images generalises better.

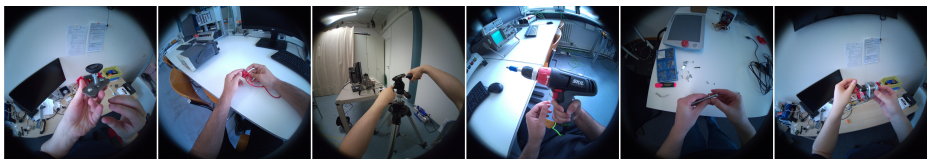


Fig. 5: Sample images from the different objects digitised during the study. The frames were extracted directly from the raw video stream and contain fisheye distortion. The 11 participants' natural interaction with 100 industrial objects was recorded using the Aria device.

4.3 Segmentation and Localisation using Eye Gaze

Part Segmentation. We explore the use of the projected eye gaze information on the RGB data to provide details on the object of interest in the image/video stream. Fig. 6 shows such an example on two different parts. Segment Anything [28] is used (zero-shot) to segment out the objects. In our observation,

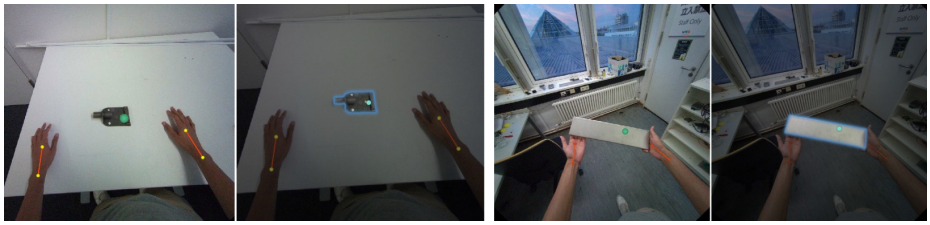


Fig. 6: An example of image segmentation using the eye gaze point (green) as the prompt input for the Segment Anything Model [28]. The eye gaze model accurately predicts the user's gaze within a few centimetres for handheld objects. The hand/wrist estimation (orange) is also visible.

medium and large parts were correctly segmented out. However, smaller parts (shortest dimension $<$ 2cm) were occasionally inappropriately segmented due to a slight error in the eye gaze prediction.

Defect Localisation. We utilise the eye gaze projection to localise the visual defects. Fig. 7 shows an example of a metal part with multiple scratches of varying depths. The gaze predicted by the proprietary model is generally accurate, with a slight tilt in the vector projection. The depth prediction for the eye gaze remains comparatively accurate. In this case, the predicted gaze was consistently off to the right of the intended location. The trend does not hold across different participants. The vector projection had a bias towards the top of the ground truth for other users, and occasionally was also correctly predicted. However, the projected eye gaze is still an effective indicator of the region of interest, with accuracy within a few cm for handheld objects.

4.4 Inspection of Large Objects and Assemblies

Egocentric vision enables the user to move freely and interact with their surroundings. We consider this to be valuable for inspecting large devices, machines, and assemblies. Fig. 8 shows an application, where the SME inspects the assembly in a robotics lab from different perspectives. The user trajectory and the corresponding sample RGB frames can be seen. Similar to previous experiments, 3D point cloud and user movement add crucial context to the use case. Eye gaze and hand interaction, in conjunction with verbal descriptions, add further information (e.g. "*This* screw is loose and must be tightened"). The trajectory information is useful when sampling a small subset of the data stream for curating a dataset, or to verify the point of view that led to a given assessment.

4.5 Object Detection and Tracking

We use the same raw data as the classification task. In this case, we train a YOLO V5 [24] model on 50% of the data and test the model on the other portion. The ability of the model to detect small handheld objects deteriorates

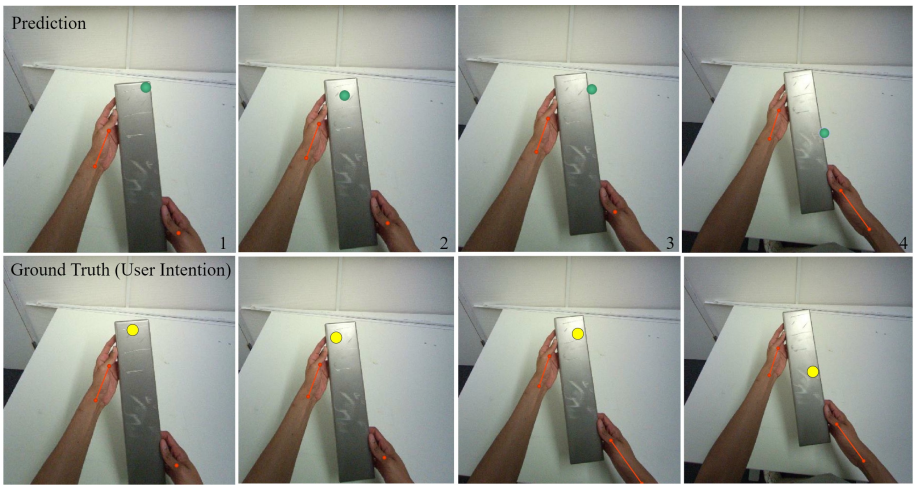


Fig. 7: Top: An example of defect localisation using the predicted eye gaze point (green) as the prompt input for indicating the defect location. **Bottom:** Intended eye gaze point (yellow) as verified by the user. The prediction model tends to slightly deviate (in this case, with a bias towards the right side) from the ground truth. The hand/wrist estimation (orange) is also visible. For reference, object dimensions are approx. 40cm x 10cm.

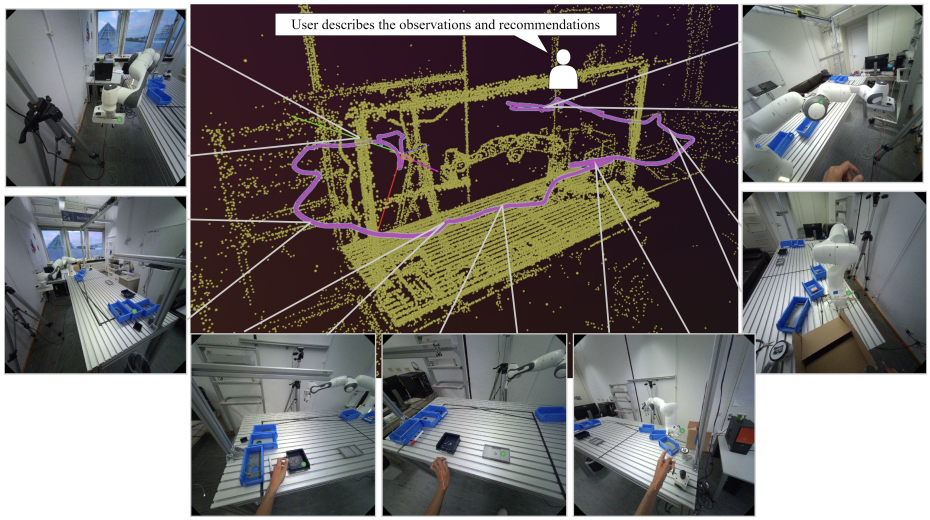


Fig. 8: Visual inspection of a large assembly from an egocentric perspective. **Centre:** A 3D point cloud representation of the assembly and the surrounding environment (yellow) along with the projected user trajectory (pink). An outline of the glasses along with the eye-gaze vector is also visible. **Encircling:** RGB camera view from the device from different points of view. The eye gaze is visible in each frame (green) and the hand/wrist pose can be seen in certain frames (orange).

for cluttered backgrounds. Similarly, the models trained on objects in one background condition do not generalise well to data in an unseen background. Additionally, we observe an egocentric-allocentric performance gap, which improves when undistorted frames are extracted from the egocentric raw data.

4.6 Optical Character Recognition

OCR is a crucial industrial application and can aid other tasks such as part re-identification. We experimented with using the extracted frames as an input to a pretrained OCR model. GLASS [44] and PaddleOCR [39] were used on image data containing printed alphanumeric information on equipment and boxes. The models were able to proficiently identify large text (at least 1cm in height at arm's distance) for natural language and well-known symbols (e.g. FedEx). The performance for smaller prints and documents was significantly worse unless the user held the objects close to the RGB camera sensor. When it comes to OCR, egocentric devices are not an optimal digitisation platform.

5 Discussion

Overview. We proposed a general methodological framework for automated data collection, labelling, and inspection using egocentric devices and SOTA multimodal foundational models. The project Aria device is a research tool, which aims to explore the future potential of smart eyewear devices via a sensor suite and Machine Perception Services for processing the raw data. The goal of our proposed pipeline is to make it intuitive and non-intrusive for SMEs to digitise their workflows and provide relevant context. In that regard, our work applies certain principles from Industry 5.0 [15], to current practices, viz. *combining advanced technologies with human skills and creativity, and improving human-machine interaction and collaboration.*

5.1 Evaluation of the Experiments.

Understanding Workflows. The biggest upside associated with using egocentric devices for industrial scenarios lies in cases where it is inconvenient or infeasible for a specialised workstation and data collection equipment to be set up, viz. §4.1, §4.4. All members were freshly introduced to the Aria device during the study and did not have any prior experience with wearable/AR/VR technology. One crucial aspect when digitising existing workflows is to encourage SMEs to share as many details as possible. In our observation, participants often skip explaining certain routine aspects of their tasks, which may not be commonly understood by others. In particular, this tended to be significant for long sequences of tasks, where the SMEs reverted to their habitual patterns. Additionally, the participants had to be regularly instructed to be mindful of their eye gaze while working on a task. Humans can use their System 1 thinking [25] when working on simpler tasks and can rely on other senses, such as touch, for

the task. The MPS output, coupled with a detailed workflow narration, provides a solid starting point for identifying areas of improvement and automation in the workflow.

Classification and Detection. The data collected using the 100 industrial objects for the classification and object detection tasks will be made publicly available. A key advantage of using always-on perception devices is the increased speed with which data can be collected. During our study, the 100 objects were successfully digitised within 8 cumulative hours (including breaks) spread over 4 days. Based on the results, the data captured from the glasses can be used for training models for part re-identification and detection tasks. However, egocentric data often includes background variance and clutter, occlusion due to hands, among other issues. This leads to poor generalisability when testing models trained on allocentric or clean data on data from an egocentric device, or vice versa. Other works on industrial use cases have also shown similar issues with handheld and cluttered background data [7].

Annotations using Eye Gaze. The predicted eye gaze can serve as an effective signal to indicate the area of interest, or the region of the field of view, being addressed at that moment. Foundational models such as SAM increase the efficacy of added context with wearable devices. Although the predicted gaze vector was marginally off w.r.t. the user intention, the depth estimation of the gaze was accurate in all scenarios, and for all participants. It is expected that this proprietary technology will be open-sourced in the near future, or there will be open-source alternatives, which would enable further research and exploration on using gaze prediction for various tasks.

Large Machinery and Assembly. Egocentric wearable devices are uniquely suitable for inspection of large machinery and assembly, where the user freely moves around the object/s of interest, makes observations and collects useful data. Especially with MPS, obtaining a 3D understanding of such setups in relation to the user and the surroundings can be highly useful. Due to the exploratory nature of our work, we focused more on examining multiple scenarios and approaches, instead of rigorously benchmarking on one particular task.

5.2 Additional Considerations

Data Protection. Egocentric devices generate and process a lot of personal data, including eye gaze, hand gestures, voice narration, etc. These sensitive personal details must be stored and processed in accordance with user preference and regional regulations [14]. Additionally, when using them on industrial applications, the devices are bound to capture sensitive business information. Hence, the organisation must establish clear guidelines and policies for using always-on devices responsibly. For our study, permission was taken from the research lab to collect data on their premises. Each participant received clear guidelines before each session, and their Personally Identifiable Information was anonymised as far as possible.

Limitations. There are multiple drawbacks when using egocentric devices for data collection and downstream applications. For instance, the camera qual-

ity on the Aria device is significantly poor (8MP, max resolution 2880 X 2880, no autofocus) compared to industrial-grade camera setups. In their current iteration, egocentric devices have a limited battery life, which limits the possible use cases that can be explored and also necessitates regular breaks in the workflow. Esp. Eyewear devices must be lightweight and cannot heat up excessively, since dissipating heat could be uncomfortable or even damaging to the wearer. This makes it a substantially weaker computing platform compared to desktop and mobile devices. Ideally, egocentric devices would be used in real-time for inference and user assistance. However, due to the unavailability of an output modality, we could not explore challenges and issues with on-device operations.

Capturing sufficient quantities of rich multimodal data involves substantial storage and computational demands. The MPS models are not open-source, which inhibits other users from exploring and developing the technology further.

Problem Formulation for Future Work. This preliminary investigation leaves sufficient room for research in several directions. It is evident that egocentric wearable devices are not a replacement for traditional vision-based inspection systems. This leads to the question- *How can egocentric vision and related technologies be optimally integrated into existing Industry 4.0 oriented workflows?* Ongoing egocentric research includes mapping egocentric and exocentric views of the same event to enhance total understanding [20, 46]. This direction would be conducive to vision-based inspection tasks, as shown in Fig 9.

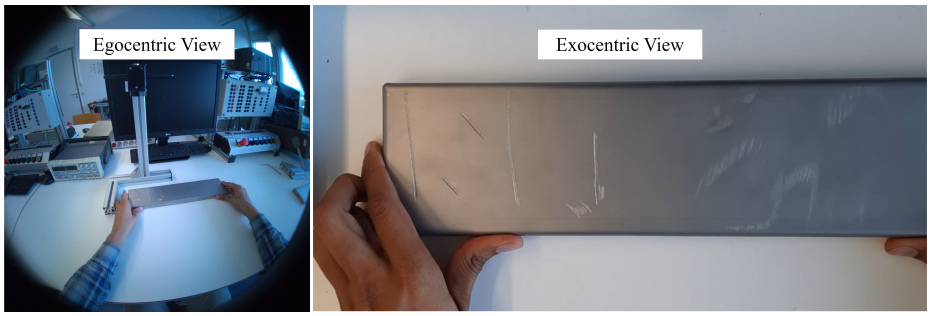


Fig. 9: An example use-case for utilising egocentric and exocentric perspectives for inspection tasks. **Left:** Egocentric view from the wearable device, which is used for general process inspection and understanding. **Right:** Exocentric view via a fixed higher quality camera and optimal lighting dedicated for image-based inspection.

Another research direction, w.r.t. algorithmic progress, is the development of foundational models for industry-specific tasks and use cases. This would require large amounts of useful and context-rich data (which would first be adequately anonymised) and compute. An important question is- *How can open-sourcing such data help in diving the community forward? Which modalities, use cases and application scenarios are needed? How do we ensure that the captured data contains sufficient diversity w.r.t. users/SMEs, working environments and use*

cases? However, models trained on this data may still only interface with the egocentric devices with a limited compute budget, due to the hardware and power management constraints.

Hence, it is essential to revisit the hardware requirements and ideal setup for industrial inspection and operator assistance. For instance, *could egocentric devices accommodate multiple RGB cameras of varying focal lengths (including a dedicated sensor for high-quality imaging), similar to modern smartphones? How can egocentric devices be set up inside the organisation to interface with local workstations for computational assistance? How can smart wearable devices be set up to interface with each other for high-speed communication and collaboration?*

Lastly, *which additional areas of industry-specific research can be addressed using the existing hardware and software stack of egocentric vision?* For example, several egocentric wearable devices (including Aria glasses) have multiple microphones for spatial audio understanding. Could sound-based inspection complement vision-based approaches? Robotics research is another area of research where egocentric data is effective [27]. How can data collected via egocentric devices be used to develop such solutions for inspection and operator assistance?

6 Conclusion

In this paper, we investigated the use of egocentric vision devices and related technologies for industrial inspection. The proposed pipeline aims to allow SMEs to digitise their workflow and collect multimodal data while adding relevant context via speech, eye gaze, and hand gestures. We explored vision-centric inspection applications using the data collected during this study. Understanding existing workflows with minimal intrusion and inspection of large assemblies and devices are some domains where egocentric vision can be uniquely productive. We also explored the conventional industrial applications, viz. part re-identification and detection, segmentation, defect localisation and OCR.

We discussed data protection and privacy considerations in industrial facilities and also highlighted the current limitations of this technology. However, egocentric vision is a relatively nascent area of research and is growing rapidly. Wearable devices are likely to significantly improve over the coming years, w.r.t. their hardware and software capabilities. Hence, we conjecture that egocentric AI could play a key role in driving industrial research forward and transitioning to Industry 5.0.

Acknowledgments. This work is funded by the German Federal Ministry of Education and Research (BMBF) and the German Aerospace Center (DLR) under the KIKERP project 0118S23055C in the KI4KMU program. We thank the Meta AI team and Reality Labs for the Project Aria initiative, including the research kit, the open-source tools and related services. The data collection for this study was carried out at the IWF research labs of TU Berlin by the AuT Project Groups 9.1 and 9.2 during the Summer Semester of 2024.

References

1. Apple: Apple vision pro (2024), https://www.apple.com/apple-vision-pro/, accessed: 2024-05-10 3
2. Bergmann, P., Fauser, M., Sattlegger, D., Steger, C.: Mvtec ad — a comprehensive real-world dataset for unsupervised anomaly detection. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9584–9592 (2019). https://doi.org/10.1109/CVPR.2019.00982 3
3. Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D.: Language models are few-shot learners (2020). https://doi.org/10.48550/ARXIV.2005.14165, https://arxiv.org/abs/2005.14165 3
4. Chai, J., Zeng, H., Li, A., Ngai, E.W.: Deep learning in computer vision: A critical review of emerging techniques and application scenarios. Machine Learning with Applications **6**, 100134 (2021). https://doi.org/https://doi.org/10.1016/j.mlwa.2021.100134, https://www.sciencedirect.com/science/article/pii/S2666827021000670 1
5. Chandola, V., Banerjee, A., Kumar, V.: Anomaly detection: A survey. ACM Comput. Surv. **41**(3) (jul 2009). https://doi.org/10.1145/1541880.1541882, https://doi.org/10.1145/1541880.1541882 2
6. Chavan, V., Koch, P., Schlüter, M., Briese, C., Krüger, J.: Active data collection and management for real-world continual learning via pretrained oracle. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. pp. 4085–4096 (June 2024) 4
7. Chavan, V., Koch, P., Schlüter, M., Briese, C.: Towards realistic evaluation of industrial continual learning scenarios with an emphasis on energy consumption and computational footprint. In: Proceedings of the International Conference on Computer Vision (ICCV) (2023) 12
8. Damen, D., Doughty, H., Farinella, G.M., Fidler, S., Furnari, A., Kazakos, E., Moltisanti, D., Munro, J., Perrett, T., Price, W., Wray, M.: Scaling egocentric vision: The epic-kitchens dataset. In: Proceedings of the European Conference on Computer Vision (ECCV) (September 2018) 3
9. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. pp. 248–255 (2009). https://doi.org/10.1109/CVPR.2009.5206848 3, 7
10. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2021), https://openreview.net/forum?id=YicbFdNTTy 3
11. Durrant-Whyte, H., Bailey, T.: Simultaneous localization and mapping: part i. IEEE Robotics & Automation Magazine **13**(2), 99–110 (2006). https://doi.org/10.1109/MRA.2006.1638022 3
12. Engel, J., Somasundaram, K., Goesele, M., Sun, A., Gamino, A., Turner, A., Talattof, A., Yuan, A., Souti, B., Meredith, B., Peng, C., Sweeney, C., Wilson, C., Barnes, D., DeTone, D., Caruso, D., Valleroy, D., Ginjupalli, D., Frost, D., Miller,

E., Mueggler, E., Oleinik, E., Zhang, F., Somasundaram, G., Solaira, G., Lanaras, H., Howard-Jenkins, H., Tang, H., Kim, H.J., Rivera, J., Luo, J., Dong, J., Straub, J., Bailey, K., Eckenhoff, K., Ma, L., Pesqueira, L., Schwesinger, M., Monge, M., Yang, N., Charron, N., Raina, N., Parkhi, O., Borschowa, P., Moulon, P., Gupta, P., Mur-Artal, R., Pennington, R., Kulkarni, S., Miglani, S., Gondi, S., Solanki, S., Diener, S., Cheng, S., Green, S., Saarinen, S., Patra, S., Mourikis, T., Whelan, T., Singh, T., Balntas, V., Baiyya, V., Dreewes, W., Pan, X., Lou, Y., Zhao, Y., Mansour, Y., Zou, Y., Lv, Z., Wang, Z., Yan, M., Ren, C., Nardi, R.D., Newcombe, R.: Project aria: A new tool for egocentric multi-modal ai research (2023), https://arxiv.org/abs/2308.13561 3, 4

13. Engel, J., Somasundaram, K., Goesele, M., Sun, A., Gamino, A., Turner, A., Talattof, A., Yuan, A., Souti, B., Meredith, B., Peng, C., Sweeney, C., Wilson, C., Barnes, D., DeTone, D., Caruso, D., Valleroy, D., Ginjupalli, D., Frost, D., Miller, E., Mueggler, E., Oleinik, E., Zhang, F., Somasundaram, G., Solaira, G., Lanaras, H., Howard-Jenkins, H., Tang, H., Kim, H.J., Rivera, J., Luo, J., Dong, J., Straub, J., Bailey, K., Eckenhoff, K., Ma, L., Pesqueira, L., Schwesinger, M., Monge, M., Yang, N., Charron, N., Raina, N., Parkhi, O., Borschowa, P., Moulon, P., Gupta, P., Mur-Artal, R., Pennington, R., Kulkarni, S., Miglani, S., Gondi, S., Solanki, S., Diener, S., Cheng, S., Green, S., Saarinen, S., Patra, S., Mourikis, T., Whelan, T., Singh, T., Balntas, V., Baiyya, V., Dreewes, W., Pan, X., Lou, Y., Zhao, Y., Mansour, Y., Zou, Y., Lv, Z., Wang, Z., Yan, M., Ren, C., Nardi, R.D., Newcombe, R.: Project aria: A new tool for egocentric multi-modal ai research (2023) 20
14. European Commission: Data protection (2023), https://commission.europa.eu/law/law-topic/data-protection_en, accessed: 2023-07-28 12
15. European Commission: Industry 5.0 (2023), https://research-and-innovation.ec.europa.eu/research-area/industrial-research-and-innovation/industry-50_en, accessed: 2023-07-28 2, 11
16. Facebook: Facebook to acquire oculus (2014), https://about.fb.com/news/2014/03/facebook-to-acquire-oculus/, accessed: 2024-05-10 3
17. Facebook Research: Introduction to project aria docs (2023), https://facebookresearch.github.io/projectaria_tools/docs/intro, accessed: 2023-07-24 6, 20
18. Goodfellow, I., Bengio, Y., Courville, A., Bengio, Y.: Deep learning, vol. 1. MIT Press (2016) 3
19. Grauman, K., Westbury, A., Byrne, E., Chavis, Z., Furnari, A., Girdhar, R., Hamburger, J., Jiang, H., Liu, M., Liu, X., Martin, M., Nagarajan, T., Radosavovic, I., Ramakrishnan, S.K., Ryan, F., Sharma, J., Wray, M., Xu, M., Xu, E.Z., Zhao, C., Bansal, S., Batra, D., Cartillier, V., Crane, S., Do, T., Doulaty, M., Erapalli, A., Feichtenhofer, C., Fragomeni, A., Fu, Q., Gebreselasie, A., González, C., Hillis, J., Huang, X., Huang, Y., Jia, W., Khoo, W., Kolář, J., Kottur, S., Kumar, A., Landini, F., Li, C., Li, Y., Li, Z., Mangalam, K., Modhugu, R., Munro, J., Murrell, T., Nishiyasu, T., Price, W., Ruiz, P., Ramazanova, M., Sari, L., Somasundaram, K., Southerland, A., Sugano, Y., Tao, R., Vo, M., Wang, Y., Wu, X., Yagi, T., Zhao, Z., Zhu, Y., Arbeláez, P., Crandall, D., Damen, D., Farinella, G.M., Fuegen, C., Ghanem, B., Ithapu, V.K., Jawahar, C.V., Joo, H., Kitani, K., Li, H., Newcombe, R., Oliva, A., Park, H.S., Rehg, J.M., Sato, Y., Shi, J., Shou, M.Z., Torralba, A., Torresani, L., Yan, M., Malik, J.: Ego4d: Around the world in 3,000 hours of egocentric video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 18995–19012 (June 2022) 3

20. Grauman, K., Westbury, A., Torresani, L., Kitani, K., Malik, J., Afouras, T., Ashutosh, K., Baiyya, V., Bansal, S., Boote, B., Byrne, E., Chavis, Z., Chen, J., Cheng, F., Chu, F.J., Crane, S., Dasgupta, A., Dong, J., Escobar, M., Forigua, C., Gebreselasie, A., Haresh, S., Huang, J., Islam, M.M., Jain, S., Khirodkar, R., Kukreja, D., Liang, K.J., Liu, J.W., Majumder, S., Mao, Y., Martin, M., Mavroudi, E., Nagarajan, T., Ragusa, F., Ramakrishnan, S.K., Seminara, L., Somayazulu, A., Song, Y., Su, S., Xue, Z., Zhang, E., Zhang, J., Castillo, A., Chen, C., Fu, X., Furuta, R., Gonzalez, C., Gupta, P., Hu, J., Huang, Y., Huang, Y., Khoo, W., Kumar, A., Kuo, R., Lakhavani, S., Liu, M., Luo, M., Luo, Z., Meredith, B., Miller, A., Oguntola, O., Pan, X., Peng, P., Pramanick, S., Ramazanova, M., Ryan, F., Shan, W., Somasundaram, K., Song, C., Southerland, A., Tateno, M., Wang, H., Wang, Y., Yagi, T., Yan, M., Yang, X., Yu, Z., Zha, S.C., Zhao, C., Zhao, Z., Zhu, Z., Zhuo, J., Arbelaez, P., Bertasius, G., Crandall, D., Damen, D., Engel, J., Farinella, G.M., Furnari, A., Ghanem, B., Hoffman, J., Jawahar, C.V., Newcombe, R., Park, H.S., Rehg, J.M., Sato, Y., Savva, M., Shi, J., Shou, M.Z., Wray, M.: Ego-exo4d: Understanding skilled human activity from first- and third-person perspectives (2024) 3, 13
21. Haffner, O., Kučera, E., Rosinová, D.: Applications of machine learning and computer vision in industry 4.0. Applied Sciences **14**(6) (2024). https://doi.org/10.3390/app14062431, https://www.mdpi.com/2076-3417/14/6/2431 3
22. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016) 7, 20
23. Huang, S., Chen, Y., Jia, J., Wang, L.: Multi-view transformer for 3d visual grounding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15524–15533 (2022) 3
24. Jocher, G., Chaurasia, A., Stoken, A., Borovec, J., NanoCode012, Kwon, Y., Michael, K., Xie, T., Fang, J., imyhxy, Lorna, Yifu, Z., Wong, C., V, A., Montes, D., Wang, Z., Fati, C., Nadar, J., Laughing, Jain, M.: ultralytics/yolov5: v7.0 - YOLOv5 SOTA Realtime Instance Segmentation (v7.0) (December 2022). https://doi.org/10.5281/zenodo.7347926, https://doi.org/10.5281/zenodo.7347926 9, 20
25. Kahneman, D.: A perspective on judgment and choice: mapping bounded rationality. American Psychologist **58**(9), 697–720 (Sep 2003). https://doi.org/10.1037/0003-066X.58.9.697, https://doi.org/10.1037/0003-066X.58.9.697, pMID: 14584987 11
26. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering (2023), https://arxiv.org/abs/2308.04079 3
27. Kim, D., Kang, B.B., Kim, K.B., Choi, H., Ha, J., Cho, K.J., Jo, S.: Eyes are faster than hands: A soft wearable robot learns user intention from the egocentric view. Science Robotics **4**(26), eaav2949 (2019). https://doi.org/10.1126/scirobotics.aav2949, https://www.science.org/doi/abs/10.1126/scirobotics.aav2949 14
28. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollar, P., Girshick, R.: Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 4015–4026 (October 2023) 3, 8, 9, 20
29. Koch, S., Matveev, A., Jiang, Z., Williams, F., Artemov, A., Burnaev, E., Alexa, M., Zorin, D., Panozzo, D.: Abc: A big cad model dataset for geometric deep

learning. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019) 3
30. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Pereira, F., Burges, C., Bottou, L., Weinberger, K. (eds.) Advances in Neural Information Processing Systems. vol. 25. Curran Associates, Inc. (2012), https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf 3, 7
31. Lecun, Y., Bengio, Y.: Convolutional networks for images, speech, and time-series (01 1995) 3
32. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. nature **521**(7553), 436–444 (2015) 1
33. Lee, Y.J., Ghosh, J., Grauman, K.: Discovering important people and objects for egocentric video summarization. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition. pp. 1346–1353 (2012). https://doi.org/10.1109/CVPR.2012.6247820 3, 4
34. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows (2021), https://arxiv.org/abs/2103.14030 3
35. Malamas, E.N., Petrakis, E.G., Zervakis, M., Petit, L., Legat, J.D.: A survey on industrial vision systems, applications and tools. Image and Vision Computing **21**(2), 171–188 (2003). https://doi.org/https://doi.org/10.1016/S0262-8856(02)00152-X, https://www.sciencedirect.com/science/article/pii/S026288560200152X 3
36. Mann, S.: Humanistic computing: "wearcomp" as a new framework and application for intelligent signal processing. Proceedings of the IEEE **86**(11), 2123–2151 (1998). https://doi.org/10.1109/5.726784 3
37. Mazzei, D., Ramjattan, R.: Machine learning for industry 4.0: A systematic review using deep learning-based topic modelling. Sensors **22**(22) (2022). https://doi.org/10.3390/s22228641, https://www.mdpi.com/1424-8220/22/22/8641 1, 3
38. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y., Li, S.W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: Dinov2: Learning robust visual features without supervision (2023) 3
39. PaddlePaddle: Paddleocr documentation (2023), https://paddlepaddle.github.io/PaddleOCR, accessed: 2024-07-28 11
40. Parthasarathy, N., Eslami, S.M.A., Carreira, J., Henaff, O.J.: Self-supervised video pretraining yields strong image representations (2023), https://openreview.net/forum?id=8onXkaNWLHA 2
41. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision (2021), https://arxiv.org/abs/2103.00020 3
42. Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I.: Robust speech recognition via large-scale weak supervision (2022), https://arxiv.org/abs/2212.04356 4, 20
43. Raina, N., Somasundaram, G., Zheng, K., Miglani, S., Saarinen, S., Meissner, J., Schwesinger, M., Pesqueira, L., Prasad, I., Miller, E., Gupta, P., Yan, M., Newcombe, R., Ren, C., Parkhi, O.M.: Egoblur: Responsible innovation in aria (2023) 6

44. Ronen, R., Tsiper, S., Anschel, O., Lavi, I., Markovitz, A., Manmatha, R.: Glass: Global to local attention for scene-text spotting. arXiv preprint arXiv:2208.03364 (2022) 11
45. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., Schramowski, P., Kundurthy, S., Crowson, K., Schmidt, L., Kaczmarczyk, R., Jitsev, J.: Laion-5b: An open large-scale dataset for training next generation image-text models (2022). https://doi.org/10.48550/ARXIV.2210.08402, https://arxiv.org/abs/2210.08402 3
46. Sener, F., Chatterjee, D., Shelepov, D., He, K., Singhania, D., Wang, R., Yao, A.: Assembly101: A large-scale multi-view video dataset for understanding procedural activities. CVPR 2022 3, 13
47. Smith, M.L., Smith, L.N., Hansen, M.F.: The quiet revolution in machine vision - a state-of-the-art survey paper, including historical review, perspectives, and future directions. Computers in Industry **130**, 103472 (2021). https://doi.org/https://doi.org/10.1016/j.compind.2021.103472, https://www.sciencedirect.com/science/article/pii/S0166361521000798 3
48. Szeliski, R.: Computer Vision: Algorithms and Applications. Springer-Verlag, Berlin, Heidelberg, 1st edn. (2010) 1, 3
49. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jegou, H.: Training data-efficient image transformers & distillation through attention. In: Meila, M., Zhang, T. (eds.) Proceedings of the 38th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 139, pp. 10347–10357. PMLR (18–24 Jul 2021), https://proceedings.mlr.press/v139/touvron21a.html 2
50. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., Lample, G.: Llama: Open and efficient foundation language models (2023). https://doi.org/10.48550/ARXIV.2302.13971, https://arxiv.org/abs/2302.13971 4, 20
51. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.: Attention is all you need. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc. (2017), https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf 3
52. Zhang, J., Huang, J., Jin, S., Lu, S.: Vision-language models for vision tasks: A survey (2024), https://arxiv.org/abs/2304.00685 3

Appendix

A. Experimental Setup

Table 1 gives the details of the configuration for the tests in Sections 3 (Methods) and 4 (Experiments). We use the same setup for all our Machine Learning (ML) trainings for a fair and unbiased comparison.

Table 1: Details of the parameters and setup used for the experiments presented in the paper involving ML training.

Parameter	**Value**
Train-Val Split	80/20 (classification)
Optimizer	SGD
lr start	0.1
lr end	0.0001
weight decay	0.0005
Batch Size	64
Transforms: Train	Resize: (224, 224), RandomHorizontalFlip
Transforms: Val	Resize: (256, 256), CenterCrop
System Memory	48GB
CPU Cores	12
GPU Count	1
GPU type	NVIDIA RTX A6000
Python version	3.8.18

Additional Context

Our code implementation borrows heavily from open source repositories of Project Aria [13], Llama [50], OpenAI Whisper [42], and others. Details on processing raw data collected using the Aria device can be found on the Project Aria documentation page [17]. The ResNet [22], YOLO [24] and Segment Anything [28] models were used for experimentation and benchmarking in this paper, because of their ease of use and widespread adoption.