An Outlook into the Future of Egocentric Vision

Chiara Plizzari[1] · Gabriele Goletto[1] · Antonino Furnari[2] · Siddhant Bansal[3] · Francesco Ragusa[2] ·
Giovanni Maria Farinella[2] · Dima Damen[3] · Tatiana Tommasi[1]

Abstract

What will the future be? We wonder! In this survey, we explore the gap between current research in egocentric vision and the ever-anticipated future, where wearable computing, with outward facing cameras and digital overlays, is expected to be integrated in our every day lives. To understand this gap, the article starts by envisaging the future through character-based stories, showcasing through examples the limitations of current technology. We then provide a mapping between this future and previously defined research tasks. For each task, we survey its seminal works, current state-of-the-art methodologies and available datasets, then reflect on shortcomings that limit its applicability to future research. Note that this survey focuses on software models for egocentric vision, independent of any specific hardware. The paper concludes with recommendations for areas of immediate explorations so as to unlock our path to the future always-on, personalised and life-enhancing egocentric vision.

Keywords Egocentric vision · Future · Survey · Localisation · Scene understanding · Recognition · Anticipation · Gaze prediction · Social understanding · Body pose estimation · Hand and hand-object interaction · Person identification · Summarisation · Dialogue · Privacy

Contents

Communicated by Limin Wang.

✉ Tatiana Tommasi
  tatiana.tommasi@polito.it

1   Politecnico di Torino, Turin, Italy

2   University of Catania, Catania, Italy

3   University of Bristol, Bristol, UK

1 Introduction

Designing and building tools able to support human activities, improve quality of life, and enhance individuals' abilities to achieve their goals is the ever-lasting aspiration of our species. Among all inventions, digital computing has already had a revolutionary effect on human history. Of particular note is mobile technology, currently integrated in our lives through hand-held devices, i.e. *mobile smart phones*. These are nowadays the de facto for outdoor navigation, capturing static and moving footage of our everyday and connecting us to both familiar and novel connections and experiences.

However, humans have been dreaming about the next-version of such mobile technology—wearable computing, for a considerable amount of time. Imaginations are present in movies, fictional novels and pop culture[1]. Notwithstanding the fast progress of Artificial Intelligence, and the hardware advances of the last 10 years, our ability to fulfil this dream is lagging behind.

In computer vision, research papers on egocentric vision have instead limited their focus to a handful of applications, where current technology can already make a difference. These are: training or monitoring in industrial settings, performing adhoc and infrequent tasks such as assembling a piece of furniture, preparing a new recipe, or playing a group game in a social setting. These showcase egocentric wearables as niche devices very distant from everyone's everyday needs. This perspective has not only limited our chances to convince others that egocentric vision is a key technology of our future, but it also restricted our ability to push the boundaries and remove obstacles to the integration of egocentric devices as the ultimate replacement of the *mobile phone* with unlocking of additional capabilities.

To make a difference, we choose a future-to-present perspective in this paper, where we start from the envisaged future then analyse the fundamental tasks that are required. This approach allows us to take a more systemic and informed perspective, highlight the gap between the expected applications and the current technological status, and provide insights into promising future research directions. While

technology forecasting is not a very common approach to research review, Firat et al. (2008) note its value in prioritising R&D. We take a scenario-based approach, amongst the options proposed in Firat et al. (2008) for future forecasting.

Our work is related to previous surveys in egocentric vision. Betancourt et al. (2015) summarised the evolution of the state of the art in egocentric vision analysis from 1997 to 2014, the year of writing of the survey. Nguyen et al. (2016) reviewed algorithms for the recognition of activities of daily living from egocentric vision. Bolaños et al. (2016) surveyed approaches for visual storytelling from the analysis of egocentric photo-streams. Del Molino et al. (2016) provided a survey of techniques for used to summarise egocentric video. Rodin et al. (2021) analysed algorithms, datasets and tasks for action anticipation in egocentric vision. Núñez-Marcos et al. (2022) summarised works in egocentric action recognition. Bandini and Zariffa (2023) considered works based on the analysis of hands in egocentric vision.

All previous survey papers, with the exception of Betancourt et al. (2015), addressed specific topics in egocentric vision. In contrast, this paper offers a holistic overview. We also offer a comprehensive and updated view of the current status of egocentric vision, covering topics of localisation (Sect. 4.1), scene understanding (Sect. 4.2), recognition (Sect. 4.3), anticipation (Sect. 4.4), gaze understanding and prediction (Sect. 4.5), social behaviour understanding (Sect. 4.6), full-body pose estimation (Sect. 4.7), hand and hand-object interactions (Sect. 4.8), person identification (Sect. 4.9), summarisation (Sect. 4.10), dialogue (Sect. 4.11), and privacy (Sect. 4.12).

The remainder of this paper is organised as follows. In Sect. 2, we present our vision of the future of egocentric vision through character-based stories and associated visuals. Section 3 relates these stories to research tasks, structuring these into familiar research questions. In Sect. 4, we survey each task with subsections dedicated to seminal works, current state-of-the-art, dedicated datasets to these tasks and limitations to future applications. In Sect. 5, we present general datasets frequently used in egocentric vision beyond a single task. Finally, in Sect. 6 we conclude by providing a perspective to key questions that need to be unlocked soon for a step-change in egocentric vision.

2 Imagining the Future

With the aim of performing an inspirational review of the current status of egocentric vision, we look into how research outputs are expected to impact our everyday life in the near future and investigate the gap still existing towards those results. The envisaged future takes the shape of *five distinctive use cases that are grounded in either a location or an occupation*. In presenting each use case, we first sum-

---

[1] Few examples: (1) Molly's Vision-Enhancing Lenses from the *Neuromancer* novel, William Gibson, 1984. (2) JVC Personal Video Glasses from the *Back to the Future II* movie, 1989. (3) Iron Man Suits with J.A.R.V.I.S. AI system from Marvel movies 2008–2015. (4) AI Earbuds and smartphone in shirt pocket from the *Her* movie, 2013. (5) E.D.I.T.H. smart glasses from the *Spider-Man: Far From Home* movie, 2019.

marise the existing relevant technology and then we introduce future narratives in the form of brief character-based stories, supporting the readers' imagination through artist-drawn sketches. The protagonists of the plots use *EgoAI, a wearable device that enables in-situ multimodal sensing from the wearer's perspective and provides ego-based assistance*. We associate story ⓟarts with research tasks (marked by section number) and later revisit the link between these use-cases and research tasks in Sect. 3.

2.1 EGO-Home

Presently, smart home technology encompasses a range of Internet of Things (IoT) devices. They either control specific domestic environmental variables (e.g. light, temperature, humidity, CO2 level, energy consumption) or manage the operation of electrical appliances on the basis of occupants' preferences. Surveillance cameras are increasingly being installed both indoors and outdoors to ensure safety and enable remote monitoring of pets, children, and the elderly. Furthermore, there has been a recent surge in the introduction of speakers assistive devices like Amazon Echo, Google Nest, and Apple HomePod, which mainly rely on audio input for interaction and event tracking. All these tools, though empowered by machine learning techniques, are dedicated to a few specific tasks and they are static in nature, covering only limited areas of the home. EgoAI will replace the set of heterogeneous sensing tools currently in operation, but also provide much more.

Sam (Fig. 1) is finally at home after a hard-working day. A good dinner is certainly needed. When Sam opens the fridge, EgoAI *automatically analyses present stock in view and suggests a tomato soup as the tomatoes look perfectly ripe. Moreover*, EgoAI *has kept track of Sam's food intake for that day and the soup sounds like the best complementary nutrition. Sam does not enjoy cooking much, but* EgoAI *switches on the 3D projection of the Remy from the movie Ratatouille to help him through the soup prep. Remy jumps around his kitchen efficiently avoiding obstacles and appears to hold the knife while he chops the tomato encouraging Sam to slice his tomatoes thinner. Remy says, "this way the tomato will cook evenly". The audio appears to come from the direction of the chopping board, where now Remy is comfortably sitting. Sam is continuously impressed by how fun it is to cook with his 3D projected friend. Sam is in doubt about the amount of spice he has added and whether more is recommended.* EgoAI *keeps track of ingredients, recommends more spice to be added and reminds Sam about the bread slice he's nearly forgotten in the toaster.*

While enjoying his warm soup, Sam asks EgoAI to take him back to that beach he visited last summer. Sam is virtually transformed to that same view he captured many months ago, and relaxes by listening to the waves hitting the shore while
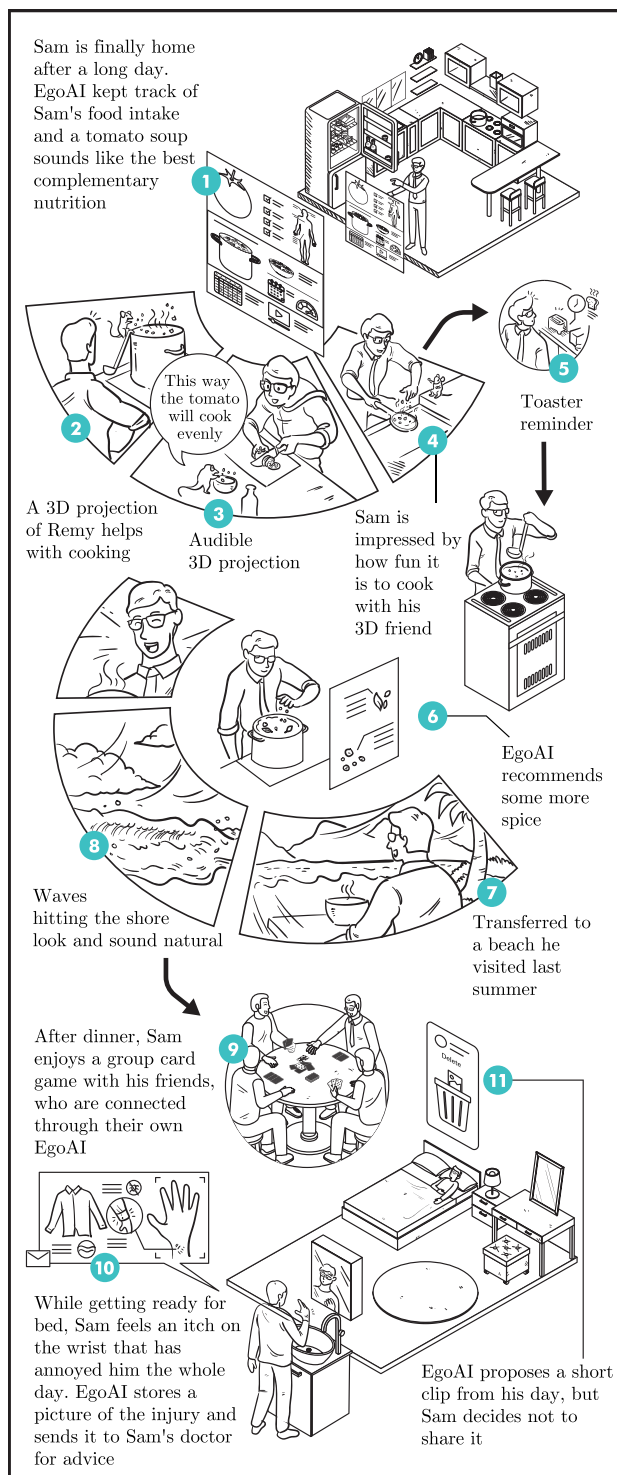
Fig. 1 EGO-Home. Character-based story envisaging the future of egocentric vision at home. Illustration of the story from Sect. 2.1. EgoAI assists Sam during dinner preparation and keeps him entertained with interactive and immersive experiences. **4.2** 3D Scene Understanding ① ② ③ ④ ⑦ ⑧ ⑨. **4.3** Object and Action Recognition ① ⑤ ⑥ ⑩. Measuring System ⑥. **4.11** Dialogue ⑥. **4.10** Summarisation and Retrieval ⑦ ⑪. **4.7** Full-body Pose, **4.8** Hand Pose and **4.6** Social Interaction ⑨. Medical Imaging ⑩. Messaging ⑩ ⑪

eating his hot soup. He laughs at the absurdity of the soup at the beach. Before heading to bed, Sam enjoys a group card game with his friends who recently moved to Australia. They are connected though their own EgoAI, which makes Sam feel as they are all physically present with him. He can hear the sound of the cards shuffling as his two friends appear seated around his table. EgoAI is a great game companion and points out a strategic move to make when he's about to play a suboptimal card.
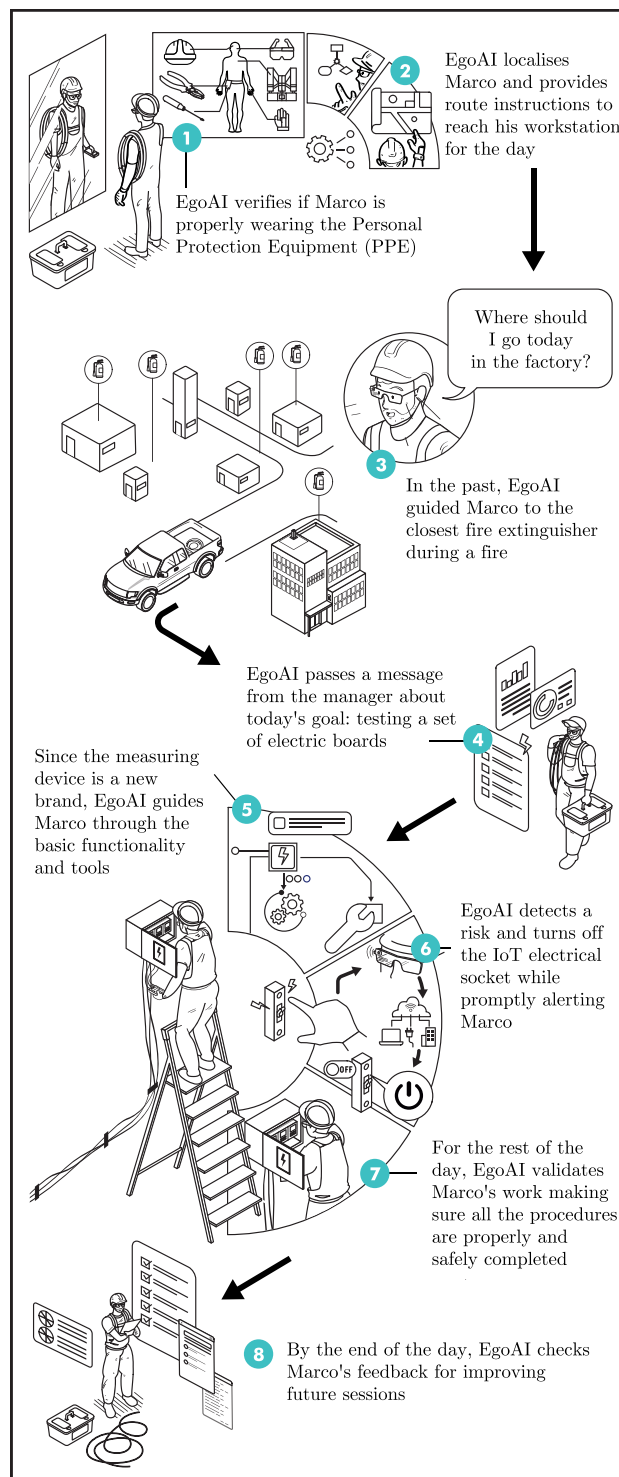
While getting ready for the night, Sam feels again that itch on the wrist that has annoyed him the whole day. EgoAI assures him that with high probability it is just the cuff of his new shirt that irritated the skin, but also offers to take care of this by sending a picture to his physician for advice. As Sam heads to bed, EgoAI proposes a short clip from his day that could be shared on social media, but Sam thinks "not today" and asks EgoAI to delete the post draft.

2.2 EGO-Worker

Current vision-based systems are being integrated in large scale workshops and factories, but these mainly rely on fixed cameras, which need to be installed in all the areas of interest and can only perceive a limited view of the scene, hence restricting their usefulness. Training and monitoring of workers is mostly offline through recorded material or over-the-shoulder advice from experienced workers. Often the knowledge is lost as one worker changes job. Feedback to workers about their performance is based on heuristic automatic or manual calculations and often does not correspond to actual performance. This is often disconnected from training and advice for how to improve performance. While technology is employed for workers' safety increasingly, this is below the levels expected with most technological advances focusing on improved productivity. EgoAI will fill this gap and make the lives of workers safer and more comfortable.

As with every morning, Marco (Fig. 2) begins his shift by looking at himself in the mirror: in this way EgoAI can verify if he is properly wearing the Personal Protection Equipment (PPE) which will guarantee his safety. After this check, Marco asks EgoAI where in the factory he is needed today. EgoAI localises Marco and provides route instructions to reach his workstation for the day, avoiding dangerous areas with suspended loads and the paths reserved for the transit of vehicles. Marco trusts EgoAI navigation abilities and always remembers that day when EgoAI swiftly guided him to the closest fire extinguisher to avoid flames spreading.

As Marco reaches his workstation, EgoAI passes a message from the manager about today's goal: testing a set of electrical boards. Since the hand-held measuring tool is a new brand, EgoAI guides Marco through the basic functionality useful to correctly test the boards. Unfortunately, Marco gets distracted and is about to probe the electrical board

EgoAI verifies if Marco is properly wearing the Personal Protection Equipment (PPE)

EgoAI localises Marco and provides route instructions to reach his workstation for the day

Where should I go today in the factory?

In the past, EgoAI guided Marco to the closest fire extinguisher during a fire

EgoAI passes a message from the manager about today's goal: testing a set of electric boards

Since the measuring device is a new brand, EgoAI guides Marco through the basic functionality and tools

EgoAI detects a risk and turns off the IoT electrical socket while promptly alerting Marco

For the rest of the day, EgoAI validates Marco's work making sure all the procedures are properly and safely completed

By the end of the day, EgoAI checks Marco's feedback for improving future sessions

Fig. 2 Character-based story envisaging the future of egocentric vision in industrial settings. Illustration of the story from Sect. 2.2. EgoAI assists Marco from the start of his day until its conclusion. Safety Compliance Assessment ①. **4.1** Localisation and Navigation ②⑤. Messaging ④. **4.8** Hand-Object Interaction ⑤. **4.4** Action Anticipation ⑥. Skill Assessment ⑦. **4.11** Visual Question Answering ⑧, **4.10** Summarisation ⑧

while it is still plugged in. EgoAI detects the risk and turns off the IoT electrical socket on which the board is connected while promptly alerting Marco.

For the rest of the day, EgoAI validates Marco's work making sure that all the procedures are properly and safely completed, answering his questions in case of doubts, and estimating his stress level to make sure that he takes breaks when needed.

By the end of the day, EgoAI thanks Marco for his very hard work particularly with all new procedures involved and checks on his feedback for better training. EgoAI sends this feedback on any misunderstandings and obstacles automatically to future training sessions and planning (Fig. 3).

2.3 EGO-Tourist

Today, travelling abroad for tourism and vacations has more than doubled in the past 20 years.[2] Technology and art, both ancient and modern, are becoming increasingly bounded, with the former increasing the spread and the possibility of interaction with the latter. Indeed, the use of technological tools such as digital audio guides or virtual tours is becoming predominant in museums and touristic sites with engagement being crucial to increasing the visitor's interest. Despite modern tools, the visitor experience still lacks a form of personalisation and necessitates active interaction from the user. EgoAI on the other hand, fills these gaps and makes travelling a fun and interactive experience.

Claire (Fig. 3) has just reached Turin as the last stop of her Italian holidays. She is thrilled to start her visit but does not know much about the city. Luckily, EgoAI is already tuned on Claire's tastes and prepared her a personalised and exciting 1-day itinerary. EgoAI knows Claire is a big fan of museums so suggests half a day to visit the famous local Egyptian one. During the visit, EgoAI activates the 3D projection of Cleopatra to guide and interact with Claire. Cleopatra leads Claire through the artworks and proposes her the most suited path. While Claire is asking Cleopatra information about a sarcophagus, she observes virtual elements being added to the scene which bring the artwork to life. Claire feels transported to ancient Egypt where she can manipulate and use the pieces as it was intended.

At the end of the visit, Claire decides to keep Cleopatra as her AR guide for lunch and asks her for a good pizza place. While enjoying her meal, Claire asks Cleopatra questions about famous Italian monuments she visited along her tour, augmenting her understand of the history behind them.

EgoAI has booked an afternoon at the thermal baths. As the next bus is not due for another 20 min, EgoAI suggests Claire a proper Italian coffee at a nearby cafè sided by a slice of bunet, a popular Turin-based dessert. Claire would
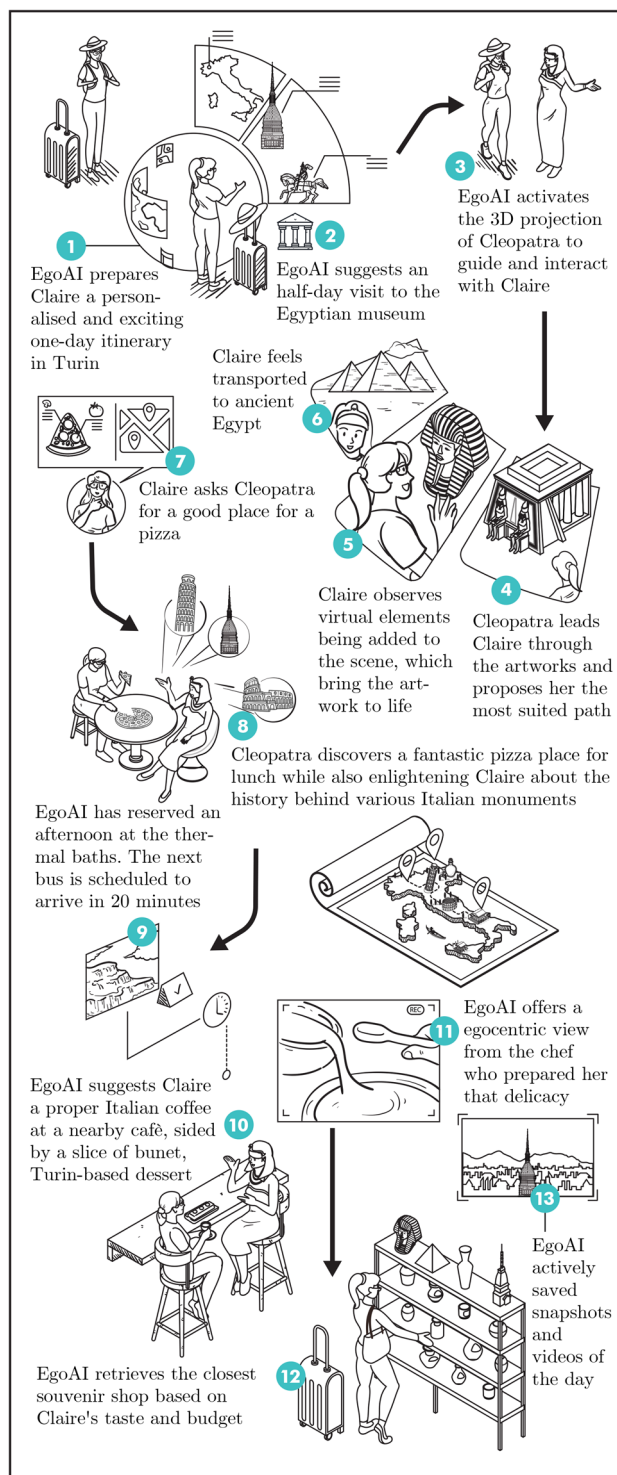
Fig. 3 Character-based story envisaging the future of egocentric vision in tourism. Illustration of the story from Sect. 2.3. EgoAI accompanies Claire throughout her itinerary in Turin. Recommendation and Personalisation ①②⑧⑨⑩⑪. **4.2** 3D Scene Understanding ②③④⑤⑥. **4.5** Gaze Prediction ⑤. **4.1** Localisation and Navigation ③④⑧⑫. Messaging ⑦. **4.11** Dialogue ⑧. **4.3** Action Recognition and Retrieval ⑪. **4.10** Summarisation ⑬

like to learn the recipe, so EgoAI offers a first-person view from the chef who prepared that delicacy earlier in the day.

After the thermal baths, EgoAI checks whether Claire is interested in buying some souvenirs for her family. EgoAI then retrieves the closest souvenir shop based on each relative's taste and the budget set by Claire.

Claire was engaged during her 1-day itinerary and did not worry about taking pictures. EgoAI actively saved relevant snapshots of the day, and videos of her favourite moments.

2.4 EGO-Police

In 2023, it has been almost 10 years since the adoption of body-worn cameras by several police departments around the world. Practical experience showed that they have a large potential in enhancing transparency and facilitating investigations, besides increasing officers' accountability and safety. Still, cameras provide only passive support to law enforcement, with data storage and post-processing analysis requiring a consistent time and cost effort. We can easily imagine how constables would benefit from AI-empowered wearable vision devices.

Judy (Fig. 4) is a police officer who uses EgoAI every day of her service. She finds it highly convenient: the device is much lighter than the usual equipment and serves as body camera, radio, phone and flashlight. Moreover, it makes her feel safe as she knows that EgoAI is constantly pinpointing her position and would send an alert to headquarters if she encounters unusual events or dangerous situations. For instance, last month Judy was assigned to a high-crime zone while searching for a suspect. EgoAI helped Judy navigate through the shortest safe path to several target places reported as possible hideouts. While patrolling the streets, one of the fellow officers shared via EgoAI a clip from a surveillance camera one block east: the suspect was moving in Judy's direction. Despite the crowds, EgoAI detected and re-identified the man before he passed Judy. She was able to swiftly arrest him.

Judy also appreciated the help of EgoAI when she had to manage an abandoned backpack at the airport. EgoAI accessed the lost-and-found database of the airport but no match was found. Then, from thermal and multi-spectral sensors, it calculated a low risk for explosive content and projected a clear red circle around the backpack with the minimal stand-off distance. EgoAI connected Judy with the bomb squad and live-shared the observed scene: the experts agreed with the initial evaluation and excluded any risk that the backpack could contain an explosive. Then, EgoAI guided Judy with exact instructions to grasp the backpack and open it. Luckily it was only containing a pair of old tennis shoes.
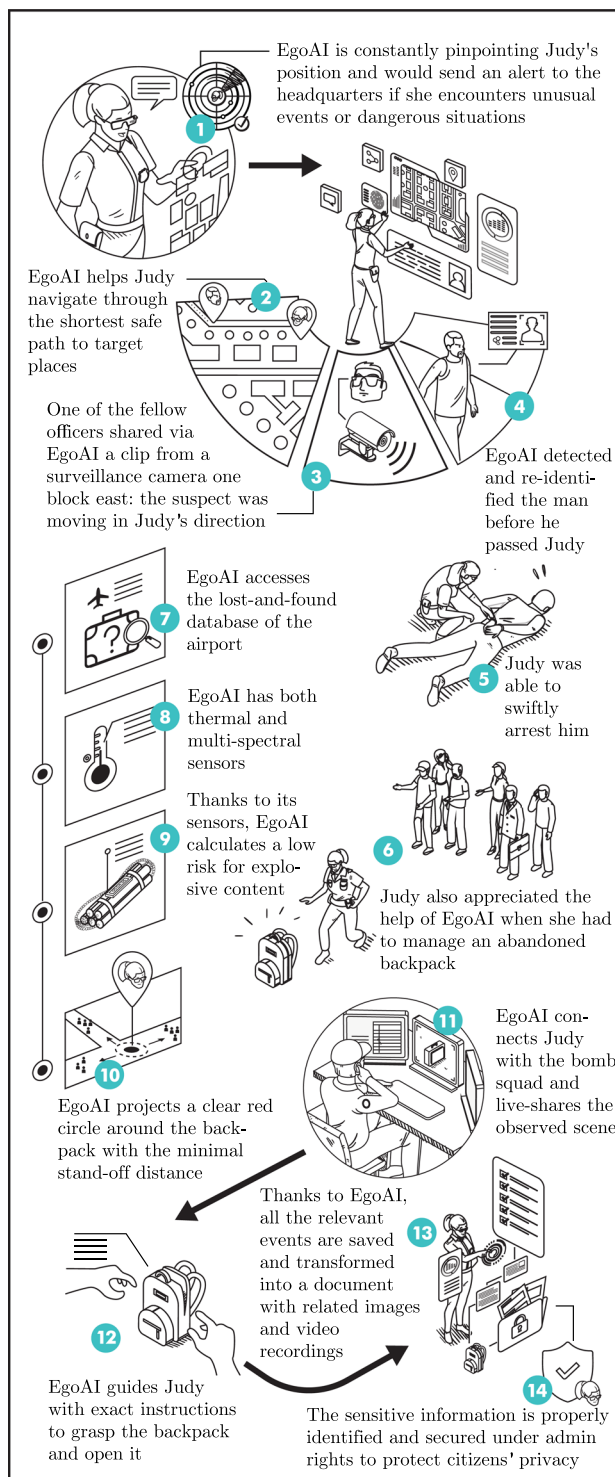
Fig. 4 Character-based story envisaging the future of egocentric vision within the police force. Illustration of the story from Sect. 2.4. EgoAI helps Judy, a police officer, during her day keeping her city safe. **4.1** Localisation and Navigation ①②. Messaging ①③⑪. **4.3** Action Recognition ②⑬. **4.9** Person Re-ID ②④. **4.3** Object Detection and Retrieval ⑦. Measuring System ⑧⑨. Decision Making ⑨. **4.2** 3D Scene Understanding ⑩. **4.8** Hand-Object Interaction ⑫. **4.10** Summarisation ⑬. **4.12** Privacy ⑭

At the end of every working day, Judy does not need to fill out any form or detailed reports. Thanks to EgoAI, relevant events are saved and transformed into a document with related images and video recordings. Importantly, the sensitive information possibly captured by EgoAI during Judy's work is properly identified and secured under admin rights to protect citizens' privacy.
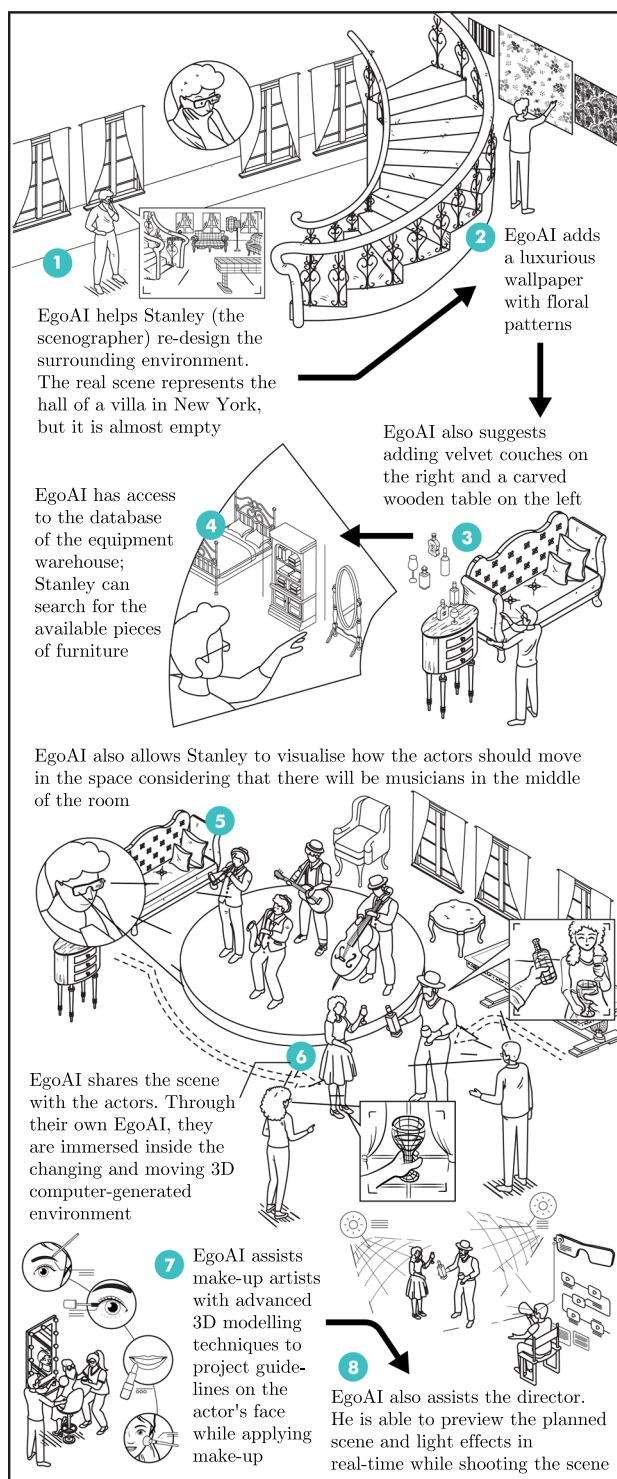
2.5 EGO-Designer

Nowadays, films are full of digital artefacts, not only sci-fi ones but also realistic dramas. These include fantasy environments as well as scenes that cannot simply be shot on-site. Current technology makes use of neon-green screens which are then removed using video editing software in post-production. However, this makes it difficult for the scenographer to visualise the final effects while shooting, and the actors must perform without having a full perception of the 3D scene around them. A movie production crew may largely benefit from egocentric devices for augmented reality (AR), digital rendering, and 3D modelling leading to a completely innovative way to experience the movie creation processes.

It is another hot day in Hollywood and Stanley (Fig. 5) has promised the movie director that the scenography will be ready first thing tomorrow. He is at the studios wearing EgoAI which is augmenting the surrounding environment: the real scene he is looking at is the reconstructed hall of a villa in New York during the 1920s. There is a fancy large spiral staircase but besides that, it is almost empty and should be designed to host a glamorous party.

EgoAI helps Stanley to virtually add a luxurious wallpaper with floral patterns and a ceiling adorned with intricate moldings. He adjusts the position of two digital chandeliers so that they cast a warm, golden glow across the room. EgoAI also suggests adding velvet couches on the right and a carved wooden table on the left with crystal decanters, champagne flutes, and a variety of liquor bottles. As EgoAI has access to the database of the equipment warehouse, Stanley can search for the available pieces of furniture which are most similar to what he has in mind so that the production assistants can position them in the scene. EgoAI also allows Stanley to visualise how the actors should move in the space around the musicians in the middle of the hall.

The scene is promptly shared with the actors. Through their EgoAI, actors are immersed inside the changing and moving 3D computer-generated environment so that they can visually engage with elements present in front of them and rehearse. Their natural acting is enhanced by their familiarity with the scene before shooting starts.

Stanley has also some suggestions for the make-up artists about the colour palettes that would stand out with the chosen lights. It will be very easy to share information with them

Fig. 5 Character-based story envisaging the future of egocentric vision in the entertainment industry, focusing on the perspective of scene and makeup designers. Illustration of the story from Sect. 2.5. EgoAI helps Stanley, the scenographer, and all the crew during movie production. **4.2** 3D Scene Understanding ①②③④⑤⑥⑦⑧. Recommendation ③. **4.3** Object Recognition and Retrieval ③④. **4.7** Full-body Pose Estimation ⑤⑥. **4.6** Social Interaction ⑥. **4.5** Gaze Prediction ⑥. **4.8** Hand-Object Interaction ⑦. Messaging ⑥⑧

as they are also using EgoAI with advanced 3D modelling techniques to project guidelines on the actor's face while applying the make-up.

At the end of the day Stanley feels satisfied and he is sure that his work will be appreciated: through EgoAI the director will be able to preview the planned scene and light effects in real-time while shooting the scene, without having to wait for playback. EgoAI has saved the industry millions of dollars, with repetitions of scenes dropping to one fourth compared to movies captured in the ancient era before EgoAI was introduced.

3 From Narratives to Research Tasks

Various research tasks can be identified in the above character-based narratives/stories. While some are only part of the future (e.g. particularly those related to augmented reality (AR)), others are currently achieved either via remote cameras (e.g. person identification) and Internet-of-Things devices (e.g. scene monitoring), or via smartphones (e.g. navigation). Despite their connectivity, local devices are typically restricted in coverage depending on where they are originally installed, while smartphones inevitably hinder interaction with the environment as they involve manual handling. It is our vision that most of the mentioned tasks will be seamlessly integrated into one egocentric device that we refer to as EgoAI in our stories. It will be person-centric, thus wearer-focused, and will also travel *anywhere* with the wearer.

In this section, we provide a mapping from the narratives above to research tasks as currently understood by the research community. We also examine whether these tasks can be performed using existing wearable devices or if new, more advanced and powerful ones are required to overcome the limitations of those currently available on the market. This sets the scene for the literature survey of the research tasks in Sect. 4.

For any task that involves *AR* technology, the need arises for in-depth **4.2** *3D scene understanding*. This is exemplified by the EGO-Home's augmented reality guides for cooking, EGO-Tourists' immersive museum experiences, and EGO-Designer's creation of imaginary scenes. Our envisaged AR is also endowed with directional audio synthesis, where auditory feedback enhances the realism of the augmented environment, as in the case of the mouse or the sounds of cards being shuffled in EGO-Home. To move within the 3D scene, **4.1** *localisation* and *navigation* emerge as recurrent tasks, both in the case of constrained spaces such as the factory in EGO-Worker as well as in open areas, as evident in EGO-Police's use of city maps. The abilities of current egocentric devices to perceive 3D scenes is continuously evolving due to the integration of additional environment

cameras (e.g., Microsoft HoloLens 2,[3] Xreal Light,[4] Magic Leap 2,[5] Project Aria Glasses[6]). These devices can scan and create a 3D model of the static environment to localise the wearer and allow them to navigate more easily. Dynamic as well as outdoor scenes still challenge these setups and this remains an active area of research for a realistic integration of 3D understanding in the future.

Inside the scene, high-level understanding of actions is carried out. Tasks like **4.3** *action recognition* undergo a paradigm shift with a transition in perspective from third-person to first-person view. In EGO-Worker, the device validates the user's actions in a workplace setting. Particularly noteworthy is **4.4** *action anticipation*, where the device can promptly prevent dangerous situations. There are currently no smart glasses on the market that are able to robustly recognise human actions in real time. Usually, data from the RGB camera and depth sensor of the glasses are collected and processed offline due to hardware limitations.

Equipped with **4.5** *gaze prediction*, EgoAI can track the user's eye movements and attend to objects seamlessly with their gaze. This capability is noted in both EGO-Tourist, where the user manipulates the artworks in the museum, and EGO-Designer, where the user virtually re-positions the objects in the virtual set. Nowadays, gaze tracking is a relatively stable feature but requires an eye calibration step before use and there remains the potential for drift over time. It has been implemented in wearable devices such as Microsoft Hololens2, Magic Leap 2, Project Aria Glasses and Apple Vision Pro.[7]

Analysing the social context of the camera wearer through **4.6** *social behaviour understanding* is also of significant importance. Social interactions are explored in EGO-Home, where users engage in interactive games with others connected through their devices. By employing **4.7** *body pose estimation* techniques from a first-person perspective, each user's pose is accurately reconstructed and seamlessly integrated into AR. Hands in particular are actively interacting with the environment and other individuals. **4.3** *action and object recognition*, **4.8** *hand-pose estimation* and *hand-object interactions* are key to EgoAI. In EGO-Police, the device intelligently comprehends user actions, providing precise instructions, like how to open the suspicious backpack. In EGO-Worker, EgoAI helps to operate a new measuring tool.

Recognising the user's identity and those of the bystanders plays a crucial role in social relationships. It has also a relevant role in security, going beyond what can be done with

---

[3] https://www.microsoft.com/en-us/hololens.

[4] https://www.xreal.com/light/.

[5] https://www.magicleap.com/magic-leap-2.

[6] https://about.meta.com/realitylabs/projectaria/.

[7] https://www.apple.com/apple-vision-pro/.

fixed cameras as in the case of **4.9** *person re-identification* described in the EGO-Police scenario. Of course, the identity, as well as users' data, should be properly safeguarded to ensure responsible use of the technology. As the wearable device can be in an "always-on" mode, it becomes imperative to address **4.12** *privacy* concerns and establish robust protection measures. Different laws regulate data protection and privacy in different countries, such as the General Data Protection Regulation (GDPR) in Europe, the California Consumer Privacy Act (CCPA), and the China Cyber Security Law (CCSL). However, the wearable glasses that are currently available in the market do not have strategies for compliance and it's left to the user to regulate the device by interactive privacy switches.

A related further issue is how to manage the ongoing and abundant stream of captured data that would be extremely costly to store in raw form. An efficient **4.10** *summarisation* and reporting process is clearly needed in multiple application scenarios. In both EGO-Home and EGO-Police, all the relevant events are saved and transformed into a report with images and video recordings. Identifying interesting events to memorise is also noted in EGO-Tourist. Thanks to EgoAI, it is also possible to *retrieve* relevant data or objects within a database by exploiting visual cues in both EGO-Police and EGO-Designer. In EGO-Worker, EgoAI can conduct *skill assessments* by monitoring whether the user correctly executes all the required procedures during their workday. Support in skill training is provided by the **4.11** *Visual Question Answering* application (VQA) that replies to Marco's questions by translating instructional videos into a step-by-step guide in the users' view.

By analysing the user's past data, the egocentric device can extract their preferences and offer *personal recommendations*. For instance, in EGO-Home, EgoAI proposes dinner recipes according to the user's preferences and eating history. Similarly, in EGO-Tourist, it suggests fitting lunch and shopping destinations based on the individual's taste.

The ability to solve several other side tasks will contribute to the success of the EgoAI device that we foresee. *Messaging* capability is recurrent throughout the stories. In EGO-Home, the user can send a picture of his wrist to the doctor, in EGO-Worker, the user received a message from the manager about his daily tasks, and in EGO-Police, EgoAI is capable of sending alerts to headquarters. This hands-free convenience is further enhanced by voice commands, allowing seamless interaction, as in EGO-Tourist when the tourist asks for additional information on the artwork. Some wearable glasses integrate vocal assistants such as Cortana,[8] Siri,[9] or Google Assistant,[10] which can interact with the device to

open applications, take photos, send messages, and more, clearly improving the user experience.

EgoAI also functions as a *measuring* instrument. In EGO-Home, EgoAI can quantify the amount of spice in the soup by leveraging its memory of the quantity previously added to the pan, or measuring the thickness of the soup from its visual appearance. Thanks to the possibility of integrating multiple sensors, such as thermal and multi-spectral cameras, it can also compute the risk of explosive content in EGO-Police by a *decision making* process. Wearable devices can also be integrated with advanced *medical imaging* techniques, enabling EgoAI to assess the severity of the condition from a picture in EGO-Home. Another assessment expected of EgoAI is related to *Safety Compliance Verification*. In EGO-Worker, EgoAI assesses whether the worker is correctly wearing a Personal Protection Equipment (PPE) through sophisticated recognition and identification techniques.

Currently, there are no devices in the market that can match the advanced features and capabilities of EgoAI. They also have strong hardware limitations that do not allow prolonged use. Even though some advanced wearable glasses provide complex and highly accurate features, the battery life is often only a few hours. This is even shorter when videos are captured constantly, as required to enhance the potential of summarisation techniques which also require a lot of computational power and memory, which can quickly drain the battery.

In this paper, we do not consider AR-specific approaches, which have their base in the computer graphics literature, but point the interested readers to recent surveys on the topic (Devagiri et al., 2022; Dargan et al., 2023) as well as a structured literature survey by Cipresso et al. (2018) and a survey of AR usability studies by Dey et al. (2018). We also exclude tasks that require perception or synthesis of audio, independent of the video—this includes speech and audio-only event perception. We are not aware of a recent survey on the topic and encourage researchers with relevant expertise to further explore this crucial modality. Moreover, we do not review system-based tasks such as personalised recommendations or measuring devices, as well as tasks related to assessment, whether for medical purposes or skill. We refer the reader to works on action assessment (Doughty et al., 2019; Parmar & Morris, 2019; Li et al., 2019b; Yu et al., 2021) and risk warning (de Santana Correia & Colombini, 2022).

Instead, in Sect. 4, we focus on a subset of all the aforementioned tasks—those that require visual understanding. We order the considered computer vision tasks from the most static to ones that respond to user engagement, primarily: scene-level understanding tasks—*localisation* and *3D understanding*, followed by tasks at the action level—*action anticipation*, *action and object recognition* and *gaze understanding and prediction*. Then, we review tasks around understanding people, particularly, *social behaviour understanding*,

---

[8] https://www.microsoft.com/en-us/cortana.

[9] https://www.apple.com/siri/.

[10] https://assistant.google.com/.

full-body pose, *hand and hand-object interaction*, and *person identification*. Moreover, we note two user engagement tasks that are recurring frequently in our narrative stories—*summarisation* and *dialogue*. Finally, we introduce *privacy* and the related approaches to preserve sensitive content captured by wearable devices. Overall, given the multi-modality nature of these tasks, we will also discuss how vision can be integrated with cues from other sensors. We visualise the connections between our use cases and these tasks in Fig. 6.

4 Research Tasks and Capabilities

For each of the egocentric vision tasks identified in Sect. 3, we now provide a structured literature review with dedicated subsections. Rather than covering the full progress of the field, we find it most informative to focus on *seminal works* that initially defined the task or changed its course as well as *state-of-the-art methods* that are currently achieving best performance. We acknowledge there are tens of works that paved the path from those seminal works to current methods but opt for not including them in this paper. We encourage interested readers to explore these intermediate works to understand the full progress of the field in each research task. Additionally, we note *datasets* specifically designed to advance the research in each of these tasks. We leave the review of more general datasets to Sect. 5. We conclude each subsection with a short reflection on the gap between current state-of-the-art and anticipated future.

4.1 Localisation

We divide localisation works into two categories: *visual place recognition* (Sect. 4.1.1) and *visual localisation* (Sect. 4.1.2). Both contribute to the broader goal of positioning the camera wearer within the surrounding environment using visual data for scene understanding and navigation, but they differ in their primary objectives. Place recognition gives a coarse estimate of 2D coordinates, whilst visual localisation determines the 6-DoF (Degrees of Freedom) of the camera pose. We also review Simultaneous Localisation and Mapping (SLAM) techniques (Sect. 4.1.3)—simultaneously building a map of unknown indoor or outdoor environments and tracking the position or trajectory of the camera.

Note that this task only differs marginally between wearable cameras, hand-held cameras and remote cameras (third-person). Additionally, cameras mounted on vehicles share similarities with wearable devices that lie in the viewpoint and perspective from which visual information is captured. These analogies allow us to broaden the scope of existing approaches beyond those exclusive to wearable devices.

4.1.1 Visual Place Recognition

Visual place recognition analyses visual cues, from either a single image or a sequence of images, to determine the place or area being observed. In egocentric vision, this relates to "contextual awareness", i.e., extracting knowledge of the user's surrounding. The most commonly used metric for evaluation is the *Recall@N*, which calculates the percentage of relevant or true positive places that are among the top *N* retrieved results. In other words, it measures how many of the correct places were successfully recognised within the top *N* ranked places.

Seminal works The early investigations of the problem of recognising the user's location from wearable devices date back to the late 90s, when image-based localisation has been mostly studied as a classification problem. Starner et al. (1998) proposed a context-aware system for assisting users while playing the "patrol" game, by recognising the room in which the player is operating. Aoki et al. (1998) presented an image matching technique for the recognition of previously visited places. Torralba et al. (2003) introduced a wearable system capable of recognising familiar locations and categorising new environments into high-level classes such as offices and corridors. They proposed to use that information as priors for object recognition (e.g., tables are more likely to exist in an office). Furnari et al. (2016) performed temporal segmentation of egocentric videos to highlight the continuous presence of the wearer in pre-defined personal locations. The work uses personal locations as cues for identifying activities.

Related to visual place recognition is the problem of visual geolocalisation—estimating the position where a given image or frame in a video was taken by comparing it with a large database of images from known locations. Visual geolocalisation is commonly approached as an image retrieval problem, with a retrieved image deemed correct if it is within a predefined range from the query's ground truth position. Jégou et al. (2010) proposed VLAD (vector of locally aggregated descriptors), an image descriptor derived from sift descriptors, bag of works and fisher kernels. Gálvez-López and Tardos (2012) presented a fast and efficient approach for place recognition using binary descriptors. A few years later, Arandjelovic et al. (2016) offered the first CNN-based approach for place recognition with weak supervision. From that work on, all methods have been using learned embeddings with some form of aggregation or pooling.

The combination of GPS and visual information to localise users in an environment has also been investigated. Capi et al. (2014) proposed an assistive system able to guide visually impaired people in urban environments, and Ahmetovic et al. (2016) proposed a smartphone app which can perform accurate and real-time localisation over large spaces.
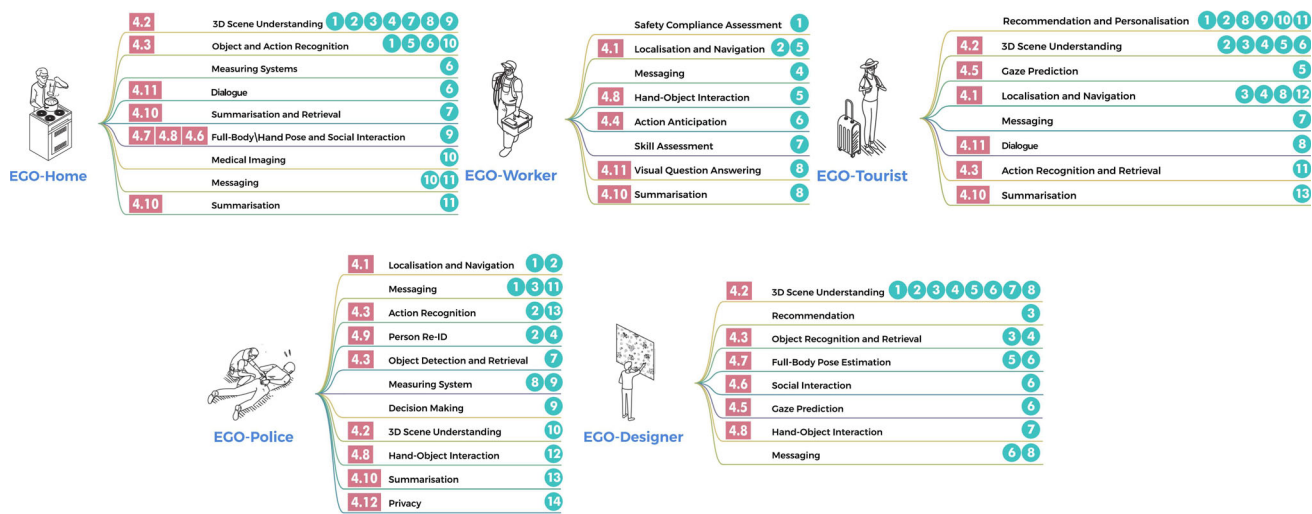
Fig. 6 Illustration of the connections between our narratives and the research tasks. For each of the use cases presented in Sect. 2, we show the corresponding research tasks, along with the specific part of the story where the tasks are occurring, indicated by the numbers corresponding to those representing sub-stories in Figs. 1, 2, 3, 4, and 5, respectively

State-of-the-art papers Current literature has shifted focus towards developing methods specifically tailored for visual geolocalisation. Most recent works aim at better training time scalability to exploit large-scale data. Berton et al. (2022) introduced CosPlace, a method that uses a classification task as a proxy to train the model that is used at inference to extract discriminative descriptors for retrieval. Zhu et al. (2023c) proposed R^{2}Former, a place recognition architecture that builds on the success of vision transformers and fuses multi-level attention information to generate global and local descriptors which are used for re-ranking. MixVPR by Alibey et al. (2023), is a new feature aggregation technique that takes in input feature maps from pretrained networks, and iteratively combines them using a stack of multi-layer perceptrons in a cascade of feature mixing.

4.1.2 Visual Localisation

Visual localisation refers to the process of determining the pose (position and orientation) of a camera with respect to a known 3D scene or environment, based on visual information. Approaches for visual localisation divide into hierarchical localisation pipelines, consisting of image retrieval, local feature extraction and matching. These are followed by 2D-3D correspondence mapping and pose estimation, and absolute pose regressors, that estimate the camera pose with a single forward pass, using only the query image. A commonly used metric for evaluating visual localisation tasks is the average of median position and orientation errors in meters and degrees, respectively.

Seminal works The work of Irschara et al. (2009) explored the transition from point cloud-based reconstruction to efficient feature-based localisation via Structure-from-Motion

(SfM). After computing a representative set of 3D point fragments that cover a 3D scene from arbitrary viewpoints, they matched directly the pose of the query image. The last stage uses the resulting 2D-3D matches for pose estimation using Random Sample Consensus (RANSAC) algorithm. Sattler et al. (2011) made significant contributions by introducing an efficient and direct matching approach between 2D query images and 3D reference data. Kendall et al. (2015) presented a deep learning-based approach for camera localisation. Their Convolutional Neural Network (CNN) architecture, called PoseNet, enabled real-time and accurate estimation of camera poses in 6-DOF, by regressing the 6-DoF camera pose from a single RGB image in an end-to-end manner with no need for additional engineering. Blanton et al. (2020) extended pose regression to multiple scenes by proposing the Multi-Scene PoseNet, where the network first classifies the particular scene related to the input image, and then uses it to index a set of scene-specific weights for regressing the pose. Also, the work of Sattler et al. (2016) contributed to large-scale image-based localisation by introducing an efficient and effective prioritised matching algorithm.

State-of-the-art papers Shavit et al. (2021) presented a novel approach using transformers for multi-scene pose regression. The approach uses encoders to focus on pose-informative features and decoders to transform encoded scene identifiers to latent pose representations. Generally, algorithms for visual localisation mostly rely on complex 3D point clouds that are expensive to build, store, and maintain over time. Do et al. (2022a) trained a CNN to detect the appearance of a sparse set of 3D scene points (scene landmarks), and showed that those predicted landmarks can yield accurate pose estimates, while being privacy preserving and requiring low data storage. Panek et al. (2022) explored dense 3D scene

models as an alternative to the sparse Structure-from-Motion point clouds as they are more flexible than SfM-based representations and can be rather compact. Moreover, storing the original images and extracting features when needed takes up less memory than storing the features.

4.1.3 Simultaneous Localisation and Mapping (SLAM)

SLAM is a technique used to build a map of an unknown environment while simultaneously estimating the camera pose within that environment. In this section, we focus on Vision SLAM (V-SLAM), which refers to those SLAM systems which use cameras as the main input sensors. In general, V-SLAM algorithms have three steps: initialisation, tracking, and mapping. The initialisation determines the global coordinates and builds an initial map. The tracking step involves the continuous estimation of the camera pose. In general, during this stage the algorithm extracts 2D-3D correspondences between the current frame and the map. Finally, the mapping step results in a sparse, semi-dense, or dense 3D reconstruction. SLAM algorithms can be mainly classified into two categories: feature-based and direct. Feature-based methods rely on sparse features for tracking, with the correspondences being used to refine poses through Structure-from-Motion techniques. Direct methods use the sensor data without preprocessing, estimating camera poses within an expectation maximisation framework. The most commonly used metric is the Root Mean Square Error (RMSE), which measures the difference between estimated and ground truth camera poses and map points, providing an overall indication of accuracy. *Seminal works* The first applications of SLAM to wearable cameras are from Davison (2003) and Mayol et al. (2005). Davison (2003) proposed a general method for real-time, single-camera V-SLAM and studied its application to the localisation of a wearable robot with active vision. The approach proposed by Mayol et al. (2005) enables prolonged periods of focused attention on specific areas of interest, followed by deliberate and controlled redirection of gaze to different parts of the scene. This reduces the need for frequent feature initialisation, and enhances overall system robustness. Castle et al. (2010) used monocular SLAM and object recognition for AR. Badino and Kanade (2011) introduced a head-wearable stereo system for structure and motion estimation. Alcantarilla et al. (2012) developed a wearable stereo system that combines SLAM with dense scene flow estimation to segment moving objects in the scene. Murillo et al. (2012) proposed to use wearable omnidirectional vision systems to augment people's navigation and recognition capabilities. Their approach involves accurate ego-motion estimation and topological/semantic localisation, enabling precise user guidance.

One problem of monocular SLAM is *scale drift*. It occurs during the initialisation of monocular SLAM, where the scale is initially set to a real or arbitrary value. However, as the camera moves and old landmarks are lost while new ones are initialised, the scale of the scene changes continuously. To address this issue in large environments, Gutierrez-Gomez and Guerrero (2016) proposed an approach that computes the true scale dynamically using visual odometry estimates from wearable single cameras. Their method relies on the characteristic oscillatory movement of the human body during walking to extract scale information, making it particularly suitable for wearable systems.

The nature of egocentric videos, characterised by sharp head rotations and predominantly forward motion, leads to rapid changes in the camera view, resulting in short and noisy feature tracks. Additionally, the dominant 3D rotation caused by natural head motion further reduces parallax, leading to triangulation errors. To address these issues, Patra et al. (2017) proposed a fast and robust egomotion estimation method for egocentric videos, using a local loop closure technique aligned with the wearer's head motion.

Suveges and McKenna (2021) proposed a semantic, non-geometric, human-centred form of SLAM, by constructing a representation of a user's everyday environment in terms of locations that they frequent and their patterns of transition between, and their behaviours within those locations. *State-of-the-art papers* With the rise of AR applications, achieving precise alignment of virtual content with the user's physical surroundings has become crucial. To accomplish this, modern devices are equipped with a range of sensors. One notable example is the HoloLens, which incorporates four tracking cameras and a time-of-flight range camera. One of its key features is spatial mapping, which allows the device to create a detailed map of its surrounding environment (Hübner et al., 2020). Using spatial mapping, the HoloLens scans the area within a 70-degree cone, capturing depth information from distances between 0.8 and 3.3 m. Based on the data, it reconstructs a mesh representation of the observed scene, which serves as a foundation for accurately placing virtual objects within the real world. Meta's Aria glasses have been also recently released with multiple sensors such as stereo cameras, dual inertial measurement units, spatialised microphones, eye tracking cameras and more. They make use of localisation and mapping techniques to build "LiveMaps", a virtual 3D representation of the world.

The combination of neural radiance fields (NeRF, Mildenhall et al. (2021)) and SLAM has also emerged as a recent trend. By utilising the capabilities of SLAM for accurate pose estimation and dense depth maps, together with the power of NeRF, it is possible to generate real-time neural scene representations (Rosinol et al., 2023). Haitz et al. (2023) proposed an acquisition pipeline that enables real-time image and pose streaming through a TCP client-server application, allowing simultaneous training of Instant-NeRF. The HoloLens acts as the image and pose server, while the client applica-

tion receives the images and writes them into a GPU image buffer. Instant-NeRF model is incrementally trained using the incoming image stream. Additionally, a fast geometric reconstruction of the scene is performed by querying the trained network based on sample rays from the training poses.

Datasets For *visual place recognition*, Furnari et al. (2016) collected a dataset of egocentric videos containing 10 personal locations of interest. More recently, Milotta et al. (2019) collected and publicly released a dataset of egocentric videos asking 12 subjects to freely visit a natural site with a total of 6 h of recording. Ragusa et al. (2020a) proposed a dataset of egocentric videos for visitor behaviour understanding, including 27 h of video acquired by 70 subjects, with labels for 26 environments which allow room-level localisation.

In *visual geo-localisation*, all previous datasets capture an autonomous driving viewpoint which is very different from that of a wearable camera or are collected using hand-held devices. This lacks characteristic head-mounted motion patterns. Up to our knowledge, no dataset is available for *visual geo-localisation* from a body-worn camera.

For *visual localisation*, Sarlin et al. (2023) created a dataset using Meta's Aria glasses at 3 locations in Seattle (Downtown, Pike Place Market, Westlake). In each location, they recorded 3 to 5 sequences following the same trajectories, for a duration of 5 to 25 min varying by location, and a total of 3 h of recordings. Each device is equipped with a GPS sensor, IMUs, grayscale SLAM cameras, and a front-facing RGB camera.

Suveges and McKenna (2021) proposed the first dataset specifically designed for SLAM applications on egocentric vision. Five videos were recorded using a head-mounted GoPro Hero 4, for a total of 4 h of videos including transition segments between locations, repeated visits by a user to multiple distinct locations, and unique labels for all visited locations.

Multiple sensors data, such as depth images, hand and eye tracking data, are essential for accurate spatial mapping and scene understanding for XR applications. Chandio et al. (2022) proposed HoloSet, a dataset captured using Microsoft Hololens 2, that contains the raw synchronised data streams from the following sensors: depth, RGB, four grayscale visible light tracking (VLC) cameras, and an IMU, along with the ground truth pose trajectory. It contains 29 sequences and 78.5k samples that cover more than 6200 m. Sarlin et al. (2022) introduced a large-scale dataset of over 100 h and covering 45'000 square meters of multi-sensor data streams (images, depth, tracking, IMU, BT, WiFi) captured using HoloLens 2 and iPhone/iPad devices in diverse environments, including a historical building, a multi-story office building, and part of a city center. Data include indoor and outdoor images with varying illumination, semantic changes, and dynamic objects.

Importantly, all previous datasets were collected specifically for localisation purposes. In these recordings, the camera wearer is only navigating the scene to capture these sequences and is not carrying out any of their daily tasks necessarily. It is acknowledged to be challenging to perform visual localisation from unscripted egocentric footage of actual activities (Suveges & McKenna, 2021). Recently, Tschernezki et al. (2023) provided 6 DoF camera positions for 99 h of the EPIC-KITCHENS dataset (Damen et al., 2022) in 45 home kitchens. Camera estimates are achieved through intelligent sampling without any additional sensors or sequences specific for localisation. However, no ground truth is available for this dataset and these camera estimates are only qualitatively evaluated.

For the future Despite progress made in recent localisation techniques for robotics and autonomous vehicles applications (Kazerouni et al., 2022; Cheng et al., 2022), the robustness of these algorithms in dynamic and changing environments as the ones captured by wearable devices require further development. For visual place recognition, current state-of-the-art performance are 64.0% recall on the Mapillary challenge. On LaMAR (Sarlin et al., 2022), the recent benchmark for localisation and mapping in the context of AR, results on single-frame localisation only achieve 45.6% / 61.3% recall at $(1°, 10 \text{ cm})/(5°, 1 \text{ m})$. Additionally, wearable devices often have limited computational resources, which can limit the complexity and accuracy of localisation algorithms. The most attractive application for localisation in head-mounted devices is AR, where the objective is to place virtual content in the physical 3D world, persisting it over time, and sharing it with other users.

Common benchmarks over the last years often rely on limited datasets with minimal scene diversity and sensors. These datasets also are typically collected specifically for localisation, through navigation-only sequences rather than capturing individuals engaged in actual activities. However, ongoing research efforts and advancements in computer vision, sensor technologies, and wearable computing such as Meta's Project Aria glasses and Microsoft HoloLens are paving the way for future applications of localisation on wearable devices, enabling promising use cases such as indoor navigation and AR experiences.

4.2 3D Scene Understanding

The goal of 3D scene understanding is to build an AI agent able to interpret the surrounding environment and explore possible interactions with it. This also involves identifying relevant objects in the scene and reasoning on their locations. The complexity of the field has attracted attention over the last few years, leading to the proposal of numerous sub-tasks and datasets. Their diversity underscores the multifaceted nature of 3D scene understanding, prompting researchers to

explore various evaluation measures tailored to specific challenges.

Seminal works The first work to explore task-relevant objects in 3D is that of Damen et al. (2014). Given a mapped environment, gaze estimation is used to cluster interaction regions into task-relevant objects and their modes of interaction. For studying human-centric interactions with the environment, Bertasius et al. (2015) proposed to utilise egocentric stereo cameras to establish an egocentric object prior within a first-person view RGBD frame, which could then be employed for 3D saliency detection. Through observations, it was discovered that humans possess a fixed size prior to salient objects, indicating that salient objects in 3D undergo consistent transformations, enabling people's visual system to perceive them with an approximately constant size. This insight led to the identification of a consistent egocentric object prior that can be characterised by its shape, size, depth, and location within the first-person view. Rhinehart and Kitani (2016) focused on learning and predicting "Action Maps" that encode the user's ability to perform activities at various locations. By mapping actions to specific regions within a scene, this technique enables the understanding and prediction of human activities in a given environment. Li et al. (2022) focused on anticipating as early as possible the target location of a person's object manipulation action in a 3D workspace. While this is a special case of trajectory forecasting, the latter is infeasible in manipulation scenarios and the hands often are located outside the field of view. Therefore, focusing on predicting the 3D target location gives a better understanding of possible interactions with objects, useful for applications such as robot planning and control. Recently, Grauman et al. (2022) proposed the task of Visual Queries with 3D Localisation (VQ3D), which focuses on retrieving the relative 3D localisation of a query object with respect to a current query frame. Another interesting problem has been proposed by Majumder et al. (2023): building the map of a previously unseen 3D environment by exploiting shared information in the egocentric audio-visual observations of participants in a natural conversation. Finally, Pan et al. (2023a) introduced the task of collision prediction and localisation from unposed egocentric videos, which aims at predicting *when* and *where* a collision with the environment might occur.

State-of-the-art papers Nagarajan and Grauman (2020) introduced a reinforcement learning approach where an embodied agent autonomously discovers the affordance landscape in new, unmapped, 3D environments, enabling interaction exploration. They rewarded the agent for quickly interacting with all objects in an environment and trained an affordance model online to segment images according to the likelihood of each of the agent's actions succeeding. Do et al. (2022b) focused on predicting depths and surface normals of the surrounding environment from a single view egocentric image. They addressed challenges derived from the use of

wearable devices such as tilted images and the presence of dynamic foreground objects by proposing an image stabilisation method which transforms tilted images to a canonical orientation for better learning. Nagarajan et al. (2023) proposed learning environment-aware video representations that encode the surrounding physical space, facilitating the prediction of local environment states at different time-steps. These states are used to train a transformer-based video encoder model, which gathers visual information from the entire video and constructs an environment memory. This memory can then be accessed to predict the local state at any specific point in the video.

Liu et al. (2022a) proposed the task of jointly recognising and localising actions of a user on a known 3D map from egocentric videos. They proposed a novel deep probabilistic model that utilised a Hierarchical Volumetric Representation (HVR) of the 3D environment and an egocentric video to infer the 3D action location and recognise the action based on contextual cues. Other works focused on object visual query localisation in the 3D space. Xu et al. (2023a) proposed a transformer-based module that incorporates object-proposal set context while considering query information. Mai et al. (2023) formalised a pipeline that better integrates 3D multiview geometry with 2D object retrieval from egocentric videos, leading to improved camera pose estimation and substantially improved VQ3D performance. The process involves three main steps: first, a sparse 3D reconstruction is performed using Structure from Motion (SfM) to estimate 3D poses and create a sparse 3D map. Second, the frames of an egocentric video and a visual crop of a query object are fed into a model that retrieves response frames and their corresponding 2D bounding boxes. Third, for each response frame, the depth is estimated and the object centroid is back-projected to 3D using the corresponding camera pose. Qian and Fouhey (2023) addressed the task of predicting the 3D location, physical properties and affordance of objects from single images. Given a set of query points, the output includes the potential 3D interaction, in terms of movable, location, rigidity, articulation, action and affordance. They achieve that using a transformer-based model which builds on a detection backbone.

Datasets General-purpose egocentric datasets such as EPIC-KITCHENS (Damen et al., 2022) and Ego4D (Grauman et al., 2022), which are reviewed in Sect. 5, can be used for scene understanding. Additionally, other task-specific datasets have been proposed. The Egocentric Depth on everyday INdoor Activities (EDINA) dataset by Do et al. (2022b) has the goal of facilitating learning the visual representation of dynamic egocentric scenes. It comprises more than 500K synchronised RGBD frames and gravity directions captured from an egocentric viewpoint with diverse daily activities, for a total of 16 h RGBD recording. EgoPAT3D (Li et al., 2022) is a large multimodality dataset of more than a mil-

lion frames of RGB-D and IMU streams, which has been designed for the task of anticipating the target location of a person's object manipulation action in a 3D workspace. The total collection contains 150 recordings, 15 household scene point clouds, 15,000 hand-object actions, 600 min of raw RGB-D/IMU data, 0.9 million hand-object action frames, and 1 million RGB-D frames for the entire dataset. Qian and Fouhey (2023) introduced the 3D Object Interaction Dataset (3DOI), with Internet videos, egocentric videos and indoor images. For the egocentric part, they sampled 2K images from EPIC-KITCHENS (Damen et al., 2022). Images come with 3D ground truth, including depth and surface normals, and 5 interactable query points, including both large and small objects. For each of them, they annotated whether the object is movable, its location, its rigidity, its articulation, the potential action that can be done with it, and its affordance (where it is possible to interact with the object).

The Aria Digital Twin (Pan et al., 2023b) is an egocentric dataset captured using the Aria glasses that contains 200 sequences of real-world activities conducted by Aria wearers in two real indoor scenes with 398 object instances (324 stationary and 74 dynamic). Each sequence includes raw data of two monochrome camera streams, one RGB camera stream, two IMU streams, complete sensor calibration, ground truth data including continuous 6-degree-of-freedom (6DoF) poses of the Aria devices, object 6DoF poses, 3D eye gaze vectors, 3D human poses, 2D image segmentations, image depth maps and photo-realistic synthetic renderings. Ravi et al. (2023) proposed ODIN (the OmniDirectional INdoor dataset), a large-scale dataset of more than 300K omnidirectional images capturing a diverse range of activities of daily living. This includes scans of the recording environments from a 3D scanner and camera-frame 3D human pose estimates, enabling its use for scene understanding purposes. Recently, Tschernezki et al. (2023) released EPIC Fields, an augmentation of EPIC-KITCHENS with 3D camera poses. It reconstructs 96% of videos in EPIC-KITCHENS, registering 19M frames in 99 h recorded in 45 kitchens, creating an opportunity to bring 3D geometry and video understanding closer together. Mur-Labadia et al. (2023) built a dataset on affordances based on the EPIC-KITCHENS dataset, EPIC-Aff, which provides interaction-grounded, multi label, metric and spatial affordance annotations. Finally, Shapovalov et al. (2023) introduced Replay, a collection of multi-view, multi-modal videos of humans interacting socially. It contains long scenes in an indoor environment, each captured in 4K resolution using 8 static DSLR cameras and 3 head-mounted GoPro cameras, along with a comprehensive microphone array. In total, it contains 66 h of footage. It is suitable for a series of tasks, such as novel-view audio/visual synthesis and 3D reconstruction.

For the future Egocentric videos provide a natural connection between the activities of the camera wearer and the surrounding 3D spatial context. Although this is an intrinsic characteristic of egocentric vision, However, motion blur, and unusual viewpoints caused by how egocentric videos are captured introduce overwhelming challenges, causing 3D reconstruction to struggle with dynamic content. As a result, much work remains before we can have a 3D understanding of dynamic phenomena, such as actions and activities. Another promising future direction is working with both egocentric and exocentric views. By combining the insights gained from both perspectives, researchers could potentially unlock a more comprehensive understanding of complex scenes and human interactions. This approach however is limited in its applicability for our anticipated EgoAI future, where exo views are unlikely to be part of our everyday lives.

4.3 Recognition

Recognition in egocentric vision is crucial as it involves understanding interactions as well as the objects the wearer interacts with and their actions. This dual focus on both actions and objects enables a comprehensive understanding of the wearer's environment and activities. We divide the works into *action* (Sect. 4.3.1) and *object* (Sect. 4.3.2) respectively.

4.3.1 Action Recognition

The goal of egocentric action recognition is to classify human actions from the egocentric point of view, i.e., the person wearing the camera is carrying out the action. In areas such as robotics and AR, egocentric action recognition is critical to enable downstream applications, such as contextual recommendations or reminders. The egocentric point of view and a wearable, hence moving in dynamic and often unpredictable ways, camera presents an higher level of complexity when compared to standard action recognition from a fixed and remote cameras. Moreover, as the camera wearer themselves are largely out of the field of view, several challenges come from the partial observability of the main actor.

One possible way to address this is to leverage complementary cues to support the visual modality. Audio, gaze and temporal dynamics via optical flow are examples of information that play a relevant role in understanding the performed actions. As managing multiple modalities may be costly, recent advancements are focusing on low-energy consumption architectures and higher-level action understanding. The task is formalised as a classification problem and generally evaluated with top-1 and top-5 accuracy.

Seminal works Early works considered the egocentric perspective to improve action recognition for robots (Johnson & Demiris, 2005) and humans (Surie et al., 2007). Spriggs et al. (2009) explored action recognition for egocentric vision with Inertial Measurement Units (IMUs) used for tempo-

rally identifying the actions. Kitani et al. (2011) authored a pioneering work about tackling action recognition from egocentric sports videos in an unsupervised manner. The research field gained large momentum after the introduction of the dataset collecting activities of daily living (ADL) (Pirsiavash & Ramanan, 2012), particularly thanks to its large set of annotations on activities, object tracks, hand positions, and interaction events. To deal with complex object interactions and long-range temporal activity structures, the authors also introduced tailored representations that included temporal pyramids and composite object models.

The work by Fathi et al. (2012b) was the first to highlight the utility of gaze: it presented a probabilistic generative model for simultaneously recognising daily actions and predicting gaze locations from egocentric videos. Li et al. (2015) proposed to combine features encoding hand pose, head motion and gaze direction together with motion and object features coming from local descriptors.

In the last years, deep learning has alleviated the burden of manually defining features. Singh et al. (2016b) was the first work to use CNNs for end-to-end learning and classification of the wearer's actions. Since then, the attention moved to learning architectures with novel pooling mechanisms (Ryoo et al., 2015) or temporal convolutions on motion fields for long-term activity recognition (Poleg et al., 2016).

Techniques that use recurrent neural networks such as Long Short-Term Memory (LSTM) (Cao et al., 2017; Verma et al., 2018) and Convolutional Long Short-Term Memory (ConvLSTM) (Sudhakaran & Lanz, 2017, 2018) have been proposed to better encode temporal information. Sudhakaran et al. (2019) proposed a new recurrent neural unit that augments LSTM with built-in spatial attention and a revised output gating. This allows to focus on features from relevant spatial parts while attention is being tracked smoothly across the video sequence. Tang et al. (2017) added an additional stream to take as input depth maps enabling the model to encode 3D information present in the scene. Kazakos et al. (2019) proposed an end-to-end trainable mid-level fusion Temporal Binding Network (TBN) on top of a convolutional network to asynchronously fuse audio, RGB and optical flow across multiple temporal windows.

The success of the transformer architecture has also given rise to a new line of works that employ transformers as a backbone for processing videos, with the most popular ones being those by Patrick et al. (2021) and Arnab et al. (2021). These works extend the vision transformer to operate on multiple frames within videos. However, they were not developed specifically for egocentric videos, and report results on both third-person and egocentric videos using the same architecture.

Still, training a deep model is data and energy intensive and several works have been focusing on reducing the related costs. Possas et al. (2018) defined a reinforcement learning based technique for understanding actions using less energy. Sigurdsson et al. (2018) proposed to jointly learn from first- and third-person videos using weak supervision. Similarly, Li et al. (2021b) introduced an approach for pretraining egocentric video models using large-scale third-person video datasets. Min and Corso (2021) presented a probabilistic approach to estimate the gaze and utilise it for action recognition, avoiding the need for expensive gaze recording equipment. Plizzari et al. (2022) showed that the visual information collected by event cameras is suited for egocentric action recognition thanks to the lack of motion blur, high temporal resolution, and reduced power consumption.

Aiming to reduce the burden and uncertainty involved in the annotations of temporal bounds, a different line of works considered the problem of recognising actions using a single timestamp originating from narrations as supervision rather than temporal bounds (Moltisanti et al., 2019).

Another approach to egocentric action recognition is to consider it as a procedural problem and learn the key steps required to perform a task upon observing multiple egocentric videos as done in Bansal et al. (2022). This work is restricted to procedural tasks but is a venue for exploration as opposed to recognising isolated actions.

State-of-the-art papers Kazakos et al. (2021) developed an approach specific to egocentric videos using an audio-visual transformer with the visual features from Patrick et al. (2021). Importantly, in this work, the action is not seen in isolation: the untrimmed video and context are explored along with a language model providing action sequencing to enhance the predictions. This approach reported significant performance improvement over prior works, with action recognition reaching 49.6% on the validation set of EPIC-KITCHENS-100.

Following the trend of Transformers, Wu et al. (2022a) proposed a memory-based approach for efficient long-term video understanding. It uses the "keys" and "values" of a transformer as memory. The queries attend to an extended set of keys and values, which come from both the current time and the past. Each layer attends further down into the past, resulting in a significantly longer receptive field. They achieve 48.4% of action recognition accuracy on the EPIC-KITCHENS-100 dataset with much less model parameters ($0.5\times$ the parameters of Patrick et al. (2021)).

Recent works focused on designing new multi-modal integration strategies, in order to build models that work well across modalities, instead of being over-optimised for each modality. Girdhar et al. (2022) proposed a transformer-based model which, leveraging the flexibility of transformers, is trained jointly on classification tasks from different modalities—2D images, 3D images and videos. They achieve an impressive top-1 performance of 47.4% on EPIC-KITCHENS-100 validation set using their largest Swin-B

transformer model. Yan et al. (2022) adapted a multi-view transformer to multi-modal inputs: they created multiple representations or "views" by tokenising spectrogram, optical flow, and RGB using tubelets of different sizes. These tokens are fed into separate encoders and further fused through a fusion module, and aggregated by a global encoder. Using both modalities, the approach achieves action recognition of 47.2% on EPIC-KITCHENS-100 validation set. They outperform the approach from Yan et al. (2022) by 1% but are still below other approaches previously published such as Girdhar et al. (2022); Kazakos et al. (2021). All the prior work, except Kazakos et al. (2021) have been developed for general action recognition and the architectures are not optimised for egocentric vision specifically.

Gong et al. (2023) studied the problem of generalisation when data from certain modalities is limited or even completely missing during inference. They proposed a method for multi-modal generalisation based on a fusion module with modality dropout training, a cross-modal contrastive alignment loss, and a cross-modal prototypical loss for better few-shot performance. To further improve the efficiency, they jointly trained a memory compression module for reducing the memory footprint. Radevski et al. (2023) proposed to distil knowledge from a high-performing but impractical multimodal ensemble into a light-weight RGB-based model. Tan et al. (2023) achieved efficient recognition by combining RGB with the head motion information from IMUs.

Wang et al. (2023c) addressed the task of unpaired multi-view video learning. To this purpose, they introduced a method that aligns multi-view pseudo-pairs with high similarities in a semantics-aware manner. They allow first-person videos to gain insights from samples of varying views or modalities. Similarly, Xue and Grauman (2023) learn fine-grained frame-wise video features that are invariant to both the ego and exo views from unpaired data. They achieved that through a self-supervised contrastive-based temporal alignment objective.

Another recent trend is to leverage over Large Language Models (LLMs), to obtain stronger representations. Zhao et al. (2023c) used LLMs to automatically generate text pairing for videos, by densely annotating rich textual descriptions. When using those to learn video-text embeddings contrastively, and then evaluating on action recognition as a downstream task, results outperformed previous state-of-the-art on EPIC-KITCHENS-100, with 51.0% accuracy on action recognition. This sets the current state-of-the-art performance on the validation set of this dataset. Language has also been used in Plizzari et al. (2023) as a robust modality for improving domain generalisation to multiple domains. Starting from the rich diversity of Ego4D in terms of both scenarios and geographical locations, they proposed to represent each video as a cross-instance reconstruction of videos from other domains. Reconstructions are paired with text narrations to guide the learning of a domain generalisable representation.

Shah et al. (2023) considers learning keysetps for procedural problem using multiple modalities. To enable AR applications, Shah et al. (2023) utilise optical flow, depth, or gaze and propose the BMC2 loss to force modalities from multiple datasets to be close in the representation space. The work improves the F1 Score on the EgoProceL dataset by 14% and achieves state-of-the-art results.

Before concluding this section, we note the relevant tasks of action segmentation and action detection. The latter is distinct from action recognition as it aims to detect the start and end of action instances in long untrimmed videos as well as predicting the action categories. Previous works on action segmentation considered graph-based temporal reasoning (Huang et al., 2020b) and segmentation from single timestamps (Li et al., 2021c), while, up to our knowledge, Wang et al. (2023a) offer the first approach to action detection specifically for the egocentric domain without exocentric pretraining. This topic requires further exploration in untrimmed egocentric videos. We refer those interested in exploring action detection to Vahdani and Tian (2023) for up-to-date methods.

Datasets The most popular datasets for action recognition, EPIC-KITCHENS-100, Ego4D and EGTEA are detailed in Sect. 5. Additionally, several specialised datasets have been proposed to explore different aspects of action recognition in egocentric videos. Kitani et al. (2011) proposed a dataset made of videos both recorded in-house and sourced from YouTube. The first video, recorded on a QUAD, consists of 124 video splices (a video splice contains 60 frames) and contains 11 ego-actions. The second video, recorded in a park, is a 25 min workout video which contains 766 video splices and contains 29 different ego-action categories. Six egocentric YouTube sports videos have also been annotated to understand actions in outdoor sports videos.

DataEgo (Possas et al., 2018) and Multimodal Egocentric Activity (Song et al., 2016) datasets have been used for evaluating methods that focus on activity recognition with limited resources or on a budget. In DataEgo (Possas et al., 2018), Images from the camera have been synchronised with readings from the accelerometer and gyroscope. In total, it contains approximately 4 h of continuous activity while its multi-modal subset has only 50 min of separate activities. The Multimodal Egocentric Activity dataset (Song et al., 2016) contains 20 distinct life-logging activities, which are recorded both indoor and outdoor with significant changes in the illumination conditions. The dataset has 200 sequences in total and each activity category has 10 sequences of 15 seconds each. It also includes other synchronised sensor data: accelerometer, gravity, gyroscope, linear acceleration, magnetic field and rotation vector. Charades-Ego (Sigurdsson et al., 2018) aims to bridge the gap between egocentric and

third-person videos, providing a dataset with paired first-person and third-person videos involving 112 individuals and 4000 paired videos. Bock et al. (2023) recently introduced WEAR, an outdoor sports dataset for both vision- and inertial-based human activity recognition. The dataset comprises data from 18 participants performing a total of 18 different workout activities with untrimmed inertial (acceleration) and camera (egocentric video) data recorded at 10 outdoor locations totalling 15 h.

4.3.2 Recognising Objects

Object recognition in egocentric vision is pivotal for applications in augmented reality and robotics but remains a challenging task. Videos recorded from the first-person point of view capture spontaneous, unscripted scenes, in densely packed environments, where objects of various scales are closely packed and often occluded.

Seminal works Ren and Gu (2010) introduced a figure-ground segmentation system for egocentric object manipulation videos captured from a wearable camera, in order to separate the moving hands and the objects in-hand from the background. Kang et al. (2011) tackled the problem of object instance discovery, defining a method for finding new objects that a person can encounter in their daily living. Fathi et al. (2011) addressed the problem of learning object models from egocentric videos of household activities, using weak supervision. For each activity sequence, the method is merely supervised by the names of the objects which are present within it. They propose a segmentation method to partition each frame into hand, object, and background categories. Bolaños and Radeva (2015) attempted object discovery— that is detecting new object instances or concepts, and assigning them a label without prior training. Damen et al. (2016) proposed a fully unsupervised approach to discover objects and their usage from multiple users in a common environment. It consists of discovering task relevant objects, building an appearance model for each, distinguishing different ways in which each discovered object has been used and discovering the spatio-temporal dependencies between object interactions.

Combined with tracking, Bertasius et al. (2017) formulated the object detection task as an interaction between the segmentation and recognition agents. Initially, the segmentation agent generates a candidate object mask for each image, and relays this mask to the recognition agent, which then tries to learn a classifier using visual semantics and spatial cues. Other works addressed the problem of object tracking. Alletto et al. (2015b) developed an approach based on visual odometry and 3D localisation for tracking objects moving around a person.

With the release of the Ego4D dataset (Grauman et al., 2022), new benchmarks involving object understanding have

been proposed. The visual Queries Localisation (VQL) task aims to retrieve given query objects from an egocentric video. The hands and objects benchmark captures how the camera-wearer changes the state of an object by using or manipulating it—particularly capturing object state change. Yu et al. (2023) proposed a new benchmark for segmenting state-changing objects in each frame of the video, given the first frame mask as reference. Zhao et al. (2023b) proposed a new benchmark for studying instance tracking in 3D scenes from egocentric videos. Finally, Herzig et al. (2022) used object knowledge to achieve action recognition, as objects can be essential for recognising actions. They presented an object-centric approach that extends video transformer layers with a block that directly incorporates object representations.

In Darkhalil et al. (2022), semi-supervised video object segmentation is evaluated on egocentric videos, focusing on active objects using a newly annotated dataset of object segmentations.

State-of-the-art papers Akiva et al. (2023) proposed a self-supervised object detection model from egocentric videos. It uses two patch-wise objectives: an objective function operates in the temporal space, enforcing similarity of multi-temporal patches, and a function in the scale space, enforcing similarity of multi-scale patches. The former captures appearance variations in time such as viewing angles and illumination conditions, and the latter captures appearance variations in scale. Wu et al. (2023) examined the problem of continual object detection in egocentric streaming videos, by a plug-and-play module inspired from the complementary learning systems theory.

On tracking, Huang et al. (2023b) proposed DETracker, a method that jointly detects and tracks deformable objects in egocentric videos. DETracker consists of three key components: the motion disentanglement network (MDN), the patch association network (PAN), and the patch memory network (PMN). MDN plays a crucial role in efficiently estimating the motion flow between successive frames, where it distinguishes between global camera motion and local object motion. This separation ensures the algorithm's robustness in the presence of significant ego motion. PAN is responsible for tracking deformable objects by breaking them down into patches and locating corresponding patches in future frames for each individual patch. PAN maintains and continually updates feature embeddings of tracked objects over an extended time window.

In the VQL setting (Grauman et al., 2022), due to random viewpoints and the large number of possible object classes that are exhibited in egocentric recordings, the target object is hard to discover and confused with high-confidence false positives. To tackle this issue, Xu et al. (2023a) proposed the CocoFormer, a detection model that incorporates a conditional projection layer. This layer is responsible for generating a transformation matrix based

on the query. Subsequently, this transformation is applied to the proposal features, resulting in query-conditioned proposal embeddings. These query-aware proposal embeddings are then inputted into a set-transformer, enabling the model to effectively leverage the global context of the associated frame. Jiang et al. (2023) proposed VQLoC, a single-stage framework. The method jointly models the query-to-frame relationship and frame-to-frame relationships across nearby video frames, and uses that information for end-to-end training. Xue et al. (2023) achieved state-of-the-art performance on the Object State Change Classification benchmark, by means of EgoTask Translation (EgoT2), a framework that takes a collection of models optimised on separate tasks and learns to translate their outputs for improved performance on all tasks jointly. Recently, several works have been proposed that make use of 3D information. Tschernezki et al. (2021) proposed a three-stream neural rendering architecture, where the streams model respectively the static background, the dynamic foreground objects, and the actor. Mai et al. (2023) formalised a pipeline that better entangles 3D multi-view geometry with 2D object retrieval from egocentric videos.

Recent state-of-the-art action recognition benchmarks make use of object-related information for better classifying actions. Zhou et al. (2023) proposed an object-guided token sampling strategy that allows to retain a small fraction of the input tokens with minimal impact on accuracy. Moreover, they introduced an object-aware attention module that enriches our feature representation with object information and improves overall accuracy. Zhang et al. (2023a) tasked the model to predict object bounding boxes and names of objects during training in order to learn grounded and fine-grained correspondence between vision and language modalities.

Datasets Several egocentric datasets focused on objects have been built. TEgO (Lee & Kacorri, 2019) contains egocentric images of 19 distinct objects for training object recognisers. HOI4D (Liu et al., 2022b) captures videos of human-object interaction with 800 object instances from 16 categories. TREK-150 (Dunnhofer et al., 2023) annotated 150 videos from EPIC-KITCHENS (Damen et al., 2022) for tracking objects from 34 categories. VISOR (Darkhalil et al., 2022) annotated 272K manual semantic masks of 257 object classes, 9.9M interpolated dense masks, 67K hand-object relations, covering 36 h from EPIC-KITCHENS (Damen et al., 2022). EgoObjects (Zhu et al., 2023a) is a large-scale egocentric dataset for fine-grained object understanding. It contains over 9,200 videos of over 30 h collected by 250 participants, 654K object annotations from 368 object categories and 14K unique object instances. EgoTracks (Tang et al., 2023a) is a new dataset for long-term egocentric visual object tracking, with more than 22,028 tracks from 5708 average 6-min videos from Ego4D (Grauman et al., 2022). PACO (Ramanathan et al., 2023b) goes beyond traditional

object masks and provide richer annotations such as part masks and attributes. It captures both egocentric and non-egocentric views. The PACO-Ego4D subset of egocentric images has 140K part masks annotated in 26.3K images across 75 object classes and 456 object-specific part classes. Kurita et al. (2023) introduced RefEgo, which is also annotations of Ego4D videos, with more than 12k video clips and 41 h for video-based object referring expression annotations. *For the future* Despite the growing interest in action recognition for egocentric videos, there are several areas that warrant attention from the computer vision community. Firstly, there are only limited approaches developed specifically for egocentric vision. Most architectures are re-purposed from third-person videos and not optimised specifically for the ego viewpoint or the camera motion. For instance, the main evident consequence of relying on third-person video pre-training is that the ability to recognise fine-grained actions in egocentric videos is still significantly lower than the corresponding performance in third-person. Secondly, even by exploiting transformer architectures and multiple modalities, state-of-the-art methods currently achieve only 51.0% activity classification accuracy (obtained by Zhao et al. (2023c) on EPIC-KITCHENS-100). It is not clear whether approaches are lacking due to the size of datasets, ambiguity (or fine-grained nature) of labels, or the need for new architectures. With multiple possible explanations and avenues for exploration, the field is only progressing slowly. Sequences of papers tend to improve performance by small margins (0.5-1%).

Thirdly, egocentric vision introduces further challenges, such as the need for modelling long temporal dependencies and learn from long-tail and class imbalanced data. Perrett et al. (2023) recently introduced a new benchmark for long-tail recognition in video, including egocentric video.

Fourthly, despite the early success of integrating gaze for egocentric action recognition, subsequent datasets do not capture the rich, though expensive, egocentric gaze. A few sequences in Ego4D (Grauman et al., 2022) include gaze but these are not labelled specifically with fine-grained actions. Gaze offers the ability to focus the attention on areas of the image and prime for the next actions. However, with the absence of large scale egocentric action recognition datasets that include gaze, this avenue of research is currently under explored.

Fifthly, the recent introduction of the EPIC-SOUNDS dataset (Huh et al., 2023) showcased the need for modality-specific annotations of action classes and temporal extents. The work showcased the disadvantage of training one modality with labels from a different modality. This perspective on multi-modalities in egocentric video can unlock new approaches for developing multi-modal architectures.

Lastly, the heavy reliance on labeled datasets for training limits the capabilities of models. Not only is labelled data

expensive to acquire, but the choice of the closed vocabulary of action classes and the granularity of actions remains subjective. Progressing from a closed subset to open labels remains an open question in most machine learning tasks including egocentric action recognition. With the advent of LLMs, this surely is the future of recognising actions, despite the lack of metrics to assess success and monitor progress.

In parallel, understanding object affordances-how objects can be used or interacted with-goes beyond basic recognition and it is crucial for applications like assistive robotics, augmented reality, and personal assistance. Such understanding will allow systems to interact more intuitively and effectively with their environment, a key advancement for EgoAI.

4.4 Anticipation

Anticipation tasks aim to predict the future state of the scene from the observation of the present. These tasks are particularly relevant in egocentric vision as they capture an uninterrupted picture of the camera wearer's interaction with the environment and objects, hence providing the opportunity to model their behaviour and understand their goals and intentions. Following Rodin et al. (2021), we divide the works focusing on anticipation tasks into three categories, depending on the target of future prediction, namely, *actions* (Sect. 4.4.1), *objects* (Sect. 4.4.2), and *trajectories* (Sect. 4.4.3). We would like to note that, while we refer to these works with "anticipation", the term "forecasting" has also been used in the literature to refer to these tasks.

4.4.1 Anticipating Actions

Action anticipation is the task of semantically predicting the next action to take place in a video. Systems able to tackle this task can provide proactive assistance to the users and improve their safety by understanding the camera wearer's goals and future interactions. Current approaches formalise action anticipation as a video classification task which aims to predict a future action from the observation of a video segment of the past. Due to the stochastic nature of the task, methods are required to produce a ranked list of outputs and they are considered successful when the ground truth future action is in the top-k predictions.

Seminal works Action anticipation has been studied both in egocentric and third-person vision with similar approaches. The problem has been introduced by Pei et al. (2011), who considered a goal-oriented scenario and used and-or-graphs to represent the different actions which might be performed by a human actor at any given point in an observed video. The feasibility of the task is demonstrated by comparing the proposed solution with human performance. Lan et al. (2014) subsequently standardised the task, evaluating it at different prediction horizons on videos from TV shows.

Koppula and Saxena (2015) showed the benefits of predicting future human poses, future human and object trajectories and future interacted objects for robotics applications. On the side of methodological advancements, Vondrick et al. (2016) first explored the idea of predicting future actions by training deep neural networks to anticipate future representations on unlabeled videos. A similar concept has been further explored by Gao et al. (2017) and Gers et al. (2000), who also leveraged Long-Short-Term Memory (LSTM) networks and a reinforcement learning criterion to anticipate future actions at different prediction horizons.

While the aforementioned works mainly considered the problem of anticipating the next action appearing in the video (short-term action anticipation), Abu Farha et al. (2018) proposed the task of predicting a longer sequence of future actions in the case of goal-driven, structured procedures (long-term action anticipation). Recently, the problem of action anticipation has been studied also in the context of egocentric videos. In particular, Damen et al. (2018) first proposed an action anticipation challenge on egocentric videos, and Furnari and Farinella (2019) systematically tackled the anticipation problem exploring the importance of egocentric cues such as object-based features.

State-of-the-art papers Girdhar and Grauman (2021) proposed Anticipative Video Transformer (AVT), an end-to-end attention-based video modelling architecture that attends to the previously observed video in order to anticipate future actions. Gu et al. (2021) leveraged transformer-based attention to aggregate features across temporal dimension, modalities, and symbiotic branches (*verb/noun* branches) respectively. Zhong et al. (2023) extended the transformer architecture to operate on multiple modalities, by unifying multi-modal data through mid-level fusion and using the obtained representations for anticipating next actions. Roy and Fernando (2022) proposed an approach that uses learned latent goals to anticipate the next action. Latent goals are accompanied by *goal closeness* and *goal consistency* losses, aiming to produce a visual representation that is closer to the latent goal and consistent throughout consecutive actions. At the moment of writing, Roy et al. (2024) achieve state-of-the-art performance on the EPIC-KITCHENS Action Anticipation challenge, by refining video representations using a transformer model computing the change in the appearance of objects and human hands due to the execution of the actions. Recently, Zhao et al. (2023a) achieved state-of-the-art long term action anticipation performance on the Ego4D dataset with a hybrid architecture integrating vision-based action recognition to infer high level, symbolic video representations and large language models for procedure planning.

4.4.2 Anticipating Objects

To provide assistance to the user at a more granular level it is useful to make future predictions of attended or manipulated physical regions appearing in the egocentric video, such as objects, scene parts, or object parts. In this way it is possible to issue alerts when specific parts of a potentially dangerous object are going to be touched or when the camera wearer is about to interact with the wrong object in a known workflow. Current approaches formulate future region prediction as object detection, heat-map prediction, or semantic mask prediction tasks. Algorithms are usually evaluated using spatial overlapping retrieval metrics such as mean Average Precision.

Seminal works Furnari et al. (2017) first introduced the problem of predicting which objects will be interacted with next. Zhang et al. (2017) proposed to predict future gaze, i.e., the spatial location which will be attended to by the user in the future. Nagarajan et al. (2019) investigated the anticipation of object affordances by predicting interaction hotspots in videos. Liu et al. (2020b) showed how predicting interaction hotspots and future hands trajectories can support more abstract tasks such as action anticipation. Notably, these works have addressed their own versions of object anticipation tasks. Grauman et al. (2022) worked toward standardisation of this task, termed short-term object-interaction anticipation. This predicts which of the objects in the scene will be interacted with by the camera wearer (noun of the future object), how the interaction will take place (verb denoting the interaction), and when the interaction will begin (time-to-contact in seconds).

State-of-the-art papers Among the anticipation algorithms based on regions, most state-of-the-art approaches mainly focus on the prediction of future interacted objects and in particular on the short-term object interaction anticipation task as defined in Grauman et al. (2022). Pasca et al. (2023) currently achieve state-of-the-art performance on the task with TransFusion, a multimodal transformer-based architecture that exploits language. In particular, TransFusion leverages pretrained image captioning and vision-language models to extract the action context from past video frames. This, together with the next video frame, is processed by the multi-modal fusion module to forecast the next object interaction. Recently, Lai et al. (2023a) proposed a state-of-the-art approach to future gaze prediction in social scenes based on the analysis of audio and video. It models audio-video correlations with a spatial fusion and a temporal fusion branch, guided by a multi-modal contrastive loss. Fused embeddings are decoded jointly to predict future gaze.

4.4.3 Anticipating Trajectories

Systems able to make future predictions in the form of trajectories will know in advance where the user may go, how the observed objects will move in the scene, and how the camera wearer's hands are going to move in the near future. Such information is crucial for all those applications that need to plan in advance, e.g., to suggest alternate routes (to avoid passing through dangerous zones) or detect unsafe operations involving the interaction between hands and objects. The metric most commonly used in trajectory prediction is the final displacement error (FDE), defined as the L2 distance between the predicted location and the ground truth.

Seminal works Park et al. (2016) first proposed the task of predicting the possible trajectories that the camera wearer may follow from egocentric video. In a complementary way, Yagi et al. (2018) studied the problem of predicting the future trajectory of other persons observed from the egocentric point of view. Liu et al. (2020b) investigated how predicting hands trajectories can be beneficial for action anticipation. Jia et al. (2022b) proposed the task of anticipating a time series of future hand masks from an egocentric video. Bao et al. (2023) explored the problem of predicting future hand trajectories in 3D with the aim to support the understanding of human intention and behaviour in AR/VR applications. While trajectory prediction tasks have not been systematically studied in the egocentric perspective, a first attempt to propose standard tasks has been done by Grauman et al. (2022), where two tasks related to the prediction of future locomotion and hands trajectories are formulated.

State-of-the-art papers Alikadic et al. (2022) presented a new method that leverages transformers to forecast future trajectories of pedestrians from egocentric views. The model predicts the trajectories by relying on previous locations and scales, dynamics poses, and ego-motions of the camera wearer. Kai et al. (2023) designed a multi-channel tensor to represent social interaction, including pedestrian pose, depth and their relative locations. They fed this input to a novel end-to-end fully convolutional transformer (Conv-Transformer) network. Hatano et al. (2023) recently proposed an approach that uses semantic information to connect bird's-eye coordinates to the egocentric viewpoint. This allows to utilise existing third-person view methods on the egocentric view, without the need to re-train.

Datasets Action anticipation and future region prediction works have often relied on action recognition datasets, namely ADL (Pirsiavash & Ramanan, 2012), EPIC-KITCHENS (Damen et al., 2022), EGTEA Gaze+ (Li et al., 2021a), and Ego4D (Grauman et al., 2022). These general-purpose datasets are described in Sect. 5. Since most action recognition datasets, such as ADL, EPIC-KITCHENS and EGTEA, do not contain significant human locomotion, early trajectory prediction works have performed evaluations on specific

datasets collected on purpose. In particular, Park et al. (2016) collected the EgoMotion dataset, a set of egocentric videos acquired in various indoor and outdoor scenes using first-person GoPro Hero 3 stereo cameras. The dataset comprises 26 scenes, 65.5k frames and 9.1 h of video covering various activities such as walking, shopping, and social interactions. The First-Person Locomotion Dataset proposed by Yagi et al. (2018) comprises about 4.5 h of egocentric videos recorded by people wearing a chest-mounted camera and walking around in diverse environments. The Ego4D (Grauman et al., 2022) is the first to offer videos suitable for anticipating actions, objects and trajectories. Limited annotations on this massive-scale dataset enable works on anticipating actions, forecasting hand and full-body trajectories.

For the future Despite the progress of research in this field, at the moment of writing, future anticipation approaches achieve limited performance. For example, the current state-of-the-art approach for action anticipation by Roy et al. (2024) only achieves a mean top-5 per class recall of 18.1% on the test set of the EPIC-KITCHENS-100 dataset. Similarly, the best performing approach to short-term object interaction anticipation (Pasca et al., 2023) achieves a top-5 mAP of 24.7% in next-active object prediction and 3.4% when also predicting the interaction verb and time-to-contact on the test set of Ego4D. These results highlight the very complex nature of anticipation tasks and the need for advances in this area. Current anticipation approaches also suffer from major limitations which prevent their widespread adoption. Most approaches assume that a "trimmed" video is sampled at a fixed time before the beginning of the action and fed to the model, which constitutes an unrealistic scenario, given that the occurrence of future actions is unknown at test time. Despite some recent work towards an untrimmed anticipation scenario (Rodin et al., 2022), the trimmed setting remains the most common one.

4.5 Gaze Understanding and Prediction

Understanding and predicting which areas of the scene the camera wearer is attending is critical for AR, assistive technologies, and human behaviour analysis. This task involves developing sophisticated algorithms that can accurately estimate the direction of a person's gaze based on the visual information captured by an eye-mounted egocentric camera. Gaze understanding enables more accurate modelling of the camera wearer's visual attention, which offers useful insights for downstream tasks including predicting which object(s) the camera wearer is focused on at a time. Methods are evaluated by their ability to produce attention maps coherent with ground truth gaze measurements. We will focus on methods estimating gaze from the perspective of the beholder. For approaches that utilise remote eye trackers to process frames

containing the wearer's face, we refer to the work by Cazzato et al. (2020).

Seminal works Predicting gaze is inherently challenging, leading initial works to explore different cues to understand the wearer's focus of attention. For instance, Yamada et al. (2011) highlighted the challenges associated with using visual saliency maps based on colour, intensity, and orientation in egocentric vision, especially when there is significant egomotion. To address this issue, Yamada et al. (2012) combined visual saliency maps with rotation- and translation-based attention maps obtained through egomotion estimation. By doing so, they aimed for more robust and accurate egocentric visual attention predictions.

The first approach for predicting gaze from egocentric videos was presented by Fathi et al. (2012b) who simultaneously tackled daily activities recognition and gaze location prediction using a common probabilistic generative model. Li et al. (2013) leveraged implicit cues in egocentric videos, such as hand location, pose, and motion, to predict gaze. Additionally, they modelled gaze behaviour to enhance prediction performance. Building on this work, Huang et al. (2018) modelled patterns in the temporal shift of gaze fixation. Their work is based on the assumption that during fixation, the gaze tends to be located on the same object, and patterns of gaze shift depend on the high-level task, which can be learned.

A more recent study by Al-Naser et al. (2019) predicted gaze based on the objects framed in the video. They used features and bounding boxes from an object detection model and combined both classic gaze point regression formulation and classification for prediction. In parallel, Tavakoli et al. (2019) analysed both top-down and bottom-up factors influencing egocentric gaze prediction. Their work confirmed the relevance of the manipulation point over hand regions and the importance of hand-object interaction for gaze prediction.

An innovative approach was introduced by Thakur et al. (2021), where the information from the video stream was combined with head movement obtained from IMU (Inertial Measurement Unit) data to improve gaze estimation. On a slightly different note, Su and Grauman (2016) focused on understanding *when* engagement with the environment happens instead of *what* the wearer is looking at. This perspective aimed to detect moments of interaction from egocentric videos.

State-of-the-art papers The current state-of-the-art method for gaze estimation on both the EGTEA Gaze+ dataset (Li et al., 2021a) and Ego4D dataset (Grauman et al., 2022) is the approach proposed by Lai et al. (2022). The authors tackled the challenge of integrating different gaze cues, such as the likelihood of scene objects to be targets, their location, and the head motion pattern related to gaze shifts, into a comprehensive analysis of visual attention. To achieve this, they developed a transformer-based model that captures the con-

nection between the global scene context and local visual cues using a Global-Local Correlation module. By combining these various gaze cues and context information, their method achieves state-of-the-art performance in predicting gaze in egocentric videos on both datasets.

Recent devices such as the Meta Aria glasses have an onboard estimation of gaze through sensors embedded with eye-tracking. Their modern eye trackers use corneal reflection, a method involving near-infrared light to illuminate the eyes, causing a reflection that is detected by a high resolution camera.

Datasets Most of the datasets used in this section are general datasets described in Sect. 5 with gaze tracking data to serve as ground truth—GTEA Gaze dataset by Fathi et al. (2012b), EGTEA Gaze+ by Li et al. (2021a), GTEA-sub by Huang et al. (2018) and Ego4D by Grauman et al. (2022). In addition, other datasets have been used in research for specific purposes, but are not publicly available. Among the limited public datasets, the Object Search Tasks (OST) Dataset by Zhang et al. (2017) includes 57 sequences of search and retrieval tasks performed by 55 subjects provided with eye-tracking data.

For the future The future of gaze prediction in wearable devices holds immense potential to revolutionise human-computer interaction and user experience. As wearable technology becomes more advanced and pervasive, integrating accurate and real-time gaze tracking capabilities will enable seamless interactions with digital content. Wearable devices with built-in gaze prediction algorithms could offer intuitive and hands-free control, improving accessibility and usability across applications. With precise gaze tracking, wearable devices can adapt their interfaces dynamically, presenting relevant information based on the user's visual focus. Despite the advances in this area, gaze analysis still poses various challenges, including the need for large annotated datasets, subjective bias due to individual differences, handling eye blinks and data attributes like occlusion and illumination (Ghosh et al., 2023; Pathirana et al., 2022). Nonetheless, even with the application of state-of-the-art techniques (Lai et al., 2022), results are not ideal, as evidenced by the F1 scores of 44.8 and 43.1 on EGTEA Gaze+ (Li et al., 2021a) and Ego4D (Grauman et al., 2022), respectively.

4.6 Social Behaviour Understanding

The wearable devices of the future will be able to support users in a variety of scenarios related to their daily lives. As humans are by nature social animals, we expect wearable systems to be able to understand the social behaviours of the camera wearers and of others they engage with. The research community has investigated this area with different topics. We organise our literature overview by considering

the works related to *understanding the relationship with the speaker* (Sect. 4.6.1), *detecting/modelling social interactions* (Sect. 4.6.2), *estimating attention towards the camera wearer* (Sect. 4.6.3) and *joint attention* (Sect. 4.6.4). We discuss the relevant publications in each sub-area and provide insights into what future directions may benefit the community with respect to this topic. It should be noted that the approaches discussed in this section aim to analyse social behaviour from the point of view of the camera wearer, which involves specific challenges and opportunities as compared to approaches based on fixed cameras. Indeed, analysing social interactions from egocentric vision gives a privileged views into behaviours directed towards the camera wearer such as facial expressions, speaking acts, and eye contact. Estimating visual attention through two synchronised wearable cameras further enables the study of joint attention which can have useful applications, including in diagnosing and monitoring social-related health conditions.

4.6.1 Modelling the Relationships with the Speakers

This area focuses on modelling the relationship between the camera wearer and speaking subjects appearing in the egocentric field of view. Previous works have addressed different objectives that improve the camera wearer's audio-visual interactions, including improving speech quality in a noisy environment, determining auditory attention towards one of a set of speakers, determining which subjects are talking to the camera wearer, and detecting speakers and transcribing their speech. Existing approaches have considered an array of similar, yet distinct, tasks related to this area, generally proposing approaches based on the processing of both audio and video. Segmentation maps or bounding boxes are usually produced to spatially detect the speaker. The evaluation is carried out by comparing the predicted areas with ground truth annotations.

Seminal works Kumano et al. (2015) first considered the use of egocentric vision to perform automatic conversation analysis. The work targeted a multi-party conversational scenario, where participants were equipped with in- and out-cameras with microphones and the gaze behaviour of each interlocutor was hence estimated via self-calibration. Donley et al. (2021) proposed the task of enhancing a target speech source and speech intelligibility in conversations held in noisy environments and recorded through egocentric devices.

State-of-the-art papers Lu and Brimijoin (2022) presented a study to assess whether head angle estimated via egocentric devices is predictive for sound source selection. Jiang et al. (2022) tackled the problem of active speaker detection by using both video and multi-channel microphone array audio. Grauman et al. (2022) presented the "Social Interactions" benchmark, which includes tasks aimed at identifying communicative acts directed towards the camera-wearer. The

most relevant for this section is the "Talking To Me" task, which focuses on classifying whether each visible face, based on a video and audio segment with tracked faces, is talking to the camera-wearer. Additionally, the researchers introduced the AV diarisation benchmark, to understand the camera-wearer's ongoing interactions with people starting from speech. Those are: localisation and tracking of the participants, active speaker detection, diarisation of each speaker's speech activity, and transcription of each speaker's speech content. For the latter, Gabeur et al. (2022) recently proposed a new model for audio-visual automatic speech recognition based on a multi-modal audio-visual transformer trained end-to-end from spectrograms and RGB frames. Ryan et al. (2023) proposed a novel task of egocentric auditory attention localisation, which identifies the person the camera wearer is talking to in a multi-people multi-conversation scenario. The task is carried out by considering audio-visual signals. The auditory signals are given by a directional array of microphones, while the visual signals are given by egocentric video.

4.6.2 Detecting and Modelling Social Interactions

Works in this area detect the presence of social relationships from egocentric images or video and potentially characterise such relationships, by highlighting the engaged subjects and classifying the behaviour (dialogue, monologue, discussion, etc.). Being able to detect and characterise social relationships can allow wearable systems to gain an understanding of the social context of the camera wearer, consider video segments relevant for later recollection (e.g., record important conversations), and track the camera wearer's social relationship for monitoring and diagnosis of potential disorders. The approaches discussed below did not follow a common task definition, instead they analysed related problems which were tackled with disparate techniques.

Seminal works Fathi et al. (2012a) proposed the first work to detect and categorise social interactions in egocentric video among a group of individuals. The location and orientation of each subject's face were used to compute a line of sight and obtain a location in space indicating the focus of attention. Head movements of egocentric cameras were also used for a better understanding of attentional focus. Narayan et al. (2014) evaluated the performances of dense trajectories to recognise social interactions performed by the camera wearer and other subjects acquired from the egocentric point of view. Alletto et al. (2015a) addressed the problem of partitioning people in an egocentric video sequence into socially related groups. Interactions are then detected with clustering and structural learning. Bambach et al. (2015) focused on hands to detect interactions, and investigated the tasks of hand detection, disambiguation, and segmentation from videos of interacting people using appearance models based on CNNs.

Yonetani et al. (2016) considered the problem of modelling dyadic interactions (interactions between two people) from paired videos, where micro-level actions and reactions such as slight shifts in attention, subtle nodding or small hand actions are detected. Yang et al. (2016) proposed the concept of "wearable social camera", a camera that summarises the video of the user's social activities. To achieve this goal, common features among different social interactions, called interaction features, are extracted and processed.

Su et al. (2016) presented a method to predict future movements of basketball players based on the analysis of social behaviours. 3D reconstruction of multiple first-person cameras and gaze information were used to automatically annotate each player's video the visual semantics. A Siamese neural network was later trained to retrieve future trajectories based on group movements. Aghaei et al. (2017) considered the problem of social style characterisation from egocentric photostreams. This is done by detecting temporal segments characterised by social interactions, detecting faces, extracting social signals and classifying the social interaction into formal or informal. Duarte et al. (2018) investigated the non-verbal visual cues to "read the intention" of other humans in social interactions from egocentric videos. Other works focused on robot-centric activity recognition (Ryoo & Matthies, 2013; Xia et al., 2015; Gori et al., 2016), where the goal is to enable an observer (e.g., a robot or a wearable camera) to understand what activity others are performing towards it. Xia et al. (2015) proposed to extract features from an ego-motion region and an independent motion region separately and combine the descriptors using multiple kernels. Gori et al. (2016) proposed a unified mid-level descriptor capable of discriminating between different types of activities. They called it Relation History Image (RHI), and it is built as the variation over time of relational information between every pair of local regions (joints or image patches) belonging to one or a pair of subjects.

More recently, Bertasius and Shi (2017) proposed to predict cooperation patterns in the near future without requiring manually labelled intention labels. To do that, they modified the output of a pretrained pose estimation network to represent the camera wearer's internal state, including visual attention and intentions. They then employed this transformed output as a supervisory signal to train another network for the cooperative basketball intention task.

State-of-the-art papers Due to limited datasets on the topic, there are only a handful of recent works proposing methods that could be deemed state-of-the-art on this task. Li et al. (2019a) addressed the problem of modelling dyadic interactions by explicitly modelling the relations between the interacting subject and the camera wearer using a dual recurrent network. That incorporates two interconnected subtasks, namely individual action representation learning and dual relation modelling. Lai et al. (2023b) recently proposed

the task of modelling persuasive behaviours during multi-player social deduction games leveraging language models. Given an utterance and its corresponding video segment, they seek to predict the persuasion strategies adopted in the utterance. They first leverage a pretrained language model as the text encoder to obtain the utterance embedding, and a vision transformer to obtain the visual embedding. They then concatenate the textual and visual features to predict the persuasion strategy.

4.6.3 Estimating Attention Towards the Camera Wearer

This line of work focuses on understanding when a subject appearing in the egocentric field of view is attending to (e.g., by looking at or talking to) the camera wearer. This ability can allow egocentric vision systems to facilitate social interactions (e.g., by notifying the camera wearer when a person is trying to make contact), improving diagnosis of potential disorders of the observed subjects by studying their attentional pattern towards the camera wearer (e.g., Autism Spectrum Disorder when the camera wearer is a doctor), and logging conversations with the camera wearer for later recollection. Current approaches have analysed different tasks, ranging from detecting eye contact to detecting people looking at or talking to the camera wearer.

Methods are evaluated by validating predictions against manually annotated ground truth or comparing algorithms' performance to human performance. Similar to other tasks, the evaluation has not yet been investigated in a systematic way.

Seminal works Ye et al. (2012) developed the first system capable of recognising eye contact between the wearer and a person in front of the camera. The system aimed to detect atypical patterns of gaze and eye contact in children to understand early signs of autism. The approach used face detection to find the child's face and then estimated its gaze in 3D space to determine whether it pointed towards the wearer. Another autism diagnosis system was developed by Petric et al. (2014), leveraging the idea that autistic children tend to interact more with technological devices than with humans. They utilised robots equipped with frontal cameras to understand joint attention patterns. Subsequently, Smith et al. (2013) focused simply on capturing whether people in images were looking towards the wearer or not, tackling what they called "gaze locking". By discretising the problem and avoiding continuous tracking of the observer, they simplified the task. To address the challenges of strong appearance diversity in human eyes, Ye et al. (2015) proposed a model that couples eye appearance with head pose for improving eye contact detection.

State-of-the-art papers The current state-of-the-art approach in eye contact detection, as proposed by Chong et al. (2020), scales the training dataset to 4.7 million human-annotated

eye contact images and leverages large-scale datasets from other tasks like face recognition to build initial representations that understand the relationship between head pose and gaze direction.

Recently, Grauman et al. (2022) introduced the "Looking At Me" task, which involves classifying whether each visible face in a video, with localised and identified social partners, is looking at the camera-wearer. Xue et al. (2023) takes a unified approach by building on the idea that various video understanding tasks are related. They propose EgoT2, a framework that combines different task-specific models to improve performance. They also integrate tasks like "Talking To Me" and "Active Speaker Detection" to understand if a specific person is talking to the wearer and who is speaking, respectively.

4.6.4 Estimating Joint Attention

Works in this area focus on modelling the joint attention of multiple subjects towards scene regions, objects, or people. The ability to estimate joint attention can allow egocentric systems to monitor and improve social interactions (e.g., by monitoring the joint focus of attention and notifying the camera wearer when it changes), improve diagnosis of potential disorders, and enhancing video curation and summarisation by detecting the most popular scenes from a set of synchronised egocentric video streams. Modelling is usually performed by predicting social saliency maps, detecting jointly attended objects, or determining a subset of subjects with coherent attention patterns. Approaches are evaluated by comparing predictions against manually annotated ground truth labels or assessing how the estimated joint attention is predictive of other subjects' attention. Previous works have considered different but related task formulations.

Seminal works Perceiving joint attention of wearers towards a common scene was introduced by Park et al. (2012). They constructed a 3D social saliency field, and located gaze concurrences by localising wearable cameras via structure-from-motion in a common coordinate system and triangulating the attention of each wearer. Subsequently, Park et al. (2013) introduced the concept of "social charges" as latent quantities driving the attention of people in a social group, defining the relationships between these charges and the primary gaze of each member. They estimated time-variant social saliency fields from observed primary gaze, enabling the prediction of gaze direction at any proximal location or time. Building on previous works, Park and Shi (2015) proposed a method to estimate the likelihood of joint attention as a function of a social formation, without relying on the gaze of group members. Using the dataset from Park et al. (2012), where the locations of group members and their joint attention were measured, their learned representation demonstrated the ability to predict the social saliency of real-world scenes.

Another line of work explores the multiple wearable camera setting to automatically edit footage in a smart manner. For example, Arev et al. (2014) used the centre of attention of different cameras as an indicator of what is important in the videos and combined it with cinematographic guidelines to produce effective summaries of the original footage. Hoshen et al. (2014) estimated the cameras looking at the same region, without reconstructing the scene in 3D.

A slightly different approach is presented by Lin et al. (2015), where the goal is to locate the person who draws attention from most wearers in a multi-camera setting. They used motion patterns to correlate people across videos to avoid appearance-related challenges, such as groups of people dressed similarly.

Kera et al. (2016) authored one of the pioneering works on discovering joint attention based on visual appearance. They proposed to locate objects that are attended by multiple camera wearers. They used multiscale spatiotemporal tubes around points of gaze as potential objects of interest and performed unsupervised clustering on them.

State-of-the-art papers The estimation of joint attention is an emerging topic, and it has mostly seen pioneering works proposing various variations on the task. More recently, Huang et al. (2020a) improved on the approach of Kera et al. (2016) on locating objects that are attended by multiple camera wearers. They tackled the challenges of cluttered scenes and noisy gaze by first temporally locating joint attention periods and then spatially segmenting the object of joint interest. Those contribute to a more reliable spatial segmentation than simply using regions in proximity to the points of gaze, which might be noisy. They achieve that by means of a hierarchical graphical model composed of multiple linear chain conditional random fields.

Datasets Numerous public datasets are now available to evaluate model performance in social behaviour understanding tasks. Among the most widely used is the First Person Social Interaction Dataset (FPSI) (Fathi et al., 2012a) dataset described in Sect. 5. The JPL First-Person Interaction dataset (Ryoo & Matthies, 2013) stands out as the first one to annotate actions performed by a robot. It includes 7 actions, comprising 4 friendly interactions, 1 neutral interaction, and 2 hostile interactions, aiming to study robot-centric activity recognition. Similarly, the datasets from Xia et al. (2015) also address robot-centric activity recognition, leveraging the depth modality in addition to other features. For dyadic interactions, the Paired Egocentric Video dataset by Yonetani et al. (2016) contains over 1000 pairs of egocentric videos capturing micro-action and reaction patterns from the perspectives of both interacting individuals. The EGO-GROUP and EGO-HPE datasets (Alletto et al., 2015a) offer videos featuring groups of people, enabling testing of group detection in egocentric vision and head pose estimation of the participants. The Focused Interaction dataset (Bano et al., 2018) provides

multimodal representations of individuals interacting, supporting the development of automatic interaction detection. Park et al. (2012) proposed a dataset for evaluating joint attention. Three video sequences were recorded: a meeting with two groups, a musical with alternating performances, and a party with multiple activities. The UTJA-M dataset (Huang et al., 2020a) captures moments of joint attention among individuals and releases the tracked gaze for all participants. Targeting the development of conversational AI, the EgoCom (Northcutt et al., 2020) and EasyCom (Donley et al., 2021) datasets offer multimodal recordings, with a primary focus on audio. While EgoCom contains nearly 40 h of recordings, EasyCom provides synchronised audio from different participants, incorporating realistic acoustic noise into the setting. Lai et al. (2023b) proposed a multi-modal dataset for studying persuasive behaviours during social games. Videos are sourced from both YouTube and the Ego4D social dataset and include text, video, and audio signals. In total, it contains 5,815 utterances from Ego4D and 20,832 utterances from YouTube.

For the future As discussed in the previous sections, the literature on social behaviour understanding is less stratified as compared to the other tasks considered in this paper. The current state-of-the-art performance varies significantly depending on the task. For the "Talking To Me" task in the Ego4D benchmark (Grauman et al., 2022), the results show 53.9% mAP and 54.3% accuracy on the test set, which is still far from human-level performance. Conversely, in the task of eye contact detection has reached acceptable performance—the state-of-the-art achieves an overall precision of 0.936 and a recall of 0.943 across 18 validation subjects. This performance is comparable to that of 10 trained human coders. While several datasets have been collected, the focus has been on diagnostic and summarisation tasks. The potential for understanding social behaviour can be a game changer in strategic interactions—whether gaming or even offering advice in live negotiations and group meetings. Such potential requires interdisciplinary research, beyond computer vision expertise and is currently at its infancy.

4.7 Full-Body Pose Estimation

The reconstruction of the wearer's body pose is crucial to enable applications such as daily life monitoring and AR. Consequently, the research community has devoted increasing attention to this field in recent years. Human pose estimation aims to create representations of the human body either in the local egocentric camera space or in a world coordinate system. Two main approaches are employed for constructing body representations: *kinematic models*, which utilise joint positions and limb orientations without capturing detailed texture and shapes, and *volumetric models*, which provide more realistic representations and capture deforma-

tions. Methods are assessed by their Mean Per Joint Position Error (MPJPE) measuring the average distance between the predicted joints and the ground truth joints. Works addressing this problem in egocentric vision tend to differ from works focusing on fixed cameras due to the limited field of view of wearable cameras, which rarely captures the full view of the human body.

Seminal works Wearable cameras, with their limited field of view primarily focused on the wearer's attention, only capture a partial view of the wearer's body. As a result, significant efforts in egocentric 3D pose estimation have been dedicated to designing systems that can overcome this limitation. One pioneering study conducted by Shiratori et al. (2011) tackled this challenge by employing a Structure-from-Motion technique. They utilised 16 outward-looking body-mounted cameras to reconstruct both the relative and global joint motion of a person in outdoor environments. The objective of their work was to develop a motion capture system that could operate effectively "in the wild".

Drawing inspiration from the concept of overcoming the wearer's body invisibility, subsequent works in the field have explored the use of cameras positioned to face downwards towards the body. One notable approach in this regard is the development of EgoCap by Rhodin et al. (2016). It involves a specially designed head-mounted stereo rig setup with downward-facing cameras. Expanding on the downward-facing camera setup, Xu et al. (2019) proposed the first real-time motion capture system utilising a single monocular fisheye camera mounted on a cap. Finally, addressing the specific setting of head-mounted displays in AR/VR, Tome et al. (2019) positioned the camera on the rim of a VR headset and generated a photorealistic dataset which served as a valuable resource for research.

In the study conducted by Wang et al. (2021a), a significant focus was placed on addressing the limitations associated with adopting a local egocentric camera reference system, particularly in applications such as animating body locomotion in a virtual environment. Recognising this restriction, the researchers proposed a novel framework that combines the local pose estimation with the world coordinate system obtained through SLAM. The aim was to achieve a temporally stable integration of both perspectives, enabling more robust and accurate results.

In parallel, chest- and head-mounted outward-looking cameras have also been employed to infer the pose of the wearer in more challenging scenarios where most of the body is out of the camera's field of view. In particular, Rogez et al. (2015b) extended the estimation of human body part joints from hands to the entire upper limb using synthetic depth data training. Jiang and Grauman (2017) went further by attempting to estimate the full-body of the camera wearer from a single outward-looking camera, leveraging dynamic motion

signatures and static scene structure to infer the "invisible" human body pose.

Unlike previous approaches focused solely on smooth and accurate poses, Yuan and Kitani (2018) introduced a method that formulates body pose estimation as a Markov decision process adopting a dynamics-based perspective. They leveraged a physics simulator to train a policy that generates physically plausible poses. In a subsequent study, Yuan and Kitani (2019) improved upon this approach by adopting a control-based methodology that not only estimates poses but also forecasts valid future poses, going beyond pure estimation. In a similar vein, Luo et al. (2021) employ a combination of kinematics- and dynamics-based modelling to achieve the first-ever estimation of physically plausible 3D human-object interactions. The authors collected their own dataset, which include 6 degrees of freedom (DoF) object poses. These object poses are factorised within the scene and subsequently utilised by their method to estimate realistic 3D human-object interactions, accounting for the physical constraints and dynamics involved.

In embodied AI, understanding social interactions holds great significance, leading to a focus on pose estimation tasks during social interactions. Ng et al. (2020) introduced a method called "You2Me" that utilises the action-reaction social interaction dynamics between the wearer and a second person as prior to estimate the wearer's pose. This work emphasises the influence of inherent synchronisation during interactions. On the other hand, Liu et al. (2021b) made a significant contribution by being the first to attempt the estimation of a second person's pose from an egocentric perspective while simultaneously grounding it in the given 3D environment.

State-of-the-art papers The current state-of-the-art performance for downward-looking fisheye datasets, such as Mo2Cap2 (Xu et al., 2019) and the dataset introduced by Wang et al. (2022), is achieved by the method proposed in Wang et al. (2023b). It integrates scene constraints into pose prediction to avoid obtaining physically unrealistic poses like body floating or penetration with the environment. The approach consists of two primary steps. Firstly, the depth modality is inferred to capture the spatial information of the scene. Secondly, the inferred depth is inpainted in areas of the image where the body occludes the scene. The inpainted depth is then combined with 2D pose features in a shared 3D voxel space. Integrating scene constraints in this common 3D voxel space allows for pose estimation while enforcing adherence to the physical constraints imposed by the scene.

At the same time, the work presented in Li et al. (2023) achieves the best performance on a set of egocentric datasets (Luo et al., 2021; Zheng et al., 2022) captured from outward-looking camera perspective, including their proposed synthetic egocentric dataset. Given that directly matching egocentric video with full-body pose is challenging

due to the frequent absence of visible body parts, the authors address the task by introducing an intermediate step of head motion estimation. This approach eliminates the requirement for a training dataset of paired egocentric video and 3D human motion, while accurately predicting head motion using SLAM and a transformer-based model. Subsequently, a diffusion model conditioned on the estimated head pose is employed to derive the full-body pose. However, in this work, evaluation sequences only contain people navigating a virtual scene, and are not undergoing any activities.

Recent research has also placed significant emphasis on the task of estimating the complete body pose of individuals within the recorder's field of view. In the study by Ye et al. (2023a), the focus is on simultaneous localisation and human mesh recovery to reconstruct the global poses of individuals featured in egocentric videos, all without relying on dense 3D reconstructions of the surroundings. The proposed method, SLAHMR, firstly predicts relative camera motions, identifies individuals, and determines their local 3D poses. Leveraging this information, the model initialises trajectories for both humans and cameras within a common world reference system, optimising them for consistency across 2D observations in the video and learned human motion priors. Zhang et al. (2023c) takes a different approach by incorporating the 3D scene and conditioning a diffusion model for human pose generation on it. The authors combine human-centric scene regions with a physics-based collision score to guide the generation of plausible human poses that avoid environment penetration. To further enhance the accuracy and diversity of poses, they employ a visibility-aware graph convolution model, enabling the learning of precise body poses for visible joints while encouraging diversity in truncated parts.

Datasets Current datasets for pose estimation from downward-facing camera systems range from simulated (Tome et al., 2019) to real-world datasets (Rhodin et al., 2016; Xu et al., 2019; Wang et al., 2021a). These datasets provide ground truth in the form of 3D poses of the camera wearer. Recently, there has been a growing interest in generating large-scale real-world datasets, such as EgoPW (Wang et al., 2022), as well as simulated datasets with a diverse range of motions, such as UnrealEgo (Akada et al., 2022).

In the context of outward-looking camera setups, interest has also been increasing. In addition to the previously mentioned works (Luo et al., 2021; Li et al., 2023; Ng et al., 2020), datasets with orthogonal characteristics continue to be released. For example, the EgoBody dataset by (Zhang et al., 2022c) plays a crucial role in modelling interactions, as it encompasses multi-modal egocentric data streams and provides 3D ground truth for multiple individuals in complex 3D scenes. Furthermore, Zheng et al. (2022) recently introduced a large-scale dataset for human motion prediction that includes gaze information. They argue that accurate

motion prediction depends on understanding human intentions, which can be studied using gaze in the egocentric setting. Additionally, the EgoHumans benchmark by Khirodkar et al. (2023) captures multiple subjects in realistic outdoor environments from multiple egocentric viewpoints, serving as a valuable resource for multi-view multi-human analysis.

For the future Although some works (Yuan & Kitani, 2019; Xu et al., 2019) have prototyped real-time egocentric body pose estimation, performance is significantly below that of third-person (or remote) cameras. Body-oriented camera methods (Wang et al., 2023b) perform slightly better than those with outward-looking camera (Li et al., 2023) obtaining a MPJPE of 118.5 mm (mm) against MPJPEs ranging from 121.1 to 152.1 mm despite being tested on different datasets. Both still exhibit MPJPE values which are far from recent ones (Tang et al., 2023b) obtained on the Humans3.6M third-person benchmark (Ionescu et al., 2013) ranging around 20mm. Even in the context of estimating the poses of other individuals within the camera wearer's field of view, the current state-of-the-art (Ye et al., 2023a) achieves a World PA First - MPJPE, i.e. an MPJPE obtained by aligning the first frame of the prediction with the ground truth—of 141.1 mm on EgoBody (Zhang et al., 2022c). This performance remains notably distant from the results achieved with third-person perspectives. The recent release of more realistic datasets (Zheng et al., 2022; Khirodkar et al., 2023) can assist in bridging the gap between research and practical solutions.

Importantly, full-body estimation during natural activities, beyond navigation and full-body motion like jumping and squatting, is yet to be explored. For example, consider a person kneeling to retrieve an object from a cupboard, where their hand is occluded by the cupboard itself. Such poses are not available in any full-body datasets currently available. To date, the task of full-body estimation is not integrated with other tasks such as action understanding, trajectory forecasting, and hand-object estimation. It is thus difficult to assess the usefulness of current techniques in isolation.

4.8 Hand and Hand-Object Interactions

The significant presence of hands in egocentric videos and their primary importance in understanding humans' behaviour in an environment have lead to a proliferation of research on hands and their interaction with objects. While other research lines focus on human-object understanding from fixed cameras, egocentric vision provides a more fine-grained view into object interactions in which hands are central. As a result, methods for hand-object interaction understanding from egocentric vision greatly differ from human-object interaction detection from fixed cameras. In the next sections, we review works that focus on estimating *hand pose* (Sect. 4.8.1) and classifying *hand ges-*

tures (Sect. 4.8.2). Additionally, we analyse works that deal with hand-object interaction aiming at understanding how the camera wearer engages with the surrounding environment and the objects present therein. We divide hand-object interaction methods into those that estimate *2D information* (Sect. 4.8.3) vs others which exploit *3D meshes of hands and objects* (Sect. 4.8.4).

4.8.1 Hand Pose Estimation

Predicting the pose of hands from an egocentric viewpoint, especially during human activities, is a challenging task due to severe occlusions caused by object manipulations, limited field of view and head motion. The goal of hand pose estimation approaches is to efficiently regress 3D hand keypoints from various input signals such as RGB images, videos, depth maps, or 3D meshes. To evaluate the quality of the predicted hand pose, evaluation measures focus on the mean error for each hand joint or re-projection errors in meshes.

Seminal works First works on hand pose estimation exploited both RGB and depth signals, thanks to the availability of RGB-D sensors like Kinect. Oikonomidis et al. (2011) were the first to study the problem without requiring special markers and a complex hardware setup. Rogez et al. (2015a) analysed hands performing daily activities from the egocentric point of view, predicting their poses through a tracking-by-detection framework. Qian et al. (2014) introduced the first real-time system capable of accurately tracking a fully articulated hand. Keskin et al. (2012) used the depth sensors to address two tasks simultaneously: hand pose estimation and hand shape classification. Sabater et al. (2021) proposed a novel skeleton-based approach which is robust for predicting hand actions in different domains. Pose features are estimated using a temporal convolutional network, and aggregated to predict hand actions. Tang et al. (2013) were the first to explore the use of synthetic data to address the articulated hand pose estimation problem in a semi-supervised manner. They aimed at minimising the synthetic-to-real domain shift by leveraging a large synthetic dataset and a small amount of labeled real data. Synthetic data have been also used to predict the 3D pose of hands by Liu et al. (2021c). They proposed a unified approach which uses labelled synthetic and unlabelled real videos for joint 3D hand and object pose estimation. Similarly, Mueller et al. (2017) used synthetic data for real-time 3D hand tracking for estimating hand poses.

State-of-the-art papers Recently, different works focused on the optimisation and refinement of 3D hand pose estimation methods. Cheng et al. (2021) introduced HandFoldingNet, a network designed for estimating 3D hand joint coordinates from an input hand point cloud. The optimisation process is achieved through a guided folding step, which computes the 3D pose by leveraging a 2D hand skeleton. The folding step is further guided by multiscale features, representing both global and local information. Yang et al. (2022) presented a shallow deep neural network that incorporates specific layers capable of iteratively refining the predicted hand pose. Hand pose estimation has expanded beyond the use of depth maps and RGB signals. Rudnev et al. (2021) were the first to address this task using an event-based camera. They proposed EventHands, an approach which regresses 3D hand poses exploiting locally-normalised event surfaces, which is a new way of accumulating events over temporal windows. Of these works, only Yang et al. (2022) evaluated their method on egocentric hand pose, though the method was tested for general views.

Several works have leveraged hand pose estimation to perform action classification (Wen et al., 2023a) and hand reconstruction through neural representation (Karunratanakul et al., 2023; Lee et al., 2023). Wen et al. (2023a) built a framework which exploits the relationship between frames and the hand poses in an end-to-end manner. Given an egocentric video, a feature extractor encodes spatial information for each frame. Sequences of per-frame features are then fed to a hierarchical temporal transformer to capture temporal information. This transformer is composed of two parts, one for predicting the 3D hand pose and the other one for estimating the action.

Karunratanakul et al. (2023) presented an approach named HARP (HAnd Reconstruction and Personalisation) designed to create personalised hand avatars from short monocular RGB videos. They proposed a method to estimate a coarse hand pose and shape and then optimise the hand mesh, the albedo and the normal map using an analysis-by-synthesis strategy that compares the input image to the reconstructed ones. The HARP representation not only enhances the quality of 3D hand pose estimation but also allows for synthesising hand poses from new viewpoints. Lee et al. (2023) proposed the first neural implicit representation of two interacting hands, called Im2Hands. This approach enables the reconstruction of two interacting hands regardless of their resolution and geometry. It achieves this through two novel attention-based modules: one for initial occupancy estimation and the other for context-aware occupancy refinement.

Recently, Tse et al. (2023) presented a novel transformer-based approach that exploits multi-view RGB images to reconstruct two hands meshes directly. In particular, the proposed approach is able to reconstruct hands avoiding the use of deep network to regress hand model parameters. A larger dataset was used in Pavlakos et al. (2024) allowing an improved transformer learning for 3D hand mesh estimation. Their annotations include 5.3K egocentric images from EPIC-KITCHENS VISOR (Darkhalil et al., 2022) and 23.2K images from Ego4D (Grauman et al., 2022).

4.8.2 Hand Gestures

Hand gestures provide key information to enable human-computer interaction for AR/VR helmets, glasses and robots. Hands can be conveniently captured by wearable devices which are equipped with cameras able to observe the scene from the first-person view. While hand pose estimation and gesture recognition have been traditionally treated as separate tasks, they are inherently related. Recognising hand gestures can be seen as a discrete version of hand pose estimation focusing on understanding the semantics of gestures. Methods are usually evaluated with standard classification measures.

Seminal works The interpretation of hand gestures for human-computer interaction has been a topic of research for a while (Pavlovic et al., 1997). A pioneer work on hand gesture recognition in the context of egocentric vision has been presented by Baraldi et al. (2014). Inspired by dense trajectories approaches introduced for action recognition, they proposed to extract dense features around regions selected by a designed hand segmentation method, enhancing temporal and spatial coherence. In addition to RGB, other signals have been also used such as depth, skeleton information or stereo-IR. De Smedt et al. (2016) exploited time series of 3D hand skeleton to extract an informative descriptor for gesture classification, which is commonly employed for interacting with devices, such as *pinch, swipe right, swipe left and tap*. Molchanov et al. (2016) classified hand gestures considering depth, RGB and stereo-IR data streams through a recurrent 3D-CNN. 3D features have been also exploited by Cao et al. (2017). They proposed a novel spatiotemporal transformer module to classify gestures from RGB videos without explicitly detecting hands (e.g., hand detection or segmentation) and estimating head motion to rectify deformations. In the context of human-computer interaction with wearable glasses, Huang et al. (2016) proposed to study pointing gestures focusing on fingers, based on the observation that pointing gesture and its fingertip trajectory are crucial to recognise hand gestures like *pointing, selecting and writing*.

State-of-the-art papers Several works addressed the hand gesture recognition task from the egocentric point of view to enable human-device interaction, especially for AR/VR devices (e.g., smart glasses). Chalasani et al. (2018) proposed a deep network comprising an encoder responsible for extracting hand features, which are then fed into an LSTM to capture temporal patterns. RGB input sequences can be of an arbitrary length and repetitive gestures. Bai and Qi (2018) proposed a method to recognise hand gestures from a single depth camera, which can be integrated into VR/AR applications. They presented a two-stage method. First, they used a CNN to estimate the hand pose from bone lengths and joint locations. Then, they classified the gesture by leveraging

hand language. The latter is composed of four basic predicates (*pointing direction, relative location, fingertip touching and finger flexion*) which are applied to the six most important areas of the hand, the 5 fingertips and the palm.

In addition to human-computer interaction, the concept has also been extended to human-robot interaction. Papanagiotou et al. (2021) proposed a multi-task approach including gesture recognition to enable human-robot collaboration on an industrial assembly line. The main component is represented by a gesture recognition module which is based on 3D CNN trained on egocentric data acquired with a GoPro camera.

Some works proposed to use multiple signals to extract richer information. Chan et al. (2016) used HandCams (i.e., a wrist-mounted camera) together with a HeadCam. By using HandCams, it is not necessary to detect hands and infer manipulation regions as in classic egocentric approaches due to the fact that hands are always in the foreground. Considering this camera setting, the authors proposed a two-streams deep CNN, with one stream dedicated to the head and the other to the hands, respectively. Extracted features from both streams are then fused through concatenation and used to predict hand states (free vs. active), object categories and hand gestures. Abavisani et al. (2019) proposed a *multimodal-training/unimodal-testing* scheme, which involves sharing the knowledge between individual modality networks (e.g., RGB, Depth and Optical Flow) in the training phase in order to derive a common representation of hand gestures. To do this, a new spatio-temporal semantic alignment loss has been proposed which is similar to the covariance matrix alignment of the source and target features maps in domain adaptation methods. At inference time, each network has learned to recognise hand gestures from its specific modality but it also gained the common knowledge from the other networks.

Based on the idea that the background is not relevant for recognising hand gestures in AR/VR applications, Chalasani and Smolic (2019) focused on hand segmentation to improve gesture recognition accuracy. They proposed a new encoder-decoder architecture capable of generating embeddings from RGB images. These embeddings were then used for hand segmentation and gesture recognition simultaneously through multiple LSTMs.

4.8.3 2D Hand-Object Interaction

2D hand-object interaction methods aim to associate each hand with one or more objects present in the scene, thereby determining their relationship (e.g., the hand is holding a plate). Formally, this task involves detecting and recognising the hands of the user, along with the objects involved in the interaction. To do so, methods have been developed to predict hand-object interactions by estimating information such as 2D bounding boxes or hand-states (i.e., contact

or no-contact). The performance of hand-object approaches is assessed by evaluating their classification and regression abilities. 2D object interaction methods also predict object state changes and object transformations.

Seminal works Relations between objects and tasks are important for critical modelling activities and behaviour. Damen et al. (2014) defined and discovered *Task Relevant Object (TRO)* which refers to an object or a part of it that a human interacts with while performing a specific task, in an unsupervised approach. Crucially, they also aimed to distinguish and classify the different *Modes of Interaction (MOI)* with these TROs. Cai et al. (2016) explored the relations between hands and objects to detect the grasped part of an object during human manipulation. They extracted object attributes such as the *thick or long* shape of a bottle, and observed how these attributes influenced the type of grasp used. Their unified model combines the prediction of the object, its attributes, the grasp type, and the action performed. Rogez et al. (2015c) formalised the problem of classifying handled objects using both RGB and depth signals. Depth provides additional information on the exact touch/contact between the hand and the objects present in the environment. Liu et al. (2017) focused on the effect of interactions on objects (e.g., a mug can be empty or full).

State-of-the-art papers Shan et al. (2020) proposed a method to detect and localise hands in the scene, distinguishing between left and right hands. Additionally, they aimed to classify objects into two classes: *active* or *passive*. In particular, if an object present in the scene is in contact with at least one hand, it is considered as *active* object, otherwise, it is considered *passive* object. They also considered 5 different contact states: *no contact, self contact, other person contact, portable object contact* and *stationary object contact (e.g., furniture)*. While originally designed for YouTube videos, a modified model with additional annotations was successfully used to automatically annotate EPIC-KITCHENS-100 (Damen et al., 2022) with hands and active objects. Grauman et al. (2022) introduced the task of object state change detection and classification. The task involves distinguishing transformative interactions from those that are purely translational. Detecting the temporal moment at which an object changes state during transformation is introduced with manual annotations.

A related task to object transformations is tracking objects in egocentric views. Dunnhofer et al. (2023) analysed the performance of state-of-the-art visual trackers in the egocentric domain highlighting challenges in the ego domain.

Although the analysis of hand-object interactions mostly involves bounding box annotations, a few works have focused on studying hand-object relations using semantic segmentation mask annotations (González-Sosa et al., 2021; Zhang et al., 2022a; Darkhalil et al., 2022; Tokmakov et al., 2023). These works focus on hands and active objects semantic seg-

mentation considering egocentric images (González-Sosa et al., 2021; Zhang et al., 2022a) or videos (Darkhalil et al., 2022; Tokmakov et al., 2023). Darkhalil et al. (2022) defined and predicted hand-object relations, including cases where the on-hand glove is in contact with an object in the environment. After segmenting active objects, a binary classifier is used to predict the state of each hand as well as the object-in-contact in each case.

Linking 2D to potential 3D information, by predicting 2D hand-object relations, Qian and Fouhey (2023) addressed the task of understanding what a user is able to do (i.e., how can I manipulate the objects in an image?) considering the environment where the user is. They introduced a transformer-based encoder-decoder which takes in input an image and a set of 2D query points to predict the potential interaction. In particular, for each query point, the transformer head predicts an interaction represented by depth, surface normal of the objects, physical properties and affordance.

4.8.4 3D Hand-Object Interaction

The task of 3D hand-object interaction predicts 3D information about the hands and objects involved in observed interactions through 3D bounding boxes, 3D meshes as well as 6 DoF hands and objects poses. Performance is measured by metrics such as 3D mean joint position error for hands, symmetric Chamfer distance for objects, differences in 6 DoF pose including translation and rotation errors as well as re-projection errors including IoU metrics of the 2D re-projections.

Seminal works Tekin et al. (2019) proposed an end-to-end framework to understand 3D human-object interactions from still RGB images. The model takes as input a single RGB image and estimates hand and object poses, recognises objects and predicts the class of the activity. Garcia-Hernando et al. (2017) used 3D hand poses, 6D object poses and RGB-D images to classify hand actions. In Chen et al. (2019) hotspots from hand touch are automatically detected and associated with actions from egocentric videos capturing the camera wearer using a sewing machine. Hasson et al. (2019) studied the problem of reconstructing hands and objects during manipulation, in the case the latter is affected by occlusions. They proposed a new architecture composed of two branches, one for the object shape and the second one for the hand mesh. Differently from previous works which focused on instance-level human-object interactions where 3D models and sizes of objects are known beforehand, Liu et al. (2022c) studied human-object interactions considering the vast diversity of objects in our daily life. They addressed this task by exploiting 4 dimensions of input data: the scene point clouds and object meshes (3D) along the time interval (1D).

State-of-the-art papers Few works have recently addressed the 3D hand-object interaction task. Chen et al. (2023) intro-

duced a geometry-driven signed distance function (gSDF) method that incorporates robust pose priors, leading to improved hand-object reconstruction by disentangling pose and shape estimation. Fan et al. (2023) proposed two novel tasks based on hand-object interactions: consistent motion reconstruction and interaction field estimation. They also presented two novel approaches to address these tasks. ArcticNet is an encoder-decoder architecture able to reconstruct the motions of both hands and the articulated object, while InterField estimates, for each hand vertex, the distance to the closest object mesh.

Temporal information has also been considered to estimate 3D hand poses and actions (Wen et al., 2023a), and it has a relevant role even in the work of Hampali et al. (2023) which proposed a novel method based on UNISURF (Oechsle et al., 2021) to reconstruct 3D objects during hand-object manipulation. Given a sequence of RGB frames in which a hand is manipulating an unknown object, the method captures both geometrical and appearance features of the object by constructing a neural implicit representation. The latter is then used to reconstruct the object. Differently from other NERF-based methods, the proposed approach assumes that the camera pose is not available.

Datasets For *hand pose estimation*, Yuan et al. (2017) acquired a large scale dataset named BigHand2.2M, which covers a wide and dense range of hand poses. The dataset contains 2.2 million depth maps annotated with hand joints, utilising six 6D magnetic sensors and inverse kinematics. Since hand poses become more complicated when involving object interactions, Ohkawa et al. (2023) published the AssemblyHands dataset which includes synchronised egocentric and exocentric images sampled from the Assembly101 dataset (Sener et al., 2022) in which users assemble toy vehicles. The dataset is composed of 3.0 million images and has been labelled with high-quality 3D hand poses, using a proposed automatic annotation model that exploits the exocentric view.

Tackling both *hand poses* and *gesture recognition*, Vakunov et al. (2020) acquired a dataset composed of real and synthetic images. The data collection has three sets created to address different aspects of the problem: 1) capturing "hands in the wild" with geographical diversity, varying lighting conditions, and diverse hand appearances, 2) covering a wide range of angles representing all physically possible hand gestures, and 3) incorporating synthetic data to enhance the study of hand poses and gestures.

Among the large datasets specifically focusing on *hand gestures* Huang et al. (2016) proposed EgoFinger which is composed of egocentric videos of different pointing gestures acquired in multiple scenarios. The dataset contains 93.729 RGB frames and it has been collected by 24 subjects in 24 indoor/outdoor scenes. To scale up research on hand gestures, Zhang et al. (2018) introduced the EgoGesture dataset,

which comprises 24,000 gesture samples (RGB and depth) acquired by 50 different subjects. The dataset contains 83 classes of static and dynamic gestures, designed specifically for interaction with wearable devices.

To study *hand-object interactions* many datasets of *real* images and videos have been proposed. Shan et al. (2020) collected the 100 Days of Hands (100DOH) Internet scale dataset to enhance size and diversity in hand-object research. It consists of 100K frames acquired over 131 days in which humans were involved in 11 categories of interactions labeled with bounding boxes around the hands and the active object, hand side and hand contact state (indicating if there is a contact between the hand and an object or not). Lu and Mayol-Cuevas (2021) introduced a dataset to study hand poses during manipulation with objects. The dataset has been captured in a multi-cam setting using two HD cameras and an iPhone 12. The authors collected 2000 pairs of hand-object interactions performed with a single right hand. GUN-71 is composed of 12K frames annotated with 71 action classes and 28 object classes. THU-READ (González-Sosa et al., 2021) is an egocentric dataset composed of 960 RGB-D videos captured from people performing 40 different daily-life interactions. The data is labeled with pixel-wise annotations of egocentric objects and hands.

Liu et al. (2022c) presented a large-scale dataset named HOI4D. It is composed of 2.4 million RGBD egocentric video frames acquired in indoor environments where people interact with 800 object instances. It has been labeled with a rich set of 2D and 3D annotations. In particular, hands are annotated with their pose, while objects have labeled segmentation masks, 3D poses and also their CAD models have been released. The MECCANO dataset by Ragusa et al. (2021, 2023b) focuses on human-object interactions in an industrial environment where 20 subjects assemble a toy model of a motorbike. It is composed of 20 videos with average duration of 20.79 min and it is multi-modal, comprising synchronised gaze signals, depth maps and RGB videos.

Unlike other datasets that primarily focused on rigid object manipulation, Fan et al. (2023) introduced the ARCTIC dataset, which is specifically designed for interactions involving hands manipulating articulated objects, such as scissors or laptops. This dataset is unique as it includes paired 3D hand and object meshes along with detailed dynamic contact information.

Only a few works have specifically focused on hand-object interactions with fine-grained information. Zhang et al. (2022a) annotated 11K egocentric hand-object interactions with semantic segmentation masks and contact boundaries. These images have been collected from three existing datasets: Ego4D (Grauman et al., 2022), EPIC-KITCHENS (Damen et al., 2018) and THU-READ (González-Sosa et al., 2021). Darkhalil et al. (2022) extended EPIC-KITCHENS-100 dataset with pixel-level annotations, obtaining 272K

semantic masks interpolated to 9.9M dense masks. As a result, they captured long-term object segmentations of the same instance which is subject to a series of transformations during hand-object interactions. Recently, Tokmakov et al. (2023) presented VOST, a dataset composed of 713 videos in which 51 different object transformations (i.e., objects which dramatically change their appearance), have been annotated with segmentation masks. Videos have been annotated at 5 fps obtaining 76K annotated frames.

Recent works investigated the use of egocentric *synthetic* data to mitigate the need of annotating real domain-specific data for model training in *hand-object interaction*. Real data with ground truth labels are difficult to obtain. Acquiring egocentric images/videos of hand-object interactions as well as manually annotating hands and objects with keypoints, 2D and 3D bounding boxes, semantic masks, relations and action descriptions are a time consuming and expensive task. Rogez et al. (2014) is a pioneer work, in which synthetic images have been generated to demonstrate their potential for the 3D hand pose detection task. They focused on hand pose estimation while humans perform object manipulation, proposing a photorealistic synthetic model of egocentric scenes to generate training data for learning depth-based pose classifiers. Hasson et al. (2019) proposed ObMan, a large scale synthetic dataset made of images of hands grasping objects. By randomising the background and selecting images from the LSUN (Yu et al., 2015) and ImageNet (Russakovsky et al., 2015) datasets, they successfully generated 20K diverse hand-object interactions.

In the pursuit of synthetic data generation and annotation, various studies have directed their attention towards creating photorealistic datasets and tackling the domain-shift problem that emerges when transitioning between real and synthetic domains. Leonardi et al. (2023) presented a comprehensive pipeline and framework for automatically generating egocentric hand-object interactions, including annotations such as depth maps, semantic segmentation masks, bounding boxes for objects and hands, as well as attributes and their respective 3D distances. In the work of Ye et al. (2023b), diffusion models have been used to generate complex hand-object interactions, allowing reasoning about *where* to interact and *how* to interact. Tendulkar et al. (2023) proposed FLEX, a framework capable of generating full-body and hands grasping poses for everyday objects. FLEX is able to synthesise a wide range of natural grasping poses, ensuring diversity and generalisation, while considering 3D geometrical constraints in complex scenes. Recently, Xu et al. (2023b) utilised tactile sensing for in-hand object reconstruction. Given the difficulty of obtaining ground truth data for object deformation, they addressed this challenge by synthesising images using the proposed simulator. Despite the significant progress towards narrowing the gap between synthetic and real domains, the problem has not been solved and there is space for future investigations.

For the future Even though there has been advancement across various research areas related to hand analysis, current approaches still come with their own set of limitations. State-of-the-art works in *hand pose* focus on the analysis of the posture considering different signals (Rudnev et al., 2021; Yang et al., 2022) or challenging scenarios in which both hands interact simultaneously (Lee et al., 2023). At this stage, methods are able to predict hand pose in a large variety of domains thanks to the availability of large datasets acquired and labelled explicitly for these domains. However, they typically fail in predicting the hands pose when hands are involved in the interaction with objects and in complex scenarios. The state-of-the-art approach for 3D hand pose estimation (Ohkawa et al., 2023) achieves a MPJPE of 23.46 mm on the test set of AssemblyHands dataset. Although these results are promising, there is still room for further research and advancements in this field.

Gestures have been studied to allow humans to interact with devices such as AR/VR glasses (Bai & Qi, 2018) or robots (Papanagiotou et al., 2021). The current state-of-the-art performance on the gesture recognition task (Chalasani et al., 2018) achieves a high accuracy of 96.9% on the test set of the EgoGesture dataset. These results confirm that the performance of gesture recognition models is comparable to that of humans when considering a discrete number of gestures. However, there is still room for exploration when larger numbers of gestures are considered. Nowadays, AR/VR devices have a gesture recognition system able to recognise simple gestures like *push, pinch, point, air tap* useful to interact with the system and with virtual objects placed in the environment (i.e., holograms). Furthermore, devices such as Xreal, implemented custom gestures to interact with the device such as *victory* or *open hand* as well as HoloLens2 and Apple Vision Pro allow to use gestures using both hands and gaze. However, the total number of gestures that can be recognised is small, even though users can usually implement custom gestures specifically for their devices.

For *hand-object interaction*, specific sets of interactions in constrained environments, such as kitchens (Darkhalil et al., 2022) and industrial workplaces (Ragusa et al., 2023b; Leonardi et al., 2022), are starting to be analysed. Despite a few initial efforts to develop approaches capable of understanding generic interactions (e.g. Shan et al. (2020)), we are still far from having robust methods that can generalise across objects and environments, for example industrial environments where hands are in contact with both large machines as well as small tools and objects (e.g. screws). One of the main problems lies in the availability of datasets explicitly labelled with human-object interactions, as it requires a significant amount of effort to acquire and manually label such data. A promising direction is to leverage automatically labelled

synthetic data, as it enables the acquisition of large-scale hand-object interactions across multiple domains and diverse sets of objects (Tendulkar et al., 2023; Leonardi et al., 2022). However, synthetic images often lack the level of photo realism of real-world images. Additionally, dealing with object transformations under manipulations is a very challenging task that has stated to be studied only recently (Darkhalil et al., 2022; Fan et al., 2023).

Actually, the state-of-the-art performance on the Hand-Object Interaction Segmentation task (Darkhalil et al., 2022) achieves a Hand Mask Average Precision of 95.6% and an Active Object Average Precision of 25.7% on the test set of the EPIC-KITCHENS VISOR dataset. On the other hand, performance on the object state change task on the test set of Ego4D (Grauman et al., 2022) reached an accuracy of 67.6% and an Average Precision of 15.5%. These results highlight the difficulty in bridging the gap between action perception and the relations between actions, objects, and the environment.

4.9 Person Identification

Person identification is very relevant for surveillance and security applications and has been extensively studied in third-person literature while it has been less investigated from the egocentric point of view. While person identification from first person cameras can leverage some algorithms already investigated from remote cameras. Particularly, in egocentric vision, person identification includes two distinct sub-tasks: recognising people in the field of view of the camera, and identifying the camera wearer. Both rely on the definition of a robust representation of faces as well as other body parts and their movement. The goal of the former is recognising whether two images depict the same person, or searching for a person within a gallery given a reference image query. More precisely, in face recognition the focus is only on faces and the gallery contains a pre-defined set of identities. Person re-identification considers instead the whole body, and the gallery contains many distractors without specific identities. Different from remote cameras, the observed individual is often truncated or highly occluded due to the camera being at human head height, compared to remote cameras which are elevated, decreasing the amount of occlusions. The performance is assessed either in terms of accuracy by considering image pairs and their prediction (each pair with positive/negative label), or by counting how many of the true match appears at the top of the ranked gallery.

The identification of the person who wears the camera can be formalised as classification in a closed-set scenario, or matching in an open-set one. It involves the wearer's hands and gait, with applications in theft prevention and personalisation.

Seminal works Initial efforts to address person identification on wearable devices primarily focused on facial recognition. For instance, Farringdon and Oni (2000) were pioneers in developing a wearable application that automatically identified and stored faces to enhance the camera wearer's memory. This approach was further expanded to various ending goals. Krishna et al. (2005) developed iCare, an interaction assistant for the visually impaired. It recognises individuals in the scene and notified the wearer through audio signals. Wang et al. (2013) targeted prosopagnostics patients, i.e. people who cannot distinguish faces, proposing a system that displayed the identity of people in the scene directly on a screen mounted on the wearable device. Thomaz et al. (2013) explored the use of face detection, image cropping, location- and motion-based filtering to remove privacy-sensitive information from collections of egocentric images while still allowing to carry out the downstream task of eating behaviour recognition. Additionally, Sajjad et al. (2020) proposed a system for enhanced law enforcement that collects data from wearable devices to identify suspects or missing individuals.

Some works proposed to explore external sources of knowledge for specific objectives. Kurze and Roselius (2011) coupled the face recognition system with data from social networks in order to automatically link the person retrieved with the corresponding online information. Chakraborty et al. (2016) considered a face recognition model running on a network of wearable cameras, extending the system developed on Google Glasses (Mandal et al., 2015) for face re-identification. Fergnani et al. (2016) proposed a method for full-body person re-identification: a metric learning approach that evaluates instance similarity by dividing images into meaningful body parts and considering a part-related weight defined from human-gaze information. Basaran et al. (2018) exploited additional metadata collected by mobile phones to reduce the search space. Their approach predicts the next moving camera where the target may appear and aggregates temporal information within sequences of body parts to re-identify individuals.

Three cross-view works are at the interface between recognising the camera wearer and recognising bystanders. Yonetani et al. (2015) considered a scenario where multiple people are wearing a camera and recording each other. In such cases, the work proposes to use motion correlation between the target person's video and the observer's video to uniquely identify instances of oneself, which can be useful for privacy filtering. Poleg et al. (2015a) leveraged head motion patterns to identify the camera wearer in other videos recorded simultaneously, both from third-person and egocentric perspectives. The head-motion signature allows the wearers to recognise themselves in videos and decide whether to keep or delete them. Fan et al. (2017) proposed to use both a third-person camera capturing the scene and multiple subjects recording from an egocentric view. The goal

is to match people across different views and identify the source of the egocentric video in the third-person one.

The identity of the camera wearer from the single egocentric perspective is challenging to discover due to the limited field of view often obscuring the wearer's body. Shiraga et al. (2012) attempted to recover this information using a complex system of stereo cameras on the user's backpack to analyse motion during walking. On the other hand, Hoshen and Peleg (2016) proposed an approach based solely on a front camera and the head motion signature present in the egocentric video. This approach tries to match the identity of the camera wearer across different videos, with potential applications in theft prevention. In a different setting, Ardeshir and Borji (2016) utilised both egocentric and top-view camera recordings to match each camera wearer with their corresponding top-view identity. Furthermore, Thapar et al. (2020b) demonstrated that even using hand motion, captured in the form of dense optical flow, can reveal the identity of the camera wearer across various activities and subjects.

State-of-the-art papers For person re-identification, Choudhary et al. (2020) recently proposed to leverage third-person large-scale datasets in the egocentric domain. They employed Neural Style Transfer (NST) to generate high-quality images with a fixed camera style from the egocentric ones. A content loss ensures that the architecture maintains coherent predictions despite style differences, and a style loss is adopted to improve the transfer capabilities of the model. The most recent work combining top-view and egocentric videos is the one presented by Ardeshir and Borji (2018). It proposes a learning method that matches corresponding pairs of bounding boxes from egocentric and top-view surveillance videos, combining also some geometrical and spatiotemporal reasoning. The former evaluates the probability of each identity being present in the field of view of the camera holder on the basis of the output of a multiple object tracking algorithm. The latter defines a cost for assigning the same identity label to a pair of bounding boxes depending on whether they are present in the same frame and if they overlap in temporally nearby frames. This approach achieves state-of-the-art performance in both self-identification and re-identification.

In the context of identifying the wearer from the egocentric perspective, Thapar et al. (2020a) introduced EgoGaitNet, a model capable of extracting the wearer's gait from the optical flow of an egocentric video, enabling the matching of videos from the same wearer. Additionally, they propose a Hybrid Symmetrical Siamese Network that can match third-person views of a subject with their egocentric videos recorded at different times, raising important privacy concerns. Tsutsui et al. (2021) explore various sources of information related to the wearer's hands to investigate the feasibility of wearer identification across different videos. They experiment with both RGB modality containing the texture of the hands and the depth modality providing information about their shape.

Furthermore, they extract the silhouette of the hands from the depth information and get a multi-modal robust representation for their experiments.

Datasets EgoSurf, is a small-scale dataset introduced by Yonetani et al. (2015). It consists of egocentric videos recorded by individuals engaged in face-to-face communications, captured in eight different scenes (four indoors and four outdoors) by two or three people. Its purpose is to match the camera wearer of a video with their third-person view from another person's egocentric recording. The Ego2Top dataset was introduced by Ardeshir and Borji (2016) and allows both self-identification and re-identification. It comprises 50 top-view and 188 egocentric videos, amounting to approximately 225 thousand frames. However, the number of distinct identities in this dataset is relatively limited.

The Egocentric Video Photographer Recognition (EVPR) dataset by Hoshen and Peleg (2016) includes videos featuring 32 subjects: it was created with two distinct types of cameras and primarily used for egocentric camera wearer identification. Thapar et al. (2020a) additionally created the IITMD-WFP and IITMD-WTP datasets to investigate potential biometric signature leakage in egocentric videos. The IITMD-WFP dataset encompasses 3 h of videos recorded by 31 different subjects, while the IITMD-WTP dataset serves as the third-person counterpart.

For the future Third-person face recognition and person re-identification research has recently took great advantage of the development of transformer modules able to produce robust feature representations (Liao & Shao, 2021; Zhang et al., 2023b), and of self-supervised pretraining (Fu et al., 2021; Zhu et al., 2022; Fu et al., 2022a). However egocentric person identification is still lagging behind. The mean Average Precision (mAP) of state of the art approaches (Wieczorek et al., 2021) over fixed camera benchmarks (Zheng et al., 2015) is close to 98% while egocentric models (Choudhary et al., 2020) barely reach 65% on small scaled egocentric datasets. In the identification of the camera wearer, the low mAP results by Fan et al. (2017) indicate that current research is still in the proof of concept phase. In particular, future developments should focus on producing larger and more diverse egocentric person Re-ID benchmarks in order to train reliable models directly from egocentric data.

4.10 Summarisation

The increasing prevalence of wearable cameras has led to a proliferation of long and unstructured video recordings documenting people's lives. However, users may not revisit much of this recorded content, and important events can be hidden among repetitive or uninteresting segments. Video summarisation is a valuable task that aims to produce a concise summary of the input recording. Current methods produced summaries in different forms. Keyframe-based summaries

involve the selection of a sequence of relevant frames to represent the most critical events or information in the video. Video skimming approaches segment the video into relevant portions and then collect them to produce a reduced version of the initial recording. Finally, fast forwarding techniques prioritise significant sections while reducing the reproduction speed of less important segments without necessarily trimming any section. The most commonly used metric for evaluating video summarisation is the F1-score.

Seminal works Traditionally, determining the most pertinent segments of a video for summarisation was achieved through heuristic criteria such as visual saliency (Itti et al., 1998), motion (Wolf, 1996), or high-dimensional curve simplification (DeMenthon et al., 1998).

Egocentric video summarisation was first introduced in Aizawa et al. (2001), which starts by segmenting the long egocentric recording based on visual motion features and then adopts brainwave signals to determine the subjective interest of the camera wearer throughout the subshots. This innovative approach, which optimises the summarisation based on the viewer's brain response has not been explored further.

In Lee et al. (2012), the approach relied on RGB data along with higher-level features to produce object-driven storyboards across multiple environments. In contrast, Lu and Grauman (2013) generated the summary as a coherent set of subshots in a story-like manner. The use of web images as a prior to skip uninformative views of objects caused by the motion of a hand-held camera was proposed in Khosla et al. (2013). Zhao and Xing (2014) were the first to consider the online aspect of summarisation, allowing the processing of arbitrarily long videos in real-time.

Concurrently, there has been a focus on developing fast-forward video summarization techniques, primarily aimed at generating hyperlapse videos. The intent here is twofold—to stabilise the captured footage and simultaneously highlight the most salient sections. Specifically, the method presented by Kopf et al. (2014) achieves this by rebuilding the three-dimensional spatial geometries within the environment and then sampling a virtual camera path from which the output video is reconstructed. Okamoto and Yanai (2014) advance this approach by incorporating semantic considerations into the video content. This method prioritises certain segments of the footage, such as crosswalks, which bear a greater significance in navigational guidance videos.

Additional input modalities were used for personalising summarisation, such as feeding textual concepts of interest in Sharghi et al. (2016), adopting the gaze modality to understand the wearer's attention in Xu et al. (2015), or using sound in the form of psychoacoustic metrics to discard moments with unpleasant noises in Bajcsy et al. (2018). Aesthetic aspects of keyframes were explored in Xiong and Grauman (2014), which used web photos as a prior to iden-tify video frames that resemble intentional snapshots, defined as snap points, and demonstrated improved performance in downstream summarisation tasks. Bettadapura et al. (2016) used quality measures together with GPS data to extract picturesque highlights from large amounts of egocentric video. Social networks were used by Ramos et al. (2020) to mine topics of interest and fast-forward the video in a semantically-coherent manner.

Egocentric touristic recordings are the main target for video summarisation. Xiong et al. (2015) experimented on egocentric sequences collected at Disneyland and proposed a storyline representation with actors, events, locations, and objects, allowing story-based queries across different tracks. Similarly, Varini et al. (2017) focused on providing a touristic summary more dependent on user preferences. This was achieved by adopting metrics based on the wearer's attention, semantic coherence with preferences, and a narrativity grade to effectively extract sub-shots of interest.

Finally, with the advent of deep learning, video summarisation began to adopt two-stream CNN models that consider motion and appearance as two complementary aspects in understanding the highlight score of a video segment (Yao et al., 2016). LSTMs were also used by Zhang et al. (2016) in a supervised video summarisation setting. By modelling variable-range dependencies among frames, the approach captures some high-level temporal understanding which is necessary to avoid relying solely on visual cues.

State-of-the-art papers The general video summarisation literature mainly trains and/or evaluates using two benchmarks: SumMe (Gygli et al., 2014) or TVSum (Song et al., 2015), where respectively just 4 out of 25 and 5 out of 50 videos are taken from an egocentric perspective. The current state-of-the-art in this setting is achieved by He et al. (2023) using a multimodal summarisation method that adopts a transformer-based architecture and an alignment-guided self-attention module to exploit the time correspondence between video and text modalities, and inter- and intra-sample contrastive losses.

However, egocentric videos pose unique challenges due to the significant head motion and long ordinary portions, with the camera wearer moving through a variety of scenes in order to perform daily activities. Consequently, the noted state-of-the-art above is not particularly designed to handle these challenges, though direct evaluation of this model on egocentric benchmarks has not been carried out.

The latest approach for personalised egocentric video summarisation is proposed in Nagar et al. (2021). The approach is customisable to adjust both the length and summary content and uses a reinforcement learning (RL) approach on top of C3D features. The RL action involves either selecting or discarding the sub-shot, and the approach uses basic rewards (distinctiveness, indicativeness, and overall length) as well as customisable ones (social interaction,

face identity, and customised target length). While developed for egocentric videos, the approach uses outdated features and using more recent architectures and/or features is yet to be assessed.

Recently, a novel query-focused approach was introduced by Wu et al. (2022b) to provide an interactive method for video summarisation. The authors developed a framework called IntentVizor, which enables the formulation of generic multi-modal queries and facilitates interactive editing of video summaries. Although the authors experimented just with textual and image queries, its underlying principle is centered around the notion of intent, referring to the high-level requirements of the user, irrespective of the modality of the input query. The intent is determined based on a learned distribution of users' needs, which takes into account various query inputs. Furthermore, the proposed Granularity Scalable Ego-Graph Convolutional Network (GSE-GCN) establishes correlations between the video features and the generic intent, thereby facilitating the extraction of noteworthy sections within the video.

Finally, Elfeki et al. (2022) is the first work on egocentric multi-stream summarisation, which summarises the videos of multiple wearable cameras intermittently sharing the field of view. The authors proposed a multi-view extension of the Determinantal Point Process, processing all the camera recordings in parallel and selecting inter-stream diverse events and the best ego-camera viewpoint for each event.

Datasets The progress in egocentric video summarisation is significantly hindered by the fact that many datasets used in previous studies have not been publicly released and remain confined to specific research projects. As a result, despite having just roughly 15% of egocentric videos, SumMe (Gygli et al., 2014) and TVSum (Song et al., 2015) are the most commonly used benchmarks. These benchmarks provide frame-level interestingness scores, enabling automatic evaluation of summarisation results without the need for user studies.

The FPVSum dataset by Ho et al. (2018) is a more recent contribution which tries to mitigate this issue. Indeed, it is composed just of egocentric videos, but only partially annotated. The authors also incorporated unlabelled egocentric data to develop a summarisation model capable of better generalisation to the egocentric domain in a semi-supervised manner.

In fast-forward summarisation methods, the Dataset of Multimodal Semantic Egocentric Videos (Silva et al., 2018) stands out as an extensive dataset. Spanning 80 h, this dataset provides valuable information about the activities being performed, the attention of the recorder, and the presence of interactions. By incorporating the recorder's interest scores across a wide range of object categories, it gives the possibility to evaluate both the smoothness and the semantic highlights of the summary.

For the future Despite Nagar et al. (2021) having developed an approach capable of handling day-long recordings using sliding windows, thus avoiding the need to feed the entire video as input to the model, this task still remains far from being considered solved. In fact, on the TVSum and SumMe datasets, the current state-of-the-art (He et al., 2023) achieves F1-scores of 63.4 and 55.0 respectively, while even the more intriguing query-focused egocentric summarisation presented in Wu et al. (2022b) only manages to achieve a F1-score of 50.9 on the egocentric query-based dataset presented in Sharghi et al. (2017). This limited performance is due to the restricted number of test query-concepts, which still falls short of reflecting the real-life scenarios depicted in the envisioned EgoAI scenarios.

4.11 Dialogue

Fostering the integration of vision and language has great potential in advancing human-machine interaction, allowing artificial agents to dialogue as they can *see* and *communicate* in a natural way. To this end, the research community has been working mainly in two directions: visual question answering (Sect. 4.11.1), and more general ego-language models (Sect. 4.11.2).

4.11.1 Visual Question Answering (VQA)

Visual Question Answering (VQA) consists of developing systems able to answer questions related to the semantic content of images and videos. Thus, the system takes a visual and a language input and produces a language output, or a referral to a part of the image or video. This is a crucial task for the development of EgoAI able to support users and skills training. VQA is also considered an effective way to investigate the reasoning capabilities of deep models as the questions can be designed to obtain not only descriptive answers but also complex predictive and explanatory outputs.

The adopted settings for VQA are mainly two: multiple-choice VQA and open-ended VQA. The first one is formalised as a classification problem and evaluated on the basis of prediction accuracy. The latter is more challenging and realistic as it requires either identifying the correct answers in a large pool of candidates or generating a free-form response. In these cases, model assessment is performed by measuring how often the ground-truth answer is selected in the top predicted choices (recall@k) or via language metrics (e.g. ROUGE) and user studies.

Seminal works Gurari et al. (2018) were the first to highlight the need for egocentric views in real VQA applications to support blind people. The task poses interesting peculiar challenges as some of the images do not contain enough information, so the associated question is not answerable. When dealing with videos, the perspective of VQA has ini-

tially shifted from third- to first-person to support the training of navigational agents in indoor environments, solving a task originally referred to as embodied question answering (Das et al., 2018; Yu et al., 2019). After having observed a sequence of egocentric visual frames of a synthetic scene, the agent is interrogated on the position of single or multiple target objects, or asked to plan a series of actions conditioned on the questions (Gordon et al., 2018). Wijmans et al. (2019) and Anderson et al. (2018) studied the same task in photo-realistic scenes obtained by involving point cloud perception and 3D simulators.

In all the referred works, each VQA episode is considered in isolation, with no memory or information persistence among the episodes. To address this limitation, Gao et al. (2021) proposed a method that decomposes long videos into separate events and exploits multi-step temporal attention. Fan (2019) moved the focus from synthetic to real-world scenarios with a human egocentric view, aiming for a better understanding of the footage of wearable cameras. This work discussed the shortages of third-person VQA methods when applied in the egocentric setting and pointed out the need for simultaneous estimation of ego-motion and third-person motion, while disentangling attention for first and third-person activities to find out relevant visual content.

State-of-the-art papers In the VQA literature for visually impaired people, Chen et al. (2022) have investigated how to ground the answers by segmenting the relevant image region. The work by Dancette et al. (2023) focused instead on how to avoid answering when the visual information is not sufficient. They proposed to train multimodal selection functions that indicate for which samples the model can be generalised, and which samples are too hard and should be abstained on.

Among the embodied question answering works with egocentric videos recorded by robotic agents, Zhu et al. (2023b) recently proposed a new reinforcement learning framework involving multiple phases of environment exploration and reasoning. Ma et al. (2023a) discussed the challenges of situated VQA in 3D scenes. Several publications have also presented novel solutions to answer questions related to long real-world videos. In particular, the approach proposed by Gao et al. (2023) decomposes traditional dense spatial-temporal self-attention into cascaded segment and region selection modules that adaptively select frames and image regions closely relevant to the question.

Another sub-part of the state-of-the-art egocentric VQA literature targets episodic memory to search for the temporal window that shows a frame relevant to the question, complemented by informative language answers. Starting from the Ego4D challenge on *Episodic Memory—Natural Language Query*, Bärmann and Waibel (2022) defined the QAEGO4D dataset with textual answers from human annotators. The authors have also benchmarked several baseline methods on the newly introduced testbed, using a temporal sequence of

feature vectors rather than on raw video data to limit the memory and computational burden. Datta et al. (2022) studied the same problem from a stream of RGB-D images and proposed a method that combines the semantic features extracted from egocentric observations into a single top-down feature map of the scene. This helps to create a consolidated spatiotemporal memory which is provided as input to an encoder-decoder architecture that grounds answers to questions.

Ego4D (Grauman et al., 2022) proposes the episodic memory challenges towards querying long-term videos with natural language questions. As an example, Ramakrishnan et al. (2023) adopted also Ego4D narrations to overcome the scarcity of query-response pairs.

The most recent research trend is on goal-oriented VQA to improve reasoning models and get a deeper task understanding from egocentric videos. Jia et al. (2022a) introduced the EgoTaskQA benchmark with questions designed to investigate actions that imply world state transitions, agents' intents in task execution, and their belief about others in collaboration. The same work presented an extensive analysis of several VQA methods highlighting the effective support provided by large language models. Wong et al. (2022) defined the affordance-centric VQA problem where the AI assistant should learn from instructional videos to provide step-by-step help in the user's view. The authors introduced a new dataset and developed a novel question-to-action model based on an encoder-decoder architecture. More precisely, the encoder is composed of multiple modules that extract features from video, script, question, and answers. The decoder performs cross-attention among the information obtained from the different modalities and produces operational localised answers (text and bounding boxes) in multiple ordered steps.

Datasets The research on Egocentric VQA is still in its infancy and rapidly evolving with many publications proposing specific subtasks and dedicated datasets. Indeed most of the references mentioned above came with a novel data collection.

The VizWiz dataset introduced by Gurari et al. (2018) consists of over 31K visual questions originating from blind people who took pictures using a mobile phone and recorded a spoken question about it, together with 10 crowd-sourced answers per visual question. This collection was also recently exploited for an international VQA challenge both on answerability evaluation and answer prediction (Massiceti et al., 2022).

The Env-QA dataset by Gao et al. (2021) was the first collection of egocentric videos covering several events, designed for the analysis of the whole trajectory of state changes. The events include interactions with the environment (e.g. move the pot, turn on the faucet), thus more skills beyond an understanding of the scene composition are needed to solve the task. It contains 23K egocentric videos with an average

length of 20 seconds, along with 85K questions querying object number, attributes and states as well as events number and their temporal order. The annotations make the dataset suitable for free-form open-ended questions. Fan (2019) introduced the EgoVQA dataset with questions related to actions (of the camera wearer and third persons), interactions and relative positions, counting, and colours. More precisely, it contains 600 question-answer pairs with visual content across 5K frames from 16 first-person videos, with each video clip lasting from 20 to 100 seconds. This dataset has been mainly used for multiple-choice QA with (five-way classification). Bärmann and Waibel (2022) defined the QAEGO4D dataset that contains 1325 egocentric videos, each of 8 min on average, and 4837 unique answers. The authors provided both target moment annotations as well as answer confidence estimations, which can both serve as an additional source of (weak) supervision. The task on this data is cast as open-ended generative QA.

Datta et al. (2022) introduced the Episodic Memory EMQA dataset built by exploiting a 3D simulator to create egocentric RGB-D maps covering indoor paths. It contains 9.7K spatial and spatio-temporal localization questions about 12 object categories and the ground truth is provided as a binary segmentation map ("answer" vs "background" pixels). The EgoTaskQA benchmark proposed by Jia et al. (2022a) contains 40K questions balanced with a 1:2 ratio of binary and open-answer. They were procedurally generated within four types of questions (descriptive, predictive, explanatory and counterfactual) to systematically test models' capabilities over spatial, temporal and causal domains of goal-oriented task understanding. The corresponding videos are reasonably long with an average of five actions per clip to cover sufficient information for action dependency inference and future prediction. The QA task is formulated as a classification problem over the whole answer vocabulary. The AQTC benchmark proposed by Wong et al. (2022) was created with a close focus on task completion and affordances. It contains 100 instructional videos with an average duration of 115 seconds and involves 25 common household appliances, with 531 multiple-choice question-answer samples. The task associated with the dataset is particularly challenging as most of the answers require a sequence of more than two multi-modal steps to guide the user in operating the observed device.

For the future Egocentric VQA is a key enabler for a wide range of assistive applications in daily life and in working environments, but the state-of-the-art is in the early proof-of-concept phase and several challenges still need to be tackled. Starting from the data, all the existing egocentric VQA testbeds focus on indoor scenes which limits the model's applicability. Moving to outdoor environments implies managing a shift in the video features and in the questions' semantics. Multi-modality is one crucial aspect of the

task, but questions and answers are interpreted as text while speech, and more in general sound, are important cues that are currently less investigated.

Regarding the emergence of powerful vision-language models, their potential for application in egocentric VQA has only begun to be explored.

Recently LLaVA (Liu et al., 2023b) extended VQA to in-the-wild conditions where the answers require extensive knowledge coverage and multilingual understanding capabilities. The obtained results showed the limitations of existing models in grasping complex semantics. A relevant aspect to consider is also the length of the required video to answer the question: Mangalam et al. (2023) introduced EgoSchema, a very long-form video question-answering dataset that can serve as a valuable probe to assess the understanding capabilities of modern vision and language systems. When benchmarking several methods, the authors showed that even models with several billions of parameters achieve QA accuracy of less than 33% on the EgoSchema multichoice question answering task, while humans achieve about 76% accuracy. Jia et al. (2022a) indicated that for the most challenging reasoning questions, the performance of large pretrained vision-language models may show a drop and that the development of tailored prompting strategies for those cases is an interesting problem to be solved. In this work the best accuracy in predicting the correct answer in the open-ended setting is 30%, while humans get 82%.

4.11.2 Ego-Language Models

To engage in a dialogue with EgoAI, it should be endowed with language abilities that go beyond that of answering visual questions. A general conversation may include descriptions, explanations, and instructions as well as narratives, summaries, and comparisons. The most recent research products in this context are Large Language Models (LLMs), trained with a huge amount of textual data and capable of chatting with a user (Touvron et al., 2023; Brown et al., 2020). The computer vision research community is now focusing on the development of large multi-modal models that integrate vision and language by building on LLMs. In particular, this topic has gained momentum in the egocentric literature. We discuss these advances in this section

Seminal works An essential skill to unlock linguistic interactions based on visual information is that of translating semantic content seamlessly between the two modalities: from video to text and from text to video. One approach to achieving that is to learning a representation space shared by the two modalities. Once trained, these embeddings may be fine-tuned for a range of downstream tasks. Motivated by this, Lin et al. (2022) explored approaches for Video-Language Pretraning (VLP) and proposed a novel video-text contrastive objective, EgoNCE for egocentric videos. It adjusts the

InfoNCE objective (Oord et al., 2018) by performing action-aware positive sampling and scene-aware negative sampling. This makes NCE specific to long egocentric videos with multiple actions.

In Surís et al. (2020), the task of relating novel words, outside the learnt vocabulary, to visual objects is explored. The authors created an episode of examples consisting of video-text pairs and the model is tasked with completing the masked word from the target example using the reference set from the episode. However, the model is only allowed to fill in the masked word by copying and pasting words from within the episode. This is how the framework learns a policy for word acquisition.

Other techniques have been developed to leverage LLMs without creating new embeddings. Lavila (Zhao et al., 2023c) focuses on automatic video narration and utilises two LLMs: a *narrator* and a *rephraser*. The *narrator* is a visually conditioned auto-regressive language model that provides pseudo labels for existing and new clips with narrations. The rephraser, on the other hand, paraphrases the output of the narrator by changing word order or replacing common nouns and verbs. The results are reliable and diverse captions, providing temporally synced dense coverage for long videos.

State-of-the-art papers For Video-Language Pre-training, Pramanick et al. (2023) proposed EgoVLPv2 that incorporates cross-modal fusion directly into the video and language backbones. The network design keeps the encoders of the two modalities separated but the cross-modal attention modules combine their information and can be reused for downstream tasks with an advantage both in performance and efficiency. In particular, the EgoVLPv2 pre-trained encoders can be leveraged both for fast retrieval and grounding tasks, which require dual and fusion encoders, respectively.

Datasets Two of the largest egocentric videos datasets, EPIC-KITCHENS (Damen et al., 2022) and Ego4D (Grauman et al., 2022) provide dense free-form text descriptions, also known as narrations. They were collected through a stop-and-narrate approach where the subject watches their video and notes about what is happening there. For EPIC-KITCHENS, the authors record audio narrations from the camera wearers themselves, in their native language. They then transcribe and translate these timestamped narrations. For Ego4D, the authors hire annotators who watch the videos and write free-form descriptions of what is happening roughly every four seconds. For each video, two descriptions are collected from different annotators. Lin et al. (2022) utilises Ego4D to curate the EgoCLIP dataset for pre-training models on video-text pairs from egocentric videos. EgoCLIP consists of 3.8M clean egocentric clip-text pairs. For selecting the clips, videos with missing narrations from Ego4D are filtered, validation and test videos are excluded, and narrations from both runs in Ego4D are used to maintain the diversity. As EgoCLIP

is derived from Ego4D, the dataset contains diverse human activities. Lin et al. (2022) further propose the EgoMCQ benchmark that aims to evaluate video-text alignment and consists of 39K questions.

For the future Due to the inherent complexity posed by egocentric videos like head motion, occlusion, and limited field of view, standard VLP methods based on CLIP (Radford et al., 2021) fall short in generalising well. So future works should improve the adaptation of existing VLMs to egocentric data and might also introduce new tailored tasks to pave the way towards EgoAI sustaining dialogues with the user.

For instance, Wang et al. (2023e) target the development of an interactive AI assistant that can perceive, reason, and collaborate with humans in the real world. The proposed HoloAssist dataset consists of 166 h of data captured by 222 participants. While capturing the data, there is a *performer* and an *instructor*. The performer works on the task while wearing the AR headset and the instructor watches the performer in real-time and verbally guides the performer. This data collection procedure allows to ground the mistakes and correct the action towards task completion.

4.12 Privacy

Since the appearance of the first mass consumer wearable cameras in the late 2000s, the research community has been aware of the increased privacy risks related to their use. Such risks are primarily due to the intrinsic mobility of wearable cameras, which allows users to operate them in an "always on" mode, thus potentially capturing, transferring and processing sensitive information about themselves and bystanders. Hence, addressing privacy issues in egocentric vision brings specific challenges, as compared to fixed cameras.

While the community has a general understanding of the aforementioned risks, privacy in egocentric vision has not been systematically investigated, which is probably due to the fact that wearable devices equipped with cameras are not yet mainstream technology. Rather, a range of seminal and exploratory works have been proposed in the last decade. In this section, we provide a comprehensive discussion of the most relevant investigations on the topic. In particular, previous research has explored privacy considerations related to wearable cameras through three distinct perspectives: studies aimed to assess the degree to which the use of wearable cameras can affect *individuals' privacy* (Sect. 4.12.1), endeavours to *redact sensitive content* captured by wearable devices (Sect. 4.12.2), and advancements in *privacy-preserving* computer vision techniques (Sect. 4.12.3).

4.12.1 Users' Studies on Individual Privacy

A line of works has studied how people perceive their privacy and that of bystanders while operating wearable cameras. In particular, Hoyle et al. (2014) performed a user study of 36 people wearing a life-logging camera for 1 week, discovering that subjects prefer to be given the option to remove sensitive images in-situ, during the image collection. Moreover, factors such as time, location, objects and people appearing in a photo determine its sensitivity, and camera wearers are generally concerned about the privacy of bystanders. Similar findings are reported in the work of Price et al. (2017), which studied the perception of privacy by different groups of life-loggers. Complementarily, Denning et al. (2014) investigated the reactions of bystanders to users wearing AR devices. The study highlights that AR devices change the bystander experience due to subtlety and ease of recording by the camera wearer, with recordings considered more or less acceptable depending on when and where they are being taken. Hoyle et al. (2015) analyses photos taken during life-logging sessions, asking camera wearers to provide motivations on whether a given image should be shared or not: impression management and respect for others' privacy were the main reasons for keeping images private.

Other works have studied the privacy implications of systems exploiting egocentric images and videos. Roesner et al. (2014) analysed the security and privacy risks of AR applications in mobile and wearable devices. The study emphasised the necessity of dedicated protocols for wearable applications to ensure the safe usage of sensitive user information such as location data from visual signals or an accurate 3D model of an indoor environment (Templeman et al., 2012). In turn, some works have investigated how egocentric video can contain subtle information which may disclose the identity of the camera wearer, as discussed in Sect. 4.9.

4.12.2 Redacting Sensitive Information

Based on the findings of previous studies (Hoyle et al., 2014; Roesner et al., 2014; Hoyle et al., 2015), some works suggested to prevent sharing of egocentric images based on the presence of specific objects or persons. For instance, Templeman et al. (2014) proposed PlaceAvoider, a system able to recognise whether a given egocentric image has been acquired in a sensitive location specified by the user, such as a bathroom or one's bedroom, in order to prevent sharing of such images. Korayem et al. (2016) proposed to prevent sharing egocentric images based on the presence of detected screens (e.g., the ones of smartphones or laptops) which are likely to include personal information such as credit card numbers or addresses. Hasan et al. (2020) investigated approaches to detect bystanders in egocentric images based on cues such as intentionally posing for a photo. Images with

detected bystanders can then be submitted for review to the camera wearer before proceeding to share them.

Other works investigated how egocentric images and videos can be transformed to protect privacy, while still allowing for downstream tasks to be performed. In particular, Hassan et al. (2017) proposed a "cartoon transform" which alters the low-level properties of the image and replaces objects with aligned clip art. Dimiccoli et al. (2018) studied how intentionally degrading image quality by blurring can improve bystanders' perception of privacy, while still allowing to perform downstream tasks such as activity recognition. Finally, Thapar et al. (2021) proposed to add subtle perturbations to egocentric video that do not affect tasks like object detection or action/activity recognition but are strong enough to prevent the identification of the camera wearer from head motion analysis.

4.12.3 Privacy Preserving by Design

This line of work investigates systems and algorithms to tackle egocentric computer vision tasks while guaranteeing that privacy-sensitive content is correctly processed, by avoiding to collect it, store it, or make it available to untrusted applications. Jana et al. (2013b) introduced the idea of "recognisers" as a software layer to provide AR applications with only the necessary high-level, anonymous information, rather than giving direct access to raw sensor data such as RGB images, which may contain private content. A similar concept is explored in Jana et al. (2013a), where a privacy protection layer based on OpenCV is used to mediate sensor input (e.g., applying sketching transform) before making it available to possibly untrusted AR applications. Ryoo et al. (2016) proposed a privacy-conscious approach which learns how to extremely subsample egocentric video resolution to preserve privacy while still allowing to perform the downstream task of activity recognition. Following Pittaluga et al. (2019), who showed how scenes can be revealed inverting structure from motion reconstructions, a line of research has investigated approaches for 6-DoF localisation that safeguards against reconstructing the original image from the information stored to support localisation (Speciale et al., 2019; Pietrantoni et al., 2023; Chelani et al., 2023, 2021; Dusmanu et al., 2021; Ng et al., 2022). Steil et al. (2019) presented a system which shuts off the video stream when sensitive visual content is detected, only to reactivate it based on the analysis of eye movements recorded by additional eye tracking cameras. Along the same lines, Khan et al. (2021) proposed a deep learning based device which detects user-customised privacy-sensitive content such as objects and faces of specific people in order to serve as a privacy filter, blocking images which do not satisfy the established privacy constraints. Qiu et al. (2023) investigated how converting images into rich text descriptions can serve as an effec-

tive privacy-preserving approach for passive dietary intake monitoring from egocentric images, as compared to directly storing the input images.

Datasets A significant portion of the studies on privacy involves the collection of in-situ data and the administration of surveys to participants (Denning et al., 2014; Hoyle et al., 2014; Price et al., 2017; Korayem et al., 2016). These works primarily focus on examining participants' attitudes towards lifelogging through the utilisation of AR glasses. In addition to analysing the gathered survey data, Hoyle et al. (2015) also examines the images generated during the data collection process. Another group of studies employs custom hardware setups to capture images in an in-situ context (Thomaz et al., 2013; Templeman et al., 2014; Speciale et al., 2019).

The First Person Social Interaction Dataset (FPSI) by Fathi et al. (2012a), described in Sect. 5, has been used for privacy analysis. Together with other data collections originally created for person identification (IITMD already covered by Sect. 4.9), it has been adopted to investigate potential biometric signature leakage in egocentric videos.

For the future The discussed seminal works exemplify the efforts of the community in pursuing privacy-aware technologies. However, the investigations are not yet systematic mainly due to the limited adoption of wearable cameras by the general public.

Previous works mainly surveyed individuals but it is important to identify issues with existing datasets, as the covered demographic and number of subjects (< 50) are generally limited. As devices are rapidly evolving, it will be crucial to evaluate the effect on privacy when moving from mobile phones and GoPros to wearable devices. This appears particularly important considering that the majority of the studies in the current literature used custom devices instead of commercially available ones (Thomaz et al., 2013; Templeman et al., 2014; Yonetani et al., 2015; Speciale et al., 2019).

Finally, several works focused on analysing recorded data to assess privacy breaches, but only a few proposed solutions for privacy preservation or privacy-aware processing and more tailored strategies are needed.

4.13 Beyond Individual Tasks

The previous sections revised 12 distinct computer vision research tasks, which are fundamental for the future of egocentric vision. Few efforts have attempted to combine these tasks to get close to the abilities of our futuristic EgoAI. For example, an approach that combines action recognition with a dynamic memory, so as to remind Sam about the bread in the toaster is beyond the reach of current methods. Similarly, combining person Re-ID with trajectory forecasting towards assisting Judy in locating a suspect along her path

has also not been explored before. Assessing performance in daily tasks, whether to advise Sam about the amount of spice in his soup or qualitatively and quantitatively assess Marco's daily performance are still futuristic tasks not yet explored in egocentric vision. The understanding of one's surroundings, actions and intentions, both independently as well as jointly with others, is key to the integration of EgoAI in our daily lives. We hope more works will expand beyond the individual research tasks to get an effective assistive device for the wearer.

We next review general datasets, suitable for multiple tasks, which can pave the way to such holistic understanding in egocentric vision.

5 General Datasets

Datasets have become the fuel of computer vision research. They offer the starting point for studying new research problems and developing artificial intelligence that can successfully support humans. The more realistic a dataset is, the higher its value in transforming our future and often the higher its challenge. Such datasets usually require increased research efforts to achieve good performance on their various metrics.

In particular, the availability of these datasets has crucially contributed to advancements in egocentric vision research. In fact, as wearable cameras are still relatively new, the videos available online are not taken from the egocentric perspective. In the previous sections, we reported datasets that were designed for one task. In this section, we review general datasets that are suitable for a variety of tasks, and present their characteristics. We compare in Table 1 the most popular publicly available egocentric datasets in terms of domains, size and modalities. We then detail available annotations to date in Table 2 and in Table 3 we relate these datasets to the tasks reviewed in Sect. 4. Next, we provide a narrative for these general datasets.

Activity of Daily Living (ADL) by Pirsiavash and Ramanan (2012) was one of the first egocentric datasets. It consists of one million frames captured in people's homes. The dataset is only scripted at a high-level by asking the camera wearers to carry out specific tasks such as watching TV or doing laundry. It is annotated with object tracks, hand positions, and interaction events. ADL has found its use in various tasks like action (Vondrick et al., 2016) and region anticipation (Furnari et al., 2017), action recognition (Pirsiavash & Ramanan, 2012) and video summarisation (Lu & Grauman, 2013). Similarly, the UTE dataset (Lee et al., 2012) is composed by videos of 4 participants involved in various activities such as eating, shopping, attending a lecture, driving and cooking. One notable difference with respect to ADL is in the video length: the average duration of a video in UTE is 3.7 h (222

Table 1 General egocentric datasets—collection characteristics

Dataset	Settings	Signals	Hours	Sequences	AVG. video duration (min)	Participants
MECCANO (Ragusa et al., 2023b)	Industrial	RGB, depth, gaze	6.9	20	20.79	20
ADL (Pirsiavash & Ramanan, 2012)	Daily activities	RGB	10.0	20	30.00	20
HOI4D (Liu et al., 2022c)	Table-top	RGB, depth	22.2	4000	0.33	9
EGTEA Gaze+† (Li et al., 2021a)	Kitchen	RGB, gaze	27.9	86	19.53	32
UTE (Lee et al., 2012)	Daily activities	RGB	37.0	10	222.00	4
EGO-CH (Ragusa et al., 2020a)	Cultural sites	RGB	37.1	180	12.37	70
FPSI (Fathi et al., 2012a)	Recreational site	RGB	42.0	8	315.00	8
KrishnaCam (Singh et al., 2016a)	Daily routine	RGB, GPS, acc	69.9	460	9.13	1
EPIC-KITCHENS-100 (Damen et al., 2022)	Kitchens	RGB, audio	100.0	700	8.57	37
Assembly101 (Sener et al., 2022)	Industrial	RGB, multi-view	167.0	1425	7.10	53
Ego4D⋆ (Grauman et al., 2022)	Multi domain	RGB, Audio, 3D, gaze, IMU, multi	3670.0	9650	24.11	931

†: For EGTEA, Audio was collected but not made public. ⋆: For Ego4D, apart from RGB, the other modalities are present for subsets of the data

Table 2 General egocentric datasets—current set of annotations

Dataset	Annotations
MECCANO (Ragusa et al., 2023b)	Temporal action segments, hand & object bounding boxes, hand-object interactions, next-active object
ADL (Pirsiavash & Ramanan, 2012)	Temporal action segments, objects bounding boxes, hand-object interactions
HOI4D (Liu et al., 2022c)	Temporal action segments, 3D hand poses and object poses, panoptic and motion segmentation, object meshes, scene point clouds
EGTEA Gaze+ (Li et al., 2021a)	Temporal action segments, hand masks, gaze
UTE (Lee et al., 2012)	Text descriptions, object segmentations
EGO-CH (Ragusa et al., 2020a)	Temporal locations, object bounding boxes, surveys, object masks
FPSI (Fathi et al., 2012a)	Temporal social interaction segments
KrishnaCam (Singh et al., 2016a)	Motion classes, virtual webcams, popular locations
EPIC-KITCHENS-100 (Damen et al., 2022)	Temporal action video segments, Temporal audio segments, narrations, hand and objects masks, hand-object interactions, camera poses
Assembly101 (Sener et al., 2022)	Temporal action segments, 3D hand poses
Ego4D⋆ (Grauman et al., 2022)	Narrations, Temporal action segments, moment queries, speaker labels, diarisation, hand bounding boxes, time to contact, active objects bounding boxes, trajectories, next-active objects bounding boxes

⋆: For Ego4D, apart from narrations, the remaining annotations are only available for subsets of the dataset depending on the benchmark

mins) compared to 30 mins for ADL. Regarding the annotations, the UTE dataset provides a paragraph summary of the videos and polygon annotations around the subjects based on the summary, which makes it suitable for studying video summarisation (Lee et al., 2012; Lu & Grauman, 2013). Further expanding the covered time range, the KrishnaCam dataset by Singh et al. (2016a) includes nine months of one student's daily activities. It consists of 7.6 million frames, spanning 70 h of video, accompanied by GPS position, acceleration, and body orientation data. Thanks to its time evolution capture of nine months, KrishnaCam can be used to study tasks such as trajectory prediction, detecting popular places and scene changes. The dataset has been also used to address online

object detection (Wang et al., 2021b) and for self-supervised representation learning (Purushwalkam et al., 2022).

Different in terms of domains and captured signals, the GTEA Gaze dataset (Fathi et al., 2012b) and its extension EGTEA Gaze+ (Li et al., 2021a) cover recipe preparation with the gaze signal in a single kitchen. The GTEA Gaze dataset by Fathi et al. (2012b) focuses on action recognition and gaze prediction and involves the use of eye-tracking glasses equipped with an infrared inward-facing gaze sensing camera to track the 2D location of the subjects' eye gaze during meal preparation activities. The dataset includes 17 sequences performed by 14 subjects making pre-specified meal recipes. It has been annotated with 25 frequently occur-

Table 3 General egocentric datasets—current set of tasks: **4.1** Localisation, **4.2** 3D Scene Understanding, **4.3** Recognition, **4.4** Anticipation, **4.5** Gaze Understanding and Prediction, **4.6** Social Behaviour Understanding, **4.7** Full-body Pose Estimation, **4.8** Hand and Hand-Object Interactions, **4.9** Person Identification, **4.10** Summarisation, **4.11** Dialogue, **4.12** Privacy

Dataset	Task											
	4.1	4.2	4.3	4.4	4.5	4.6	4.7	4.8	4.9	4.10	4.11	4.12
MECCANO (Ragusa et al., 2023b)			✓	✓	✓			✓				
ADL (Pirsiavash & Ramanan, 2012)			✓	✓						✓		
HOI4D (Liu et al., 2022c)								✓				
EGTEA Gaze+ (Li et al., 2021a)			✓	✓	✓			✓				
UTE (Lee et al., 2012)								✓		✓		
EGO-CH (Ragusa et al., 2020a)	✓											
FPSI (Fathi et al., 2012a)						✓				✓		✓
KrishnaCam (Singh et al., 2016a)				✓								
EPIC-KITCHENS-100 (Damen et al., 2022)		✓	✓	✓				✓			✓	✓
Assembly101 (Sener et al., 2022)			✓					✓				
Ego4D (Grauman et al., 2022)			✓	✓	✓	✓		✓		✓	✓	

ring actions, such as "take", "pour", and "spread" indicating their starting and ending frames. This dataset was later extended as EGTEA Gaze+ by Li et al. (2021a) with 28 h of cooking activities, including video, gaze tracking data, and action annotations of 106 actions, along with pixel-level hand masks. The dataset has been used to address different tasks such as anticipation (Furnari & Farinella, 2019; Girdhar & Grauman, 2021; Zhong et al., 2023), action recognition (Kazakos et al., 2021; Fathi et al., 2012b), procedural learning (Bansal et al., 2022), and future hand masks prediction (Jia et al., 2022b).

A few datasets targeted multi-person egocentric social interactions. The First Person Social Interaction (FPSI) dataset by Fathi et al. (2012a) was collected over 3 days, by a group of 8 individuals that visited Disney theme parks, recording over 42 h of multi-person videos using head-mounted cameras. The group often splits into smaller sub-groups during the day, resulting in unique experiences in each video. The dataset consists of over two million images, manually labelled for six types of social interactions: dialogue, discussion, monologue, walk dialogue, walk discussion, and background activities. The dataset has proven useful for video summarisation (Nagar et al., 2021; Rathore et al., 2019; Poleg et al., 2015b) and privacy preservation (Fathi et al., 2012a; Thapar et al., 2020a).

Ragusa et al. (2020a) proposed the EGO-CH dataset to study visits to cultural heritage sites. It includes 27 h of video recorded from 70 subjects. Annotations are provided for 26 environments and over 200 Points of Interest (POIs), featuring temporal labels indicating the environment in which the visitor is located and the currently observed PoI with bounding box annotations. It has been used by the authors to tackle room-based localisation, PoI recognition, image retrieval and survey generation—i.e. predicting the responses in the survey from the egocentric video. Furthermore, the dataset has been used to address object detection (Pasqualino et al., 2022b, a), image-based localisation (Orlando et al., 2020) and semantic object segmentation (Ragusa et al., 2020b).

While these datasets explore various aspects of egocentric vision, their small scale and focus on a single environment or a handful of individuals poses challenges when training deep learning models or attempting to generalise to other locations or subjects. To this end, the EPIC-KITCHENS dataset by Damen et al. (2018) was proposed as a significantly larger egocentric video dataset introduced in 2018 and subsequently extended with the latest version EPIC-KITCHENS-100 by Damen et al. (2022). The dataset comprises 100 h of unscripted video recordings captured by 37 participants from 4 countries in their own kitchens. It is unique in its instructions to participants, so as to start recording before entering the kitchen and only to pause when stepping out. This offered the first unscripted nature where participants go around their environments unhindered forming their own goals. The dataset has been annotated temporally with action segments. It consists of 90K action segments, 20K unique narrations, 97 verb classes, and 300 noun classes. It has since been extended with three additional annotations. First, EPIC-KITCHENS Video Object Segmentations and Relations (VISOR) (Darkhalil et al., 2022) provided pixel-level annotations focusing on hands, objects and hand-object interaction labels. VISOR offers 272K manual semantic masks of 257 object classes, 9.9M interpolated dense masks, 67K hand-object relations. Second, EPIC-SOUNDS (Huh et al., 2023) annotates the temporally distinguishable audio segments, purely from the audio stream of videos in EPIC-KITCHENS. It includes 78.4k categorised segments of audible events and actions,

distributed across 44 classes as well as 39.2k non-categorised segments. Third, EPIC Fields (Tschernezki et al., 2023) successfully registered and provided camera poses for 99 out of the 100 h of EPIC-KITCHENS. This is achieved through a proposed pipeline of frame filtering so as to attend to transitions between hotspots. The camera poses offer the chance for combining all aforementioned annotations with 3D understanding and are likely to unlock new potential on this dataset.

Since its introduction, EPIC-KITCHENS has become the de facto dataset for egocentric action recognition (Kazakos et al., 2019; Xiong et al., 2022; Yan et al., 2022; Girdhar et al., 2022), privacy (Thapar et al., 2020b), and anticipation (Furnari & Farinella, 2019; Girdhar & Grauman, 2021; Gu et al., 2021; Roy & Fernando, 2022; Liu et al., 2020b; Jia et al., 2022b; Pasca et al., 2023; Zhong et al., 2023). New tasks have also been defined around EPIC-KITCHENS, particularly related to domain adaptations with its capture in multiple locations and over time (Munro & Damen, 2020; Kim et al., 2021; Sahoo et al., 2021), video retrieval (Zhao et al., 2023c; Lin et al., 2022), manipulations (Shaw et al., 2022) as well as niche topics like object-level reasoning (Baradel et al., 2018) and learning words in other languages from visual representations (Surís et al., 2020).

A couple of datasets focus on industry-like setting. MEC-CANO by Ragusa et al. (2021, 2023b) is an egocentric procedural dataset capturing subjects building a toy motorbike model. The dataset includes synchronised gaze, depth and RGB. Its 20 object classes cover components, tools, instructions booklet. It has been used to address tasks like action recognition (Deng et al., 2023), active object detection (Fu et al., 2022b), hand-object interactions (Tango et al., 2022) and procedural learning (Bansal et al., 2022). Similarly, Assembly101 (Sener et al., 2022) is a procedural activity dataset with 4321 videos of individuals assembling and disassembling 101 "take-apart" toy vehicles. The dataset showcases diverse variations in action orders, mistakes, and corrections. It contains over 100K coarse and 1M fine-grained action segments, along with 18M 3D hand poses. This dataset has found its use in action recognition (Wen et al., 2023b), anticipation (Zatsarynna & Gall, 2023) and hand pose estimation (Zheng et al., 2023b; Ohkawa et al., 2023). Additionally, HOI4D dataset by Liu et al. (2022c) consists of 2.4M RGB-D video frames and 4000 sequences, featuring 9 participants interacting with 800 object instances from 16 categories. The dataset provides annotations for panoptic and motion segmentation, 3D hand pose, category-level object pose and includes reconstructed object meshes and scene point clouds. This dataset has proven useful for object segmentation and shape reconstruction (Liu et al., 2023c; Zhang et al., 2023d; Wen et al., 2022), action segmentation (Reza et al., 2023; Zhang et al., 2023d), hand-object manipulation synthesis (Zheng et al., 2023a; Ye et al., 2023b), hand action

detection (Hung-Cuong et al., 2023) and 3D hand pose estimation (Ye et al., 2023b).

The most impressive and massive-scale dataset to date is Ego4D by Grauman et al. (2022), with 3670 h of daily-life activity videos spanning hundreds of unscripted scenarios (household, outdoor, workplace, leisure, etc.) captured by 931 unique camera wearers from 74 worldwide locations and 9 countries. It primarily comprises videos, with subsets of the dataset containing audio, eye gaze, and 3D meshes of the environment. The dataset was released with a set of benchmarks and train/val/test split annotations that focus on the past (querying an episodic memory), the present (hand-object manipulation, audio-visual conversation, and social interactions), as well as the future (forecasting activities and trajectories).

Due to the massive-scale and unconstrained nature of Ego4D, it has proved to be useful for various tasks including action recognition (Liu et al., 2022a; Lange et al., 2023), action detection (Wang et al., 2023a), visual question answering (Bärmann & Waibel, 2022), active speaker detection (Wang et al., 2023d), natural language localisation (Liu et al., 2023a), natural language queries (Ramakrishnan et al., 2023), gaze estimation (Lai et al., 2022), persuasion modelling for conversational agents (Lai et al., 2023b), audio visual object localisation (Huang et al., 2023a), hand-object segmentation (Zhang et al., 2022b) and action anticipation (Ragusa et al., 2023a; Pasca et al., 2023; Mascaró et al., 2023). New tasks have also been introduced thanks to the diversity of Ego4D, e.g. modality binding (Girdhar et al., 2023), part-based segmentation (Ramanathan et al., 2023a), long-term object tracking (Tang et al., 2023a), relational queries (Yang et al., 2023) and action generalisation over scenarios (Plizzari et al., 2023). Additionally, due to its unprecedented scale, it has broken grounds in training robot models with a series of publications (Nair et al., 2022; Radosavovic et al., 2022; Ma et al., 2023b), transforming the field of learning from demonstrations. The potential for the Ego4D dataset is yet to be fully explored.

As noted at the start of this section, egocentric datasets have a key role in research advancement. By reviewing our initial forecast of the future in Sect. 2, some scenarios received more attention than others in producing large-scale datasets. EGO-Home (Sect. 2.1) is partially overlapping with datasets such as EPIC-KITCHENS-100, Ego4D and EGTEA Gaze+. However, these datasets mostly focus on the home activities of cooking, cleaning and playing games. They do not cover parts related to down time (i.e. rest), or grooming or personal health, mostly due to privacy concerns. EGO-Worker (Sect. 2.2) is related to datasets such as MECCANO, Assembly-101 and HOI4D. However, these do not cover the holistic aspects of a worker's daily activities and are yet to explore the critical aspects of safety and feedback. EGO-Tourist (Sect. 2.3) is related to the EGO-CH dataset

of visitors in heritage sites. However, the scale remains very small despite the presence of large-scale touring videos on YouTube that could be utilise for city-wide touring.

EGO-Police (Sect. 2.4) does not correspond to any publicly available datasets. Despite the wide usage of chest-mounted cameras within the police forces worldwide, such data is very sensitive, particularly across boarders. Relevant datasets currently are far from being utilised for advancing research in egocentric vision. Finally, we call for more datasets in egocentric understanding from the entertainment industry, given the huge potential of transforming this domain as noted by the Ego-Designer scenario we presented (Sect. 2.5).

6 Conclusion

This paper aimed to provide a future-to-present perspective into egocentric vision. Looking ahead, we envisage a wearable device that we call EgoAI, holding the potential to redefine our daily lives. We showcased its seamless integration into our everyday existence through character-based futuristic scenarios, indoors and outdoors, at work, at home and even during holidays.

We demonstrated the need for this device to be *multi-sensored* and *multi-tasked*. While our focus is on cameras and visual cues, the future clearly requires it to be multimodal in its capabilities, whether for perceiving the surroundings and understanding what is happening in the observed scene, or interacting with the camera wearer. At the same time, without the ability to solve multiple fundamental vision tasks it will not be possible to get a competent egocentric assistant.

Additionally, we believe that further developing generative tasks in egocentric vision will play a pivotal role towards building EgoAI. Consider, for instance, scenarios where Marco could benefit from a device guiding him through his work procedure by illustrating the sequential steps within his environment. Similarly, EgoAI could spark Stanley's creativity by proposing diverse scenographies projected onto his current surroundings. Current egocentric methods employ generative approaches in limited contexts, ranging from predicting future head motion (Jia et al., 2022b) to anticipating gaze (Zhang et al., 2017) and modelling hand-object interactions (Ye et al., 2023b). Only a handful of works explore cross-view third-to-first-person image (Liu et al., 2020a) and video (Liu et al., 2021a) synthesis. One recent work that closely aligns with our use cases is from Yang et al. (2024). They introduced a universal video generator that predicts future frames based on both low- and high-level textual action prompts. When run sequentially it can also effectively simulates long-horizon interactions. This quality makes it well-suited for generating a visual representation of work procedures tailored to Marco's needs.

In this paper, we reviewed 12 research tasks in egocentric vision: *localisation*, *3D scene understanding*, *recognition*, *anticipation*, *gaze understanding and prediction*, *social behaviour understanding*, *full-body pose estimation*, *hand and hand-object interactions*, *person identification*, *summarisation*, *dialogue* and *privacy*. For each task, we presented an overview considering appreciated seminal works that set the research path, and we provided an insight into the current state-of-the-art methods, publicly available datasets and directions for future innovations. While the literature builds on previous research based on fixed cameras, each of these tasks present challenges which are unique to egocentric vision, and in particular to the mobile nature of wearable cameras and the need for a user-specific understanding of the scene. On the other hand, egocentric vision brings new opportunities for human-centric applications as discussed at large in this paper. We anticipate that future research will focus on bridging the gap between egocentric approaches and those based on third-person vision in the spirit of convergence towards a unified technology. Towards achieving this goal, the newly introduced Ego-Exo4D dataset recorded using both egocentric and up to 4 exocentric cameras has recently been introduced (Grauman et al., 2023).

We highlight that these tasks cannot exist independently—i.e. it is infeasible that we will be learning one deep model per task. This is not only because a model-per-task is extremely inefficient, but because these tasks are co-dependent and the prediction of one task would inform plausible predictions of another. We encourage researchers to explore the taxonomy of research tasks in egocentric vision. Moreover, future works should also consider open set settings so that each task is able to manage novelty to avoid relying too much on pre-defined label sets and enhance model trustworthiness.

Regarding efficiency, this survey stopped short of exploring the need for real-time sensing and learning, although it is evident that we need to build models capable of performing all the mentioned tasks in real-time or with very minimal latency. Ideally, future egocentric devices should be always connected online, while respecting all privacy and protection concerns. We encourage other researchers to analyse these aspects as without efficient sensing, efficient computing and real-time interactions, the future would remain a dream in fiction novels and sci-fi films. At the same time, without privacy-aware models, sensors and systems, the future would fail to deliver on its users expectations.

We hope this paper offers a stepping stone for researchers to make the future of egocentric vision a reality. We are seeking input from researchers in the field to strengthen and complete this survey, so it can serve as a useful reference to incoming researchers who wish to explore and contribute to egocentric vision.

References

Abavisani, M., Joze, H.R.V., & Patel, V.M. (2019). Improving the performance of unimodal dynamic hand-gesture recognition with multimodal training. In *CVPR*.

Abu Farha, Y., Richard, A., & Gall, J. (2018). When will you do what? Anticipating temporal occurrences of activities. In *CVPR*.

Aghaei, M., Dimiccoli, M., Ferrer, C.C., & Radeva, P. (2017). Social style characterization from egocentric photo-streams. In *ICCV workshop*.

Ahmetovic, D., Gleason, C., Ruan, C., Kitani, K., Takagi, H., & Asakawa, C. (2016). NavCog: A navigational cognitive assistant for the blind. In *International conference on human–computer interaction with mobile devices and services*

Aizawa, K., Ishijima, K., & Shiina, M. (2001). Summarizing wearable video. In *ICIP*.

Akada, H., Wang, J., Shimada, S., Takahashi, M., Theobalt, C., & Golyanik, V. (2022). UnrealEgo: A new dataset for robust egocentric 3D human motion capture. In *ECCV*.

Akiva, P., Huang, J., Liang, K.J., Kovvuri, R., Chen, X., Feiszli, M., Dana, K., & Hassner, T. (2023). Self-supervised object detection from egocentric videos. In *ICCV*.

Alcantarilla, P.F., Yebes, J.J., Almazán, J., & Bergasa, L.M. (2012). On combining visual SLAM and dense scene flow to increase the robustness of localization and mapping in dynamic environments. In *ICRA*.

Ali-bey, A., Chaib-draa, B., & Giguère, P. (2023). MixVPR: Feature mixing for visual place recognition. In *WACV*.

Alikadic, A., Saito, H., & Hachiuma, R. (2022). Transformer networks for future person localization in first-person videos. In *International symposium on visual computing*.

Alletto, S., Serra, G., & Cucchiara, R. (2015b). Egocentric object tracking: An odometry-based solution. In *ICIAP*.

Alletto, S., Serra, G., Calderara, S., & Cucchiara, R. (2015). Understanding social relationships in egocentric vision. *Pattern Recognition, 48*(12), 4082–4096.

Al-Naser, M., Siddiqui, S.A., Ohashi, H., Ahmed, S., Katsuyki, N., Takuto, S., & Dengel, A. (2019). OGaze: Gaze prediction in egocentric videos for attentional object selection. In *International conference on digital image computing: Techniques and applications*.

Anderson, P., Wu, Q., Teney, D., Bruce, J., Johnson, M., Sunderhauf, N., Reid, I., Gould, S., & van den Hengel, A. (2018). Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *CVPR*.

Aoki, H., Schiele, B., & Pentland, A. (1998). Recognizing personal location from video. In *Workshop on perceptual user interfaces*.

Arandjelovic, R., Gronat, P., Torii, A., Pajdla, T., & Sivic, J. (2016). NetVLAD: CNN architecture for weakly supervised place recognition. In *CVPR*.

Ardeshir, S., & Borji, A. (2016). Ego2Top: matching viewers in egocentric and top-view videos. In *ECCV*.

Ardeshir, S., & Borji, A. (2018). Integrating egocentric videos in top-view surveillance videos: Joint identification and temporal alignment. In *ECCV*.

Arev, I., Park, H. S., Sheikh, Y., Hodgins, J., & Shamir, A. (2014). Automatic editing of footage from multiple social cameras. *Transactions on Graphics, 33*(4), 1–11.

Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., & Schmid, C. (2021). ViViT: A video vision transformer. In *ICCV*.

Badino, H., & Kanade, T. (2011). A head-wearable short-baseline stereo system for the simultaneous estimation of structure and motion. In *International conference on machine vision applications*.

Bai, Y., & Qi, Y. (2018). A robust method for hands gesture recognition from egocentric depth sensor. In *International conference on virtual reality and visualization*.

Furlan, V. S., Bajcsy, R. Nascimento, E. R. (2018). Fast forwarding egocentric videos by listening and watching. In *CVPR workshop*.

Bambach, S., Lee, S., Crandall, D.J., & Yu, C. (2015). Lending a hand: Detecting hands and recognizing activities in complex egocentric interactions. In *ICCV*.

Bandini, A., & Zariffa, J. (2023). Analysis of the hands in egocentric vision: A survey. *TPAMI, 45*(6), 6846–6866.

Bano, S., Suveges, T., Zhang, J., & Mckenna, S. J. (2018). Multimodal egocentric analysis of focused interactions. *Access, 6*, 37493–37505.

Bansal, S., Arora, C., & Jawahar, C. (2022). My view is the best view: Procedure learning from egocentric videos. In *ECCV*.

Bao, W., Chen, L., Zeng, L., Li, Z., Xu, Y., Yuan, J., & Kong, Y. (2023). Uncertainty-aware state space transformer for egocentric 3D hand trajectory forecasting. In *ICCV*.

Baradel, F., Neverova, N., Wolf, C., Mille, J., & Mori, G. (2018). Object level visual reasoning in videos. In *ECCV*.

Baraldi, L., Paci, F., Serra, G., Benini, L., & Cucchiara, R. (2014). Gesture recognition in ego-centric videos using dense trajectories and hand segmentation. In *CVPR Workshop*.

Bärmann, L., & Waibel, A. (2022). Where did I leave my keys? Episodic-memory-based question answering on egocentric videos. In *CVPR workshop*.

Basaran, E., Tesfaye, Y.T., & Shah, M. (2018). EgoReID dataset: Person Re-identification in videos acquired by mobile devices with first-person point-of-view. arXiv:1812.09570

Bertasius, G., & Shi, J. (2017). using cross-model egosupervision to learn cooperative basketball intention. In *ICCV workshop*.

Bertasius, G., Park, H.S., & Shi, J. (2015). Exploiting egocentric object prior for 3D saliency detection. arXiv:1511.02682

Bertasius, G., Soo Park, H., Yu, S.X., & Shi, J. (2017). Unsupervised learning of important objects from first-person videos. In *ICCV*.

Berton, G., Masone, C., & Caputo, B. (2022). Rethinking visual geo-localization for large-scale applications. In *CVPR*.

Betancourt, A., Morerio, P., Regazzoni, C. S., & Rauterberg, M. (2015). The evolution of first person vision methods: A survey. *Transactions on Circuits and Systems for Video Technology, 25*(5), 744–760.

Bettadapura, V., Castro, D., & Essa, I. (2016). Discovering picturesque highlights from egocentric vacation videos. In *WACV*.

Blanton, H., Greenwell, C., Workman, S., & Jacobs, N. (2020). Extending absolute pose regression to multiple scenes. In *CVPR workshop*.

Bock, M., Kuehne, H., Van Laerhoven, K., & Moeller, M. (2023). WEAR: An outdoor sports dataset for wearable and egocentric activity recognition. arXiv:2304.05088

Bolaños, M., & Radeva, P. (2015). Ego-object discovery. arXiv:1504.01639

Bolaños, M., Dimiccoli, M., & Radeva, P. (2016). Toward storytelling from visual lifelogging: An overview. *Transactions on Human-Machine Systems, 47*(1), 77–90.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., 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., 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. (2020). Language models are few-shot learners. In *NeurIPS*.

Cai, M., Kitani, K. M., & Sato, Y. (2016). Understanding hand-object manipulation with grasp types and object attributes. In *Robotics: science and systems*.

Cao, C., Zhang, Y., Wu, Y., Lu, H., & Cheng, J. (2017). Egocentric gesture recognition using recurrent 3D convolutional neural networks with spatiotemporal transformer modules. In *ICCV*.

Capi, G., Kitani, M., & Ueki, K. (2014). Guide robot intelligent navigation in urban environments. *Advanced Robotics, 28*(15), 1043–1053.

Castle, R. O., Klein, G., & Murray, D. W. (2010). Combining monoSLAM with object recognition for scene augmentation using a wearable camera. *Image and Vision Computing, 28*(11), 1548–1556.

Cazzato, D., Leo, M., Distante, C., & Voos, H. (2020). When I look into your eyes: A survey on computer vision contributions for human gaze estimation and tracking. *Sensors, 20*(13), 3739.

Chakraborty, A., Mandal, B., & Galoogahi, H.K. (2016). Person re-identification using multiple first-person-views on wearable devices. In *WACV*.

Chalasani, T., & Smolic, A. (2019). Simultaneous segmentation and recognition: Towards more accurate ego gesture recognition. In *ICCV workshop*.

Chalasani, T., Ondrej, J., & Smolic, A. (2018). Egocentric gesture recognition for head-mounted AR devices. In *International symposium on mixed and augmented reality adjunct*.

Chan, C.S., Chen, S.Z., Xie, P., Chang, C.C., & Sun, M. (2016). Recognition from hand cameras: A revisit with deep learning. In *ECCV*.

Chandio, Y., Bashir, N., Anwar, F.M. (2022). HoloSet—A dataset for visual-inertial pose estimation in extended reality: Dataset. In *Conference on embedded networked sensor systems*.

Chelani, K., Kahl, F., & Sattler, T. (2021). How privacy-preserving are line clouds? Recovering scene details from 3D lines. In *CVPR*.

Chelani, K., Sattler, T., Kahl, F., & Kukelova, Z. (2023). Privacy-preserving Representations are not Enough: Recovering scene content from camera poses. In *CVPR*.

Chen, C., Anjum, S., & Gurari, D. (2022). Grounding answers for visual questions asked by visually impaired people. In *CVPR*.

Chen, Z., Chen, S., Schmid, C., & Laptev, I. (2023). gSDF: Geometry-driven signed distance functions for 3D hand-object reconstruction. In *CVPR*.

Cheng, W., Park, J.H., & Ko, J.H. (2021). HandFoldingNet: A 3D hand pose estimation network using multiscale-feature guided folding of a 2D hand skeleton. In *ICCV*.

Cheng, J., Zhang, L., Chen, Q., Hu, X., & Cai, J. (2022). A review of visual SLAM methods for autonomous driving vehicles. *Engineering Applications of Artificial Intelligence, 114*, 104992.

Chen, L., Nakamura, Y., Kondo, K., & Mayol-Cuevas, W. (2019). Hotspot modeling of hand-machine interaction experiences from a head-mounted RGB-D camera. *Transactions on Information and Systems, E102.D*(2), 319–330.

Chong, E., Clark-Whitney, E., Southerland, A., Stubbs, E., Miller, C., Ajodan, E. L., Silverman, M. R., Lord, C., Rozga, A., Jones, R. M., & Rehg, J. M. (2020). Detection of eye contact with deep neural networks is as accurate as human experts. *Nature Communications, 11*(1), 6386.

Choudhary, A., Mishra, D., & Karmakar, A. (2020). Domain adaptive egocentric person Re-identification. In *International conference on computer vision and image processing*.

Cipresso, P., Giglioli, I. A. C., Raya, M. A., & Riva, G. (2018). The past, present, and future of virtual and augmented reality research: A network and cluster analysis of the literature. *Frontiers in Psychology, 9*, 309500.

Damen, D., Doughty, H., Farinella, G.M., Fidler, S., Furnari, A., Kazakos, E., Moltisanti, D., Munro, J., Perrett, T., Price, W., & Price, W. (2018). Scaling egocentric vision: The EPIC-KITCHENS dataset. In *ECCV*.

Damen, D., Leelasawassuk, T., Haines, O., Calway, A., & Mayol-Cuevas, W. (2014). You-do, I-learn: Discovering task relevant objects and their modes of interaction from multi-user egocentric video. In *BMVC*.

Damen, D., Doughty, H., Farinella, G. M., Furnari, A., Ma, J., Kazakos, E., Moltisanti, D., Munro, J., Perrett, T., Price, W., & Wray, M. (2022). Rescaling egocentric vision: Collection, pipeline and challenges for EPIC-KITCHENS-100. *IJCV, 130*, 33–55.

Damen, D., Leelasawassuk, T., & Mayol-Cuevas, W. (2016). You-Do, I-Learn: Egocentric unsupervised discovery of objects and their modes of interaction towards video-based guidance. *CVIU, 149*, 98–112.

Dancette, C., Whitehead, S., Maheshwary, R., Vedantam, R., Scherer, S., Chen, X., Cord, M., & Rohrbach, M. (2023). Improving selective visual question answering by learning from your peers. In *CVPR*.

Dargan, S., Bansal, S., Kumar, M., Mittal, A., & Kumar, K. (2023). Augmented reality: A comprehensive review. *Archives of Computational Methods in Engineering, 30*(2), 1057–1080.

Darkhalil, A., Shan, D., Zhu, B., Ma, J., Kar, A., Higgins, R., Fidler, S., Fouhey, D., & Damen, D. (2022). EPIC-KITCHENS VISOR benchmark: VIdeo segmentations and object relations. In *NeurIPS*.

Das, A., Datta, S., Gkioxari, G., Lee, S., Parikh, D., & Batra, D. (2018). Embodied question answering. In *CVPR*.

Datta, S., Dharur, S., Cartillier, V., Desai, R., Khanna, M., Batra, D., & Parikh, D. (2022). Episodic memory question answering. In *CVPR*.

Davison, A.J. (2003). Real-time simultaneous localisation and mapping with a single camera. In *ICCV*.

de Santana, Correia A., & Colombini, E. L. (2022). Attention, please! A survey of neural attention models in deep learning. *Artificial Intelligence Review, 55*(8), 6037–6124.

De Smedt, Q., Wannous, H., & Vandeborre, J.P. (2016). Skeleton-based dynamic hand gesture recognition. In *CVPR workshop*.

Del Molino, A. G., Tan, C., Lim, J. H., & Tan, A. H. (2016). Summarization of egocentric videos: A comprehensive survey. *Transactions on Human–Machine Systems, 47*(1), 65–76.

DeMenthon, D., Kobla, V., & Doermann, D. (1998). Video summarization by curve simplification. In *International conference on multimedia*

Deng, A., Yang, T., & Chen, C. (2023). A large-scale study of spatiotemporal representation learning with a new benchmark on action recognition. In *ICCV*.

Denning, T., Dehlawi, Z., & Kohno, T. (2014). In situ with bystanders of augmented reality glasses: Perspectives on recording and privacy-mediating technologies. In *Conference on human factors in computing systems*.

Devagiri, J. S., Paheding, S., Niyaz, Q., Yang, X., & Smith, S. (2022). Augmented reality and artificial intelligence in industry: Trends, tools, and future challenges. *Expert Systems with Applications, 207*, 118002.

Dey, A., Billinghurst, M., Lindeman, R. W., & Swan, J. E. (2018). A systematic review of 10 years of augmented reality usability studies: 2005 to 2014. *Frontiers in Robotics and AI, 5*, 37.

Dimiccoli, M., Marín, J., & Thomaz, E. (2018). Mitigating bystander privacy concerns in egocentric activity recognition with deep learning and intentional image degradation. *ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 1*(4), 1–18.

Do, T., Miksik, O., DeGol, J., Park, H.S., & Sinha, S.N. (2022a). Learning to detect scene landmarks for camera localization. In *CVPR*.

Do, T., Vuong, K., & Park, H.S. (2022b). Egocentric scene understanding via multimodal spatial rectifier. In *CVPR*.

Donley, J., Tourbabin, V., Lee, J.S., Broyles, M., Jiang, H., Shen, J., Pantic, M., Ithapu, V.K., & Mehra, R. (2021). EasyCom: An augmented reality dataset to support algorithms for easy communication in noisy environments. arXiv:2107.04174

Doughty, H., Mayol-Cuevas, W., & Damen, D. (2019). The pros and cons: Rank-aware temporal attention for skill determination in long videos. In *CVPR*.

Duarte, N. F., Raković, M., Tasevski, J., Coco, M. I., Billard, A., & Santos-Victor, J. (2018). Action anticipation: Reading the intentions of humans and robots. *Robotics and Automation Letters, 3*(4), 4132–4139.

Dunnhofer, M., Furnari, A., Farinella, G. M., & Micheloni, C. (2023). Visual object tracking in first person vision. *IJCV, 131*(1), 259–283.

Dusmanu, M., Schonberger, J.L., Sinha, S.N., & Pollefeys, M. (2021). Privacy-preserving image features via adversarial affine subspace embeddings. In *CVPR*.

Elfeki, M., Wang, L., & Borji, A. (2022). Multi-stream dynamic video Summarization. In *WACV*.

Fan, C. (2019). EgoVQA—An egocentric video question answering benchmark dataset. In *ICCV workshop*.

Fan, C., Lee, J., Xu, M., Kumar Singh, K., Jae Lee, Y., Crandall, D.J., & Ryoo, M.S. (2017). Identifying first-person camera wearers in third-person videos. In *CVPR*.

Fan, Z., Taheri, O., Tzionas, D., Kocabas, M., Kaufmann, M., Black, M.J., & Hilliges, O. (2023). ARCTIC: A dataset for dexterous bimanual hand-object manipulation. In *CVPR*.

Farringdon, J., & Oni, V. (2000). Visual augmented memory (VAM). In *International symposium on wearable computers*.

Fathi, A., Hodgins, J.K., & Rehg, J.M. (2012a). Social interactions: A first-person perspective. In *CVPR*.

Fathi, A., Li, Y., & Rehg, J.M., et al. (2012b). Learning to recognize daily actions using gaze. In *ECCV*.

Fathi, A., Ren, X., & Rehg, J.M. (2011). Learning to recognize objects in egocentric activities. In *CVPR*.

Fergnani, F., Alletto, S., Serra, G., De Mira, J., & Cucchiara, R. (2016). Body part based re-identification from an egocentric perspective. In *CVPR workshop*.

Firat, A.K., Woon, W.L., & Madnick, S. (2008). Technological forecasting—A review. Composite Information Systems Laboratory (CISL), Massachusetts Institute of Technology (pp. 1–19).

Fu, D., Chen, D., Bao, J., Yang, H., Yuan, L., Zhang, L., Li, H., & Chen, D. (2021). Unsupervised Pre-training for Person Re-identification. In *CVPR*.

Fu, D., Chen, D., Yang, H., Bao, J., Yuan, L., Zhang, L., Li, H., Wen, F., & Chen, D. (2022a). Large-scale pre-training for person Re-identification with noisy labels. In *CVPR*.

Fu, Q., Liu, X., & Kitani, K. (2022b). Sequential voting with relational box fields for active object detection. In *CVPR*.

Furnari, A., & Farinella, G. M. (2019). What would you expect? Anticipating egocentric actions with rolling-unrolling LSTMs and modality attention. In *ICCV*.

Furnari, A., Farinella, G.M., & Battiato, S. (2016). Temporal segmentation of egocentric videos to highlight personal locations of interest. In *ECCV workshop*.

Furnari, A., Battiato, S., Grauman, K., & Farinella, G. M. (2017). Next-active-object prediction from egocentric videos. *Journal of Visual Communication and Image Representation, 49*, 401–411.

Gabeur, V., Seo, P.H., Nagrani, A., Sun, C., Alahari, K., & Schmid, C. (2022). AVATAR: Unconstrained audiovisual speech recognition. In *INTERSPEECH*.

Gálvez-López, D., & Tardos, J. D. (2012). Bags of binary words for fast place recognition in image sequences. *Transactions on Robotics, 28*(5), 1188–1197.

Gao, D., Wang, R., Bai, Z., & Chen, X. (2021). Env-QA: A video question answering benchmark for comprehensive understanding of dynamic environments. In *ICCV*.

Gao, J., Yang, Z., & Nevatia, R. (2017). RED: Reinforced encoder–decoder networks for action anticipation. In *BMVC*.

Gao, D., Zhou, L., Ji, L., Zhu, L., Yang, Y., & Shou, M.Z. (2023). MIST: Multi-modal iterative spatial-temporal transformer for long-form video question answering. In *CVPR*.

Garcia-Hernando, G., Yuan, S., Baek, S., & Kim, T.K. (2017). First-person hand action benchmark with RGB-D videos and 3D hand pose annotations. In *CVPR*.

Gers, F. A., Schmidhuber, J., & Cummins, F. (2000). Learning to forget: Continual prediction with LSTM. *Neural Computation, 12*(10), 2451–2471.

Ghosh, S., Dhall, A., Hayat, M., Knibbe, J., & Ji, Q. (2023). Automatic gaze analysis: A survey of deep learning based approaches. In *TPAMI*.

Girdhar, R., & Grauman, K. (2021). Anticipative video transformer. In *ICCV*.

Girdhar, R., El-Nouby, A., Liu, Z., Singh, M., Alwala, K.V., Joulin, A., & Misra, I. (2023). ImageBind: One Embedding space to bind them all. In *CVPR*.

Girdhar, R., Singh, M., Ravi, N., van der Maaten, L., Joulin, A., & Misra, I. (2022). Omnivore: A single model for many visual modalities. In *CVPR*.

Gong, X., Mohan, S., Dhingra, N., Bazin, J.C., Li, Y., Wang, Z., & Ranjan, R. (2023). MMG-Ego4D: Multi-modal generalization in egocentric action recognition. In *CVPR*.

González-Sosa, E., Robledo, G., Morín, D.G., Perez-Garcia, P., & Villegas, Á. (2021). Real time egocentric object segmentation: Thu-read labeling and benchmarking results. In *CVPR workshop*.

Gordon, D., Kembhavi, A., Rastegari, M., Redmon, J., Fox, D., & Farhadi, A. (2018). IQA: Visual question answering in interactive environments. In *CVPR*.

Gori, I., Aggarwal, J., Matthies, L., & Ryoo, M. S. (2016). Multi-type activity recognition in robot-centric scenarios. *Robotics and Automation Letters, 1*(1), 593–600.

Grauman, K., Westbury, A., Byrne, E., Chavis, Z., Furnari, A., Girdhar, R., Hamburger, J., Jiang, H., Liu, M., Liu, X., & Martin, M. (2022). Ego4D: Around the world in 3000 hours of egocentric video. In *CVPR*.

Grauman, K., Westbury, A., Torresani, L., Kitani, K., Malik, J., Afouras, T., Ashutosh, K., Baiyya, V., Bansal, S., Boote, B., & Byrne, E.

(2023). Ego-Exo4D: Understanding skilled human activity from first- and third-person perspectives. arXiv:2311.18259

Gu, X., Qiu, J., Guo, Y., Lo, B., & Yang, G.Z. (2021). TransAction: ICL-SJTU submission to EPIC-kitchens action anticipation challenge 2021. arXiv:2107.13259.

Gurari, D., Li, Q., Stangl, A.J., Guo, A., Lin, C., Grauman, K., Luo, J., & Bigham, J.P. (2018). VizWiz grand challenge: Answering visual questions from blind people. In *CVPR*.

Gutierrez-Gomez, D., & Guerrero, J. (2016). True scaled 6 DoF egocentric localisation with monocular wearable systems. *Image and Vision Computing, 52*, 178–194.

Gygli, M., Grabner, H., Riemenschneider, H., & Van Gool, L. (2014). Creating summaries from user videos. In *ECCV*.

Haitz, D., Jutzi, B., Ulrich, M., Jäger, M., & Hübner, P. (2023). Combining HoloLens with instant-NeRFs: Advanced real-time 3D mobile mapping. arXiv:2304.14301

Hampali, S., Hodan, T., Tran, L., Ma, L., Keskin, C., & Lepetit, V. (2023). In-hand 3D object scanning from an RGB sequence. In *CVPR*.

Hasan, R., Crandall, D., Fritz, M., & Kapadia, A. (2020). Automatically detecting bystanders in photos to reduce privacy risks. In *Symposium on security and privacy*.

Hassan, E.T., Hasan, R., Shaffer, P., Crandall, D., & Kapadia, A. (2017). Cartooning for enhanced privacy in lifelogging and streaming videos. In *CVPR workshop*.

Hasson, Y., Varol, G., Tzionas, D., Kalevatykh, I., Black, M.J., Laptev, I., & Schmid, C. (2019). Learning joint reconstruction of hands and manipulated objects. In *CVPR*.

Hatano, M., Hachiuma, R., & Saito, H. (2023). Trajectory prediction in first-person video: Utilizing a pre-trained bird's-eye view model. In *International conference on computer vision theory and applications*.

He, B., Wang, J., Qiu, J., Bui, T., Shrivastava, A., & Wang, Z. (2023). Align and attend: Multimodal summarization with dual contrastive losses. In *CVPR*.

Herzig, R., Ben-Avraham, E., Mangalam, K., Bar, A., Chechik, G., Rohrbach, A., Darrell, T., & Globerson, A. (2022). Object-region video transformers. In *CVPR*.

Ho, H.I., Chiu, W.C., & Wang, Y.C.F. (2018). Summarizing first-person videos from third persons' points of views. In *ECCV*.

Hoshen, Y., & Peleg, S. (2016). An egocentric look at video photographer identity. In *CVPR*.

Hoshen, Y., Ben-Artzi, G., & Peleg, S. (2014). Wisdom of the crowd in egocentric video curation. In *CVPR workshop*.

Hoyle, R., Templeman, R., Anthony, D., Crandall, D., & Kapadia, A. (2015). Sensitive lifelogs: A privacy analysis of photos from wearable cameras. In *Conference on human factors in computing systems*.

Hoyle, R., Templeman, R., Armes, S., Anthony, D., Crandall, D., & Kapadia, A. (2014). Privacy behaviors of lifeloggers using wearable cameras. In *International joint conference on pervasive and ubiquitous computing*.

Huang, Y., Cai, M., Li, Z., & Sato, Y. (2018). Predicting gaze in egocentric video by learning task-dependent attention transition. In *ECCV*.

Huang, M., Li, X., Hu, J., Peng, H., & Lyu, S. (2023b). Tracking multiple deformable objects in egocentric videos. In *CVPR*.

Huang, Y., Liu, X., Zhang, X., & Jin, L. (2016). A pointing gesture based egocentric interaction system: Dataset. approach and application. In *CVPR workshop*.

Huang, Y., Sugano, Y., & Sato, Y. (2020b). Improving action segmentation via graph-based temporal reasoning. In *CVPR*.

Huang, C., Tian, Y., Kumar, A., & Xu, C. (2023a). Egocentric audio-visual object localization. In *CVPR*.

Huang, Y., Cai, M., & Sato, Y. (2020). An ego-vision system for discovering human joint attention. *Transactions on Human–Machine Systems, 50*(4), 306–316.

Hübner, P., Clintworth, K., Liu, Q., Weinmann, M., & Wursthorn, S. (2020). Evaluation of HoloLens tracking and depth sensing for indoor mapping applications. *Sensors, 20*(4), 1021.

Huh, J., Chalk, J., Kazakos, E., Damen, D., & Zisserman, A. (2023). Epic-sounds: A large-scale dataset of actions that sound. In *ICASSP*.

Hung-Cuong, N., Nguyen, T. H., Scherer, R., & Le, V. H. (2023). YOLO series for human hand action detection and classification from egocentric videos. *Sensors, 23*(6), 3255.

Ionescu, C., Papava, D., Olaru, V., & Sminchisescu, C. (2013). Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. *TPAMI, 36*(7), 1325–1339.

Irschara, A., Zach, C., Frahm, J.M., & Bischof, H. (2009). From structure-from-motion point clouds to fast location recognition. In *CVPR*.

Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *TPAMI, 20*(11), 1254–1259.

Jana, S.S., Molnar, D.A., Moshchuk, A., Dunn, A.M., Livshits, B., Wang, H.J., & Ofek, E. (2013b). Enabling fine-grained permissions for augmented reality applications with recognizers. In *USENIX security symposium*.

Jana, S., Narayanan, A., & Shmatikov, V. (2013a). A scanner darkly: Protecting user privacy from perceptual applications. In *Symposium on security and privacy*.

Jégou, H., Douze, M., Schmid, C., & Pérez, P. (2010). Aggregating local descriptors into a compact image representation. In *CVPR*.

Jia, B., Lei, T., Zhu, S.C., & Huang, S. (2022a). EgoTaskQA: Understanding human tasks in egocentric videos. In *NeurIPS*.

Jia, W., Liu, M., & Rehg, J.M. (2022b). Generative adversarial network for future hand segmentation from egocentric video. In *ECCV*.

Jiang, H., & Grauman, K. (2017). Seeing invisible poses: Estimating 3D body pose from egocentric video. In *CVPR*.

Jiang, H., Murdock, C., & Ithapu, V.K. (2022). Egocentric deep multi-channel audio-visual active speaker localization. In *CVPR*.

Jiang, H., Ramakrishnan, S.K., & Grauman, K. (2023). Single-stage visual query localization in egocentric videos. In *NeurIPS*.

Johnson, M., & Demiris, Y. (2005). Perceptual perspective taking and action recognition. *International Journal of Advanced Robotic Systems, 2*(4), 32.

Kai, C., Haihua, Z., Dunbing, T., & Kun, Z. (2023). Future pedestrian location prediction in first-person videos for autonomous vehicles and social robots. *Image and Vision Computing, 134*, 104671.

Kang, H., Hebert, M., & Kanade, T. (2011). Discovering object instances from scenes of Daily Living. In *ICCV*.

Karunratanakul, K., Prokudin, S., Hilliges, O., & Tang, S. (2023). HARP: personalized hand reconstruction from a monocular RGB video. In *CVPR*.

Kazakos, E., Huh, J., Nagrani, A., Zisserman, A., & Damen, D. (2021). With a little help from my temporal context: Multimodal egocentric action recognition. In *BMVC*.

Kazakos, E., Nagrani, A., Zisserman, A., & Damen, D. (2019). EPIC-fusion: Audio-visual temporal binding for egocentric action recognition. In *ICCV*.

Kazerouni, I. A., Fitzgerald, L., Dooly, G., & Toal, D. (2022). A survey of state-of-the-art on visual SLAM. *Expert Systems with Applications, 205*, 117734.

Kendall, A., Grimes, M., & Cipolla, R. (2015). PoseNet: A convolutional network for real-time 6-DOF camera relocalization. In *ICCV*.

Kera, H., Yonetani, R., Higuchi, K., & Sato, Y. (2016). Discovering objects of joint attention via first-person sensing. In *CVPR workshop*.

Keskin, C., Kıraç, F., Kara, Y.E., & Akarun, L. (2012). Hand pose estimation and hand shape classification using multi-layered randomized decision forests. In *ECCV*.

Khan, M., Fernandes, G., Vaish, A., Manuja, M., & Maes, P. (2021). Wearable system for personalized and privacy-preserving egocentric visual context detection using on-device deep learning. In *International conference on user modeling, adaptation, and personalization*.

Khirodkar, R., Bansal, A., Ma, L., Newcombe, R., Vo, M., & Kitani, K. (2023). EgoHumans: An egocentric 3D multi-human benchmark. In *ICCV*.

Khosla, A., Hamid, R., Lin, C.J., & Sundaresan, N. (2013). Large-scale video summarization using web-image priors. In *CVPR*.

Kim, D., Tsai, Y.H., Zhuang, B., Yu, X., Sclaroff, S., Saenko, K., & Chandraker, M. (2021). Learning cross-modal contrastive features for video domain adaptation. In *ICCV*.

Kitani, K.M., Okabe, T., Sato, Y., & Sugimoto, A. (2011). Fast unsupervised ego-action learning for first-person sports videos. In *CVPR*.

Kopf, J., Cohen, M. F., & Szeliski, R. (2014). First-person hyper-lapse videos. *Transactions on Graphics, 33*(4), 1–10.

Koppula, H. S., & Saxena, A. (2015). Anticipating human activities using object affordances for reactive robotic response. *TPAMI, 38*(1), 14–29.

Korayem, M., Templeman, R., Chen, D., Crandall, D., & Kapadia, A. (2016). Enhancing lifelogging privacy by detecting screens. In *Conference on human factors in computing systems*.

Krishna, S., Little, G., Black, J., & Panchanathan, S. (2005). A wearable face recognition system for individuals with visual impairments. In *International conference on computers and accessibility*.

Kumano, S., Otsuka, K., Ishii, R., & Yamato, J. (2015). Automatic gaze analysis in multiparty conversations based on collective first-person vision. In *International conference and workshops on automatic face and gesture recognition*.

Kurita, S., Katsura, N., & Onami, E. (2023). RefEgo: Referring expression comprehension dataset from first-person perception of Ego4D. In *ICCV*.

Kurze, M., & Roselius, A. (2011). Smart glasses linking real live and social network's contacts by face recognition. In *Augmented humans international conference*.

Lai, B., Liu, M., Ryan, F., & Rehg, J. (2022). In the eye of transformer: Global-local correlation for egocentric gaze estimation. In *BMVC*.

Lai, B., Ryan, F., Jia, W., Liu, M., & Rehg, J.M. (2023a). Listen to look into the future: Audio-visual egocentric gaze anticipation. arXiv:2305.03907

Lai, B., Zhang, H., Liu, M., Pariani, A., Ryan, F., Jia, W., Hayati, S.A., Rehg, J., & Yang, D. (2023b). Werewolf among us: Multimodal resources for modeling persuasion behaviors in social deduction games. In *Findings of the association for computational linguistics*.

Lan, T., Chen, T.C., & Savarese, S. (2014). A hierarchical representation for future action prediction. In *ECCV*.

Lange, M.D., Eghbalzadeh, H., Tan, R., Iuzzolino, M.L., Meier, F., & Ridgeway, K. (2023). EgoAdapt: A multi-stream evaluation study of adaptation to real-world egocentric user video. arXiv:2307.05784.

Lee, K., & Kacorri, H. (2019). hands holding clues for object recognition in teachable machines. In *Conference on human factors in computing systems*.

Lee, Y.J., Ghosh, J., & Grauman, K. (2012). Discovering important people and objects for egocentric video summarization. In *CVPR*.

Lee, J., Sung, M., Choi, H., & Kim, T.K. (2023). Im2Hands: Learning attentive implicit representation of interacting two-hand shapes. In *CVPR*.

Leonardi, R., Ragusa, F., Furnari, A., & Farinella, G.M. (2022). Egocentric human-object interaction detection exploiting synthetic data. In *ICIAP*.

Leonardi, R., Ragusa, F., Furnari, A., & Farinella, G.M. (2023). Exploiting multimodal synthetic data for egocentric human-object interaction detection in an industrial scenario. arXiv:2306.12152

Li, Z., Abu Farha, Y., & Gall, J. (2021c). Temporal action segmentation from timestamp supervision. In *CVPR*.

Li, H., Cai, Y., & Zheng, W.S. (2019a). Deep dual relation modeling for egocentric interaction recognition. In *CVPR*.

Li, Y., Cao, Z., Liang, A., Liang, B., Chen, L., Zhao, H., & Feng, C. (2022). Egocentric prediction of action target in 3D. In *CVPR*.

Li, Y., Fathi, A., & Rehg, J.M. (2013). Learning to predict gaze in egocentric video. In *ICCV*.

Li, Z., Huang, Y., Cai, M., & Sato, Y. (2019b). Manipulation-skill assessment from videos with spatial attention network. In *ICCV workshop*.

Li, J., Liu, K., & Wu, J. (2023). Ego-body pose estimation via ego-head pose estimation. In *CVPR*.

Li, Y., Nagarajan, T., Xiong, B., & Grauman, K. (2021b). Ego-exo: Transferring visual representations from third-person to first-person videos. In *CVPR*.

Li, Y., Ye, Z., & Rehg, J.M. (2015). Delving into egocentric actions. In *CVPR*.

Liao, S., & Shao, L. (2021). TransMatcher: Deep image matching through transformers for generalizable person Re-identification. In *NeurIPS*.

Li, Y., Liu, M., & Rehg, J. (2021). In the eye of the beholder: Gaze and actions in first person video. *TPAMI, 45*(06), 6731–6747.

Lin, Y., Abdelfatah, K., Zhou, Y., Fan, X., Yu, H., Qian, H., & Wang, S. (2015). Co-interest person detection from multiple wearable camera videos. In *ICCV*.

Lin, K.Q., Wang, J., Soldan, M., Wray, M., Yan, R., Xu, Z., Gao, D., Tu, R.C., Zhao, W., Kong, W., Cai, C., HongFa, W., Damen, D., Ghanem, B., Liu, W., & Shou, M.Z. (2022). Egocentric video-language pretraining. In *NeurIPS*.

Liu, S.W., Jiang, H., Xu, J., Liu, S., & Wang, X. (2021c). Semi-supervised 3D hand-object poses estimation with interactions in time. In *CVPR*.

Liu, H., Li, C., Wu, Q., & Lee, Y.J. (2023b). Visual instruction tuning. In *NeurIPS*.

Liu, Y., Liu, Y., Jiang, C., Lyu, K., Wan, W., Shen, H., Liang, B., Fu, Z., Wang, H., & Yi, L. (2022b). HOI4D: A 4D Egocentric dataset for category-level human-object interaction. In *CVPR*.

Liu, Y., Liu, Y., Jiang, C., Lyu, K., Wan, W., Shen, H., Liang, B., Fu, Z., Wang, H., & Yi, L. (2022c). HOI4D: A 4D egocentric dataset for category-level human-object interaction. In *CVPR*.

Liu, M., Ma, L., Somasundaram, K., Li, Y., Grauman, K., Rehg, J.M., & Li, C. (2022a). Egocentric activity recognition and localization on a 3D map. In *ECCV*.

Liu, G., Tang, H., Latapie, H., & Yan, Y. (2020a). Exocentric to egocentric image generation via parallel generative adversarial network. In *ICASSP*.

Liu, G., Tang, H., Latapie, H. M., Corso, J. J., & Yan, Y. (2021). Cross-view exocentric to egocentric video synthesis. In *International conference on multimedia*.

Liu, M., Tang, S., Li, Y., & Rehg, J.M. (2020b). Forecasting human-object interaction: Joint prediction of motor attention and actions in first person video. In *ECCV*.

Liu, Y., Wei, P., & Zhu, S.C. (2017). Jointly recognizing object fluents and tasks in egocentric videos. In *ICCV*.

Liu, M., Yang, D., Zhang, Y., Cui, Z., Rehg, J.M., & Tang, S. (2021b). 4D Human body capture from egocentric video via 3D scene grounding. In *3DV*.

Liu, X., Zhang, J., Hu, R., Huang, H., Wang, H., & Yi, L. (2023c). Self-supervised category-level articulated object pose estimation with part-level SE(3) equivariance. In *ICLR*.

Liu, B., Zheng, S., Fu, J., & Cheng, W. H. (2023). Anchor-based detection for natural language localization in ego-centric videos. In *International conference on consumer electronics*.

Lu, Z., & Grauman, K. (2013). Story-driven summarization for egocentric video. In *CVPR*.

Lu, Y., & Mayol-Cuevas, W.W. (2021). Understanding egocentric hand-object interactions from hand pose estimation. arXiv:2109.14657

Lu, H., & Brimijoin, W. O. (2022). Sound source selection based on head movements in natural group conversation. *Trends in Hearing, 26*, 23312165221097788.

Luo, Z., Hachiuma, R., Yuan, Y., & Kitani, K. (2021). Dynamics-regulated kinematic policy for egocentric pose estimation. In *NeurIPS*.

Ma, Y.J., Sodhani, S., Jayaraman, D., Bastani, O., Kumar, V., & Zhang, A. (2023b). VIP: Towards universal visual reward and representation via value-implicit pre-training. In *ICLR*.

Ma, X., Yong, S., Zheng, Z., Li, Q., Liang, Y., Zhu, S.C., & Huang, S. (2023a). SQA3D: Situated question answering in 3D scenes. In *ICLR*.

Mai, J., Hamdi, A., Giancola, S., Zhao, C., & Ghanem, B. (2023). EgoLoc: Revisiting 3D object localization from egocentric videos with visual queries. In *ICCV*.

Majumder, S., Jiang, H., Moulon, P., Henderson, E., Calamia, P., Grauman, K., & Ithapu, V.K. (2023). Chat2Map: Efficient scene mapping from multi-ego conversations. In *CVPR*.

Mandal, B., Chia, S.C., Li, L., Chandrasekhar, V., Tan, C., & Lim, J.H. (2015). A wearable face recognition system on google glass for assisting social interactions. In *ACCV*.

Mangalam, K., Akshulakov, R., & Malik, J. (2023). EgoSchema: A diagnostic benchmark for very long-form video language understanding. In *NeurIPS*.

Mascaró, E.V., Ahn, H., & Lee, D. (2023). Intention-conditioned long-term human egocentric action anticipation. In *WACV*.

Massiceti, D., Anjum, S., & Gurari, D. (2022). VizWiz grand challenge workshop at CVPR 2022. In *SIGACCESS accessibility and computing*.

Mayol, W.W., Davison, A.J., Tordoff, B.J., & Murray, D.W. (2005). Applying active vision and SLAM to wearables. In *Robotics research*.

Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R., & Ng, R. (2021). NeRF: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM, 65*(1), 99–106.

Milotta, F. L., Furnari, A., Battiato, S., Signorello, G., & Farinella, G. M. (2019). Egocentric visitors localization in natural sites. *Journal of Visual Communication and Image Representation, 65*, 102664.

Min, K., & Corso, J.J. (2021). Integrating human gaze into attention for egocentric activity recognition. In *WACV*.

Molchanov, P., Yang, X., Gupta, S., Kim, K., Tyree, S., & Kautz, J. (2016). Online detection and classification of dynamic hand gestures with recurrent 3D convolutional neural networks. In *CVPR*.

Moltisanti, D., Fidler, S., & Damen, D. (2019). Action recognition from single timestamp supervision in untrimmed videos. In *CVPR*.

Mueller, F., Bernard, F., Sotnychenko, O., Mehta, D., Sridhar, S., Casas, D., & Theobalt, C. (2017). GANerated hands for real-time 3D hand tracking from monocular RGB. In *CVPR*.

Munro, J., & Damen, D. (2020). Multi-modal domain adaptation for fine-grained action recognition. In *CVPR*.

Murillo, A.C., Gutiérrez-Gómez, D., Rituerto, A., Puig, L., & Guerrero, J.J. (2012). Wearable omnidirectional vision system for personal localization and guidance. In *CVPR workshop*.

Mur-Labadia, L., Guerrero, J.J., & Martinez-Cantin, R. (2023). Multi-label affordance mapping from egocentric vision. In *ICCV*.

Nagarajan, T., & Grauman, K. (2020). Learning affordance landscapes for interaction exploration in 3D environments. In *NeurIPS*.

Nagarajan, T., Feichtenhofer, C., & Grauman, K. (2019). Grounded human-object interaction hotspots from video. In *ICCV*.

Nagarajan, T., Ramakrishnan, S.K., Desai, R., Hillis, J., & Grauman, K. (2023). EgoEnv: Human-centric environment representations from egocentric video. In *NeurIPS*.

Nagar, P., Rathore, A., Jawahar, C., & Arora, C. (2021). Generating personalized summaries of day long egocentric videos. *TPAMI, 45*(6), 6832–6845.

Nair, S., Rajeswaran, A., Kumar, V., Finn, C., & Gupta, A. (2022). R3M: A universal visual representation for robot manipulation. In *CoRL*.

Narayan, S., Kankanhalli, M.S., & Ramakrishnan, K.R. (2014). Action and interaction recognition in first-person videos. In *CVPR workshop*.

Ng, T., Kim, H.J., Lee, V.T., DeTone, D., Yang, T.Y., Shen, T., Ilg, E., Balntas, V., Mikolajczyk, K., & Sweeney, C. (2022). NinjaDesc: Content-concealing visual descriptors via adversarial learning. In *CVPR*.

Ng, E., Xiang, D., Joo, H., & Grauman, K. (2020). You2Me: Inferring body pose in egocentric video via first and second person interactions. In *CVPR*.

Nguyen, T. H. C., Nebel, J. C., & Florez-Revuelta, F. (2016). Recognition of activities of daily living with egocentric vision: A review. *Sensors, 16*(1), 72.

Northcutt, C., Zha, S., Lovegrove, S., & Newcombe, R. (2020). EgoCom: A multi-person multi-modal egocentric communications dataset. *TPAMI, 45*(6), 6783–6793.

Núñez-Marcos, A., Azkune, G., & Arganda-Carreras, I. (2022). Egocentric vision-based action recognition: A survey. *Neurocomputing, 472*, 175–197.

Oechsle, M., Peng, S., & Geiger, A. (2021). UNISURF: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction. In *ICCV*.

Ohkawa, T., He, K., Sener, F., Hodan, T., Tran, L., & Keskin, C. (2023). AssemblyHands: Towards egocentric activity understanding via 3D hand pose estimation. In *CVPR*.

Oikonomidis, I., Kyriazis, N., & Argyros, A.A. (2011). Efficient model-based 3D tracking of hand articulations using kinect. In *BMVC*.

Okamoto, M., & Yanai, K. (2014). Summarization of egocentric moving videos for generating walking route guidance. In *Pacific-rim symposium on image and video technology*.

Oord, A.V.D., Li, Y., & Vinyals, O. (2018). Representation learning with contrastive predictive coding. arXiv:1807.03748

Orlando, S.A., Furnari, A., & Farinella, G.M. (2020). Virtual to real unsupervised domain adaptation for image-based localization in cultural sites. In *International conference on image processing, applications and systems*.

Pan, B., Shen, B., Rempe, D., Paschalidou, D., Mo, K., Yang, Y., & Guibas, L.J. (2023a). COPILOT: Human-environment collision prediction and localization from egocentric videos. In *ICCV*.

Pan X, Charron N, Yang Y, Peters S, Whelan T, Kong C, Parkhi O, Newcombe R, & Ren, Y.C. (2023b). Aria digital twin: A new benchmark dataset for egocentric 3D machine perception. In *ICCV*.

Panek, V., Kukelova, Z., & Sattler, T. (2022). MeshLoc: Mesh-based visual localization. In *ECCV*.

Papanagiotou, D., Senteri, G., & Manitsaris, S. (2021). Egocentric gesture recognition using 3D convolutional neural networks for the spatiotemporal adaptation of collaborative robots. *Frontiers in Neurorobotics, 15*, 703545.

Park, H.S., Hwang, J.J., Niu, Y., & Shi, J. (2016). Egocentric future localization. In *CVPR*.

Park, H., Jain, E., & Sheikh, Y. (2012). 3D social saliency from head-mounted cameras. In *NeurIPS*.

Park, H.S., Jain, E., & Sheikh, Y. (2013). Predicting primary gaze behavior using social saliency fields. In *ICCV*.

Park, H.S., & Shi, J. (2015). Social saliency prediction. In *CVPR*.

Parmar, P., & Morris, B. T. (2019). What and How Well You Performed? A multitask learning approach to action quality assessment. In *CVPR*.

Pasca, R.G., Gavryushin, A., Kuo, Y.L., Hilliges, O., & Wang, X. (2023). Summarize the past to predict the future: Natural language descriptions of context boost multimodal object interaction. arXiv:2301.09209

Pasqualino, G., Furnari, A., & Farinella, G.M. (2022a). A multi camera unsupervised domain adaptation pipeline for object detection in cultural sites through adversarial learning and self-training. In *CVIU*, (p. 103487).

Pasqualino, G., Furnari, A., & Farinella, G.M. (2022b). Unsupervised multi-camera domain adaptation for object detection in cultural sites. In *ICIAP*.

Pathirana, P., Senarath, S., Meedeniya, D., & Jayarathna, S. (2022). Eye gaze estimation: A survey on deep learning-based approaches. *Expert Systems with Applications, 199*, 116894.

Patra, S., Aggarwal, H., Arora, H., Banerjee, S., & Arora, C. (2017). Computing egomotion with local loop closures for egocentric videos. In *WACV*.

Patrick, M., Campbell, D., Asano, Y.M., Metze, I.M.F., Feichtenhofer, C., Vedaldi, A., & Henriques, J.F. (2021). Keeping your eye on the ball: Trajectory attention in video transformers. In *NeurIPS*.

Pavlakos, G., Shan, D., Radosavovic, I., Kanazawa, A., Fouhey, D., & Malik, J. (2024). Reconstructing hands in 3d with transformers. arXiv:2312.05251

Pavlovic, V. I., Sharma, R., & Huang, T. S. (1997). Visual interpretation of hand gestures for human–computer interaction: A review. *TPAMI, 19*(7), 677–695.

Pei, M., Jia, Y., & Zhu, S.C. (2011). Parsing video events with goal inference and intent prediction. In *ICCV*.

Perrett, T., Sinha, S., Burghardt, T., Mirmehdi, M., & Damen, D. (2023). Use your head: Improving long-tail video recognition. In *CVPR*.

Petric, F., Hrvatinić, K., Babić, A., Malovan, L., Miklić, D., Kovačić, Z., Cepanec, M., Stošić, J., Šimleša, S. (2014). Four tasks of a robot-assisted autism spectrum disorder diagnostic protocol: First clinical tests. In *Global humanitarian technology conference*.

Pietrantoni, M., Humenberger, M., Sattler, T., & Csurka, G. (2023). SegLoc: Learning segmentation-based representations for privacy-preserving visual localization. In *CVPR*.

Pirsiavash, H., & Ramanan, D. (2012). Detecting activities of daily living in first-person camera views. In *CVPR*.

Pittaluga, F., Koppal, S.J., Kang, S.B., & Sinha, S.N. (2019). Revealing scenes by inverting structure from motion reconstructions. In *CVPR*.

Plizzari, C., Perrett, T., Caputo, B., & Damen, D. (2023). What can a cook in Italy teach a mechanic in India? Action recognition generalisation over scenarios and locations. In *ICCV*.

Plizzari, C., Planamente, M., Goletto, G., Cannici, M., Gusso, E., Matteucci, M., & Caputo, B. (2022). E2(GO)MOTION: Motion augmented event stream for egocentric action recognition. In *CVPR*.

Poleg, Y., Arora, C., & Peleg, S. (2015a). Head motion signatures from egocentric videos. In Cremers, D., Reid, I., Saito, H., Yang, M.H. (eds.) *ACCV*.

Poleg, Y., Ephrat, A., Peleg, S., & Arora, C. (2016). Compact CNN for indexing egocentric videos. In *WACV*.

Poleg, Y., Halperin, T., Arora, C., & Peleg, S. (2015b). EgoSampling: Fast-forward and stereo for egocentric videos. In *CVPR*.

Possas, R., Caceres, S.P., & Ramos, F. (2018). Egocentric activity recognition on a budget. In *ICPR*.

Pramanick, S., Song, Y., Nag, S., Lin, K.Q., Shah, H., Shou, M.Z., Chellappa, R., & Zhang, P. (2023). EgoVLPv2: Egocentric video-language pre-training with fusion in the backbone. In *ICCV*.

Price, B. A., Stuart, A., Calikli, G., Mccormick, C., Mehta, V., Hutton, L., Bandara, A. K., Levine, M., & Nuseibeh, B. (2017). Logging you, logging me: a replicable study of privacy and sharing behaviour in groups of visual lifeloggers. *ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 1*(2), 1–18.

Purushwalkam, S., Morgado, P., & Gupta, A. (2022). The challenges of continuous self-supervised learning. In *ECCV*.

Qian, S., & Fouhey, D.F. (2023). Understanding 3D object interaction from a single image. In *ICCV*.

Qian, C., Sun, X., Wei, Y., Tang, X., & Sun, J. (2014). Realtime and robust hand tracking from depth. In *CVPR*.

Qiu, J., Lo, F. P. W., Gu, X., Jobarteh, M., Jia, W., Baranowski, T., Steiner, M., Anderson, A., McCrory, M., Sazonov, E., Sun, M., Frost, G., & Lo, B. (2023). Egocentric image captioning for privacy-preserved passive dietary intake monitoring. *Transactions on Cybernetics, 54*, 679–692.

Radevski, G., Grujicic, D., Blaschko, M., Moens, M.F., & Tuytelaars, T. (2023). Multimodal distillation for egocentric action recognition. In *ICCV*.

Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., & Clark, J., et al (2021). Learning transferable visual models from natural language supervision. In *ICML*.

Radosavovic, I., Xiao, T., James, S., Abbeel, P., Malik, J., & Darrell, T. (2022). Real-world robot learning with masked visual pre-training. In *CoRL*.

Ragusa, F., Farinella, G.M., & Furnari, A. (2023a). StillFast: An end-to-end approach for short-term object interaction anticipation. In *CVPR workshop*.

Ragusa, F., Furnari, A., & Farinella, G.M. (2023b). MECCANO: A multimodal egocentric dataset for humans behavior understanding in the industrial-like domain. In *CVIU*.

Ragusa, F., Furnari, A., Livatino, S., & Farinella, G.M. (2021). The MECCANO dataset: Understanding human-object interactions from egocentric videos in an industrial-like domain. In *WACV*.

Ragusa, F., Mauro, D.D., Palermo, A., Furnari, A., & Farinella, G.M. (2020b). Semantic object segmentation in cultural sites using real and synthetic data. In *ICPR*.

Ragusa, F., Furnari, A., Battiato, S., Signorello, G., & Farinella, G. M. (2020). EGO-CH: Dataset and fundamental tasks for visitors behavioral understanding using egocentric vision. *Pattern Recognition Letters, 131*, 150–157.

Ramakrishnan SK, Al-Halah Z, & Grauman K (2023). NaQ: Leveraging narrations as queries to supervise episodic memory. In *CVPR*.

Ramanathan, V., Kalia, A., Petrovic, V., Wen, Y., Zheng, B., Guo, B., Wang, R., Marquez, A., Kovvuri, R., Kadian, A., & Mousavi, A. (2023b). PACO: Parts and attributes of common objects. In *CVPR*.

Ramanathan, V., Kalia, A., Petrovic, V., Wen, Y., Zheng, B., Guo, B., Wang, R., Marquez, A., Kovvuri, R., Kadian, A., Mousavi, A., Song, Y., Dubey, A., & Mahajan, D. (2023a). PACO: Parts and attributes of common objects. In *CVPR*.

Ramos, W., Silva, M., Araujo, E., Neves, A., & Nascimento, E. (2020). Personalizing fast-forward videos based on visual and textual features from social network. In *WACV*.

Rathore, A., Nagar, P., Arora, C., & Jawahar, C. (2019). Generating 1 min summaries of day long egocentric videos. In *International conference on multimedia*.

Ravi, S., Climent-Perez, P., Morales, T., Huesca-Spairani, C., Hashemifard, K., & Flórez-Revuelta, F. (2023). ODIN: An omnidirectional indoor dataset capturing activities of daily living from multiple synchronized modalities. In *CVPR*.

Ren X, & Gu C (2010) Figure-ground segmentation improves handled object recognition in egocentric video. In *CVPR*.

Reza, S., Sundareshan, B., Moghaddam, M., & Camps, O.I. (2023). Enhancing transformer backbone for egocentric video action segmentation. In *CVPR workshop*.

Rhinehart, N., & Kitani, K.M. (2016). Learning action maps of large environments via first-person vision. In *CVPR*.

Rhodin, H., Richardt, C., Casas, D., Insafutdinov, E., Shafiei, M., Seidel, H. P., Schiele, B., & Theobalt, C. (2016). EgoCap: Egocentric marker-less motion capture with two fisheye cameras. *Transactions on Graphics, 35*(6), 1–11.

Rodin, I., Furnari, A., Mavroeidis, D., & Farinella, G.M. (2022). Untrimmed action anticipation. In *ICIAP*.

Rodin, I., Furnari, A., Mavroeidis, D., & Farinella, G. M. (2021). Predicting the future from first person (egocentric) vision: A survey. *CVIU, 211*, 103252.

Roesner, F., Kohno, T., & Molnar, D. (2014). Security and privacy for augmented reality systems. *Communications of the ACM, 57*(4), 88–96.

Rogez, G., Khademi, M., Supancic III J.S., Montiel, J.M.M., & Ramanan, D. (2015a). 3D hand pose detection in egocentric RGB-D images. In *ECCV workshop*.

Rogez, G., Khademi, M., Supancic, J.S., Montiel, J.M.M., & Ramanan, D. (2014). 3D hand pose detection in egocentric RGB-D images. In *ECCV workshop*.

Rogez, G., Supancic III, J.S., & Ramanan, D. (2015c). Understanding everyday hands in action from RGB-D images. In *ICCV*.

Rogez, G., Supancic, J.S., & Ramanan, D. (2015b). First-person pose recognition using egocentric workspaces. In *CVPR*.

Rosinol, A., Leonard, J.J., & Carlone, L. (2023). NeRF-SLAM: Real-time dense monocular SLAM with neural radiance fields. In *IROS*.

Roy, D., & Fernando, B. (2022). Action anticipation using latent goal learning. In *WACV*.

Roy, D., Rajendiran, R., & Fernando, B. (2024). Interaction region visual transformer for egocentric action anticipation. In *WACV*.

Rudnev, V., Golyanik, V., Wang, J., Seidel, H.P., Mueller, F., Elgharib, M., & Theobalt, C. (2021). EventHands: Real-time neural 3D hand pose estimation from an event stream. In *ICCV*.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., & Fei-Fei, L. (2015). ImageNet large scale visual recognition challenge. *IJCV, 115*(3), 211–252.

Ryan, F., Jiang, H., Shukla, A., Rehg, J.M., & Ithapu, V.K. (2023). Egocentric auditory attention localization in conversations. In *CVPR*.

Ryoo, M.S., & Matthies, L. (2013). First-person activity recognition: What are they doing to me? In *ICPR*.

Ryoo, M.S., Rothrock, B., & Matthies, L. (2015). Pooled motion features for first-person videos. In *CVPR*.

Ryoo, M.S., Rothrock, B., Fleming, C., & Yang, H.J. (2016). Privacy-preserving human activity recognition from extreme low resolution. In *Conference on artificial intelligence*.

Sabater, A., Alonso, I., Montesano, L., & Murillo, A. C. (2021). Domain and view-point agnostic hand action recognition. *Robotics and Automation Letters, 6*, 7823–7830.

Sahoo, A., Shah, R., Panda, R., Saenko, K., & Das, A. (2021). Contrast and mix: Temporal contrastive video domain adaptation with background mixing. In *NeurIPS*.

Sajjad, M., Nasir, M., Muhammad, K., Khan, S., Jan, Z., Sangaiah, A. K., Elhoseny, M., & Baik, S. W. (2020). Raspberry Pi assisted face recognition framework for enhanced law-enforcement services in smart cities. *Future Generation Computer Systems, 108*, 995–1007.

Sarlin, P.E., DeTone, D., Yang, T.Y., Avetisyan, A., Straub, J., Malisiewicz, T., Bulo, S.R., Newcombe, R., Kontschieder, P., & Balntas, V. (2023). OrienterNet: Visual localization in 2D public maps with neural matching. In *CVPR*.

Sarlin, P.E., Dusmanu, M., Schönberger, J.L., Speciale, P., Gruber, L., Larsson, V., Miksik, O., & Pollefeys, M. (2022). LaMAR: Benchmarking localization and mapping for augmented reality. In *ECCV*.

Sattler, T., Leibe, B., & Kobbelt, L. (2011). Fast image-based localization using direct 2D-to-3D matching. In *ICCV*.

Sattler, T., Leibe, B., & Kobbelt, L. (2016). Efficient and effective prioritized matching for large-scale image-based localization. *TPAMI, 39*(9), 1744–1756.

Sener, F., Chatterjee, D., Shelepov, D., He, K., Singhania, D., Wang, R., & Yao, A. (2022). Assembly101: A large-scale multi-view video dataset for understanding procedural activities. In *CVPR*.

Shah, A., Lundell, B., Sawhney, H., & Chellappa, R. (2023). STEPs: Self-supervised key step extraction and localization from unlabeled procedural videos. In *ICCV*.

Shan, D., Geng, J., Shu, M., & Fouhey, D.F. (2020). Understanding human hands in contact at internet scale. In *CVPR*.

Shapovalov, R., Kleiman, Y., Rocco, I., Novotny, D., Vedaldi, A., Chen, C., Kokkinos, F., Graham, B., & Neverova, N. (2023). Replay: Multi-modal multi-view acted videos for casual holography. In *ICCV*.

Sharghi, A., Gong, B., & Shah, M. (2016). Query-focused extractive video summarization. In *ECCV*.

Sharghi, A., Laurel, J.S., & Gong, B. (2017). Query-focused video summarization: Dataset, evaluation, and a memory network based approach. In *CVPR*.

Shavit, Y., Ferens, R., & Keller, Y. (2021). Learning multi-scene absolute pose regression with transformers. In *ICCV*.

Shaw, K., Bahl, S., & Pathak, D. (2022). VideoDex: Learning dexterity from internet videos. In *CoRL*.

Shiraga, K., Trung, N. T., Mitsugami, I., Mukaigawa, Y., & Yagi, Y. (2012). Gait-based person authentication by wearable cameras. In *International conference on networked sensing systems*.

Shiratori, T., Park, H. S., Sigal, L., Sheikh, Y., & Hodgins, J. K. (2011). Motion capture from body-mounted cameras. *Transactions on Graphics, 30*(4), 1–10.

Sigurdsson, G.A., Gupta, A., Schmid, C., Farhadi, A., & Alahari, K. (2018). Actor and observer: Joint modeling of first and third-person videos. In *CVPR*.

Silva, M., Ramos, W., Ferreira, J., Chamone, F., Campos, M., & Nascimento, E.R. (2018). A weighted sparse sampling and smoothing frame transition approach for semantic fast-forward first-person videos. In *CVPR*.

Singh, S., Arora, C., & Jawahar, C. (2016b). First person action recognition using deep learned descriptors. In *CVPR*.

Singh, K.K., Fatahalian, K., & Efros, A.A. (2016a). KrishnaCam: Using a longitudinal, single-person, egocentric dataset for scene understanding tasks. In *WACV*.

Smith, B.A., Yin, Q., Feiner, S.K., & Nayar, S.K. (2013). Gaze locking: Passive eye contact detection for human-object interaction. In *Symposium on user interface software and technology*.

Song, S., Chandrasekhar, V., Mandal, B., Li, L., Lim, J.H., Babu, G.S., San, P.P., & Cheung, N.M. (2016). Multimodal multi-stream deep learning for egocentric activity recognition. In *CVPR workshop*.

Song, Y., Vallmitjana, J., Stent, A., & Jaimes, A. (2015). TVSum: Summarizing web videos using titles. In *CVPR*.

Speciale, P., Schönberger, J.L., Kang, S.B., Sinha, S.N., & Pollefeys, M. (2019). Privacy preserving image-based localization. In *CVPR*.

Spriggs, E.H., De La Torre, F., & Hebert, M. (2009). Temporal segmentation and activity classification from first-person sensing. In *CVPR workshop*.

Starner, T., Schiele, B., & Pentland, A. (1998) Visual contextual awareness in wearable computing. In *International symposium on wearable computers*.

Steil, J., Koelle, M., Heuten, W., Boll, S., & Bulling, A. (2019). PrivacEye: Privacy-preserving head-mounted eye tracking using egocentric scene image and eye movement features. In *Symposium on eye tracking research and applications*.

Su, Y.C., & Grauman, K. (2016). Detecting engagement in egocentric video. In *ECCV*.

Su, S., Hong, J.P., Shi, J., & Park, H.S. (2016). Social behavior prediction from first person videos. arXiv:1611.09464

Sudhakaran, S., & Lanz, O. (2017). Convolutional long short-term memory networks for recognizing first person interactions. In *ICCV workshop*.

Sudhakaran, S., & Lanz, O. (2018). Attention is all we need: Nailing down object-centric attention for egocentric activity recognition. In *BMVC*.

Sudhakaran, S., Escalera, S., & Lanz, O. (2019). LSTA: Long short-term attention for egocentric action recognition. In *CVPR*.

Surie, D., Pederson, T., Lagriffoul, F., Janlert, L. E., & Sjölie, D. (2007). Activity recognition using an egocentric perspective of everyday objects. In *International conference on ubiquitous intelligence and computing*.

Surís, D., Epstein, D., Ji, H., Chang, S.F., & Vondrick, C. (2020). Learning to learn words from visual scenes. In *ECCV*.

Suveges, T., & McKenna, S. (2021). Egomap: Hierarchical first-person semantic mapping. In *ICPR*.

Tan, S., Nagarajan, T., & Grauman, K. (2023). EgoDistill: Egocentric head motion distillation for efficient video understanding. In *NeurIPS*.

Tang, H., Liang, K., Grauman, K., Feiszli, M., & Wang, W. (2023a) EgoTracks: A long-term egocentric visual object tracking dataset. In *NeurIPS*.

Tang, Z., Qiu, Z., Hao, Y., Hong, R., & Yao, T. (2023b). 3D human pose estimation with spatio-temporal criss-cross attention. In *CVPR*.

Tang, Y., Tian, Y., Lu, J., Feng, J., & Zhou, J. (2017). Action recognition in RGB-D egocentric videos. In *ICIP*.

Tang, D., Yu, T.H., & Kim, T.K. (2013). Real-time articulated hand pose estimation using semi-supervised transductive regression forests. In *ICCV*.

Tango, K., Ohkawa, T., Furuta, R., & Sato, Y. (2022). Background mixup data augmentation for hand and object-in-contact detection. In *ECCV workshop*.

Tavakoli, H.R., Rahtu, E., Kannala, J., & Borji, A. (2019). Digging deeper into egocentric gaze prediction. In *WACV*.

Tekin, B., Bogo, F., & Pollefeys, M. (2019). H+O: Unified egocentric recognition of 3D hand-object poses and interactions. In *CVPR*.

Templeman, R., Korayem, M., Crandall, D.J., & Kapadia, A. (2014). PlaceAvoider: Steering first-person cameras away from sensitive spaces. In *Network and distributed system security symposium*.

Templeman, R., Rahman, Z., Crandall, D.J., & Kapadia, A. (2012). PlaceRaider: Virtual theft in physical spaces with smartphones. arXiv:1209.5982.

Tendulkar, P., Surís, D., & Vondrick, C. (2023). FLEX: Full-body grasping without full-body grasps. In *CVPR*.

Thakur, S. K., Beyan, C., Morerio, P., & Del Bue, A. (2021). Predicting gaze from egocentric social interaction videos and IMU data. In *International conference on multimodal interaction*.

Thapar, D., Arora, C., & Nigam, A. (2020a). Is sharing of egocentric video giving away your biometric signature? In Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M. (Eds.), *ECCV*.

Thapar, D., Nigam, A., & Arora, C. (2020b). Recognizing camera wearer from hand gestures in egocentric videos. In *International conference on multimedia*.

Thapar, D., Nigam, A., & Arora, C. (2021). Anonymizing egocentric videos. In *ICCV*.

Thomaz, E., Parnami, A., Bidwell, J., Essa, I., & Abowd, G.D. (2013). Technological approaches for addressing privacy concerns when recognizing eating behaviors with wearable cameras. In *International joint conference on pervasive and ubiquitous computing*.

Tokmakov, P., Li, J., & Gaidon, A. (2023). Breaking the "Object" in video object segmentation. In *CVPR*.

Tome, D., Peluse, P., Agapito, L., & Badino, H. (2019). xR-EgoPose: Egocentric 3D human pose from an HMD camera. In *ICCV*.

Torralba, A., Murphy, K.P., Freeman, W.T., & Rubin, M.A. (2003). Context-based vision system for place and object recognition. In *ICCV*.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., & Bikel, D. (2023). Llama 2: Open foundation and fine-tuned chat models. arXiv:2307.09288

Tschernezki, V., Darkhalil, A., Zhu, Z., Fouhey, D., Larina, I., Larlus, D., Damen, D., & Vedaldi, A. (2023). EPIC fields: Marrying 3D geometry and video understanding. In *NeurIPS*.

Tschernezki, V., Larlus, D., & Vedaldi, A. (2021). NeuralDiff: Segmenting 3D objects that move in egocentric videos. In *3DV*.

Tse, T.H.E., Mueller, F., Shen, Z., Tang, D., Beeler, T., Dou, M., Zhang, Y., Petrovic, S., Chang, H.J., Taylor, J., & Doosti, B. (2023). Spectral graphormer: Spectral graph-based transformer for egocentric two-hand reconstruction using multi-view color images. In *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)* (pp. 14666–14677).

Tsutsui, S., Fu, Y., & Crandall, D. J. (2021). Whose hand is this? Person identification from egocentric hand gestures. In *WACV*.

Vahdani, E., & Tian, Y. (2023). Deep learning-based action detection in untrimmed videos: A survey. *TPAMI, 45*(4), 4302–4320.

Vakunov, A., Chang, C.L., Zhang, F., Sung, G., Grundmann, M., & Bazarevsky, V. (2020). MediaPipe hands: On-device real-time hand tracking. In *CVPR workshop*.

Varini, P., Serra, G., & Cucchiara, R. (2017). Personalized egocentric video summarization of cultural tour on user preferences input. *Transactions on Multimedia, 19*(12), 2832–2845.

Verma, S., Nagar, P., Gupta, D., & Arora, C. (2018). Making third person techniques recognize first-person actions in egocentric videos. In *ICIP*.

Vondrick, C., Pirsiavash, H., & Torralba, A. (2016). Anticipating visual representations from unlabeled video. In *CVPR*.

Wang, X., Cheng, F., Bertasius, G., & Crandall, D.J. (2023d). LoCoNet: Long-short context network for active speaker detection. arXiv:2301.08237

Wang, X., Kwon, T., Rad, M., Pan, B., Chakraborty, I., Andrist, S., Bohus, D., Feniello, A., Tekin, B., Frujeri, F.V., & Joshi, N. (2023e). HoloAssist: An egocentric human interaction dataset for interactive AI assistants in the real world. In *ICCV*.

Wang, J., Liu, L., Xu, W., Sarkar, K., & Theobalt, C. (2021a). Estimating egocentric 3D human pose in global space. In *ICCV*.

Wang, J., Liu, L., Xu, W., Sarkar, K., Luvizon, D., & Theobalt, C. (2022). Estimating egocentric 3D human pose in the wild with external weak supervision. In *CVPR*.

Wang, J., Luvizon, D., Xu, W., Liu, L., Sarkar, K., & Theobalt, C. (2023b). Scene-aware egocentric 3D human pose estimation. In *CVPR*.

Wang, H., Singh, M.K., & Torresani, L. (2023a). Ego-only: Egocentric action detection without exocentric transferring. In *ICCV*.

Wang, J., Wang, X., Shang-Guan, Y., & Gupta, A.K. (2021b). Wanderlust: Online continual object detection in the real world. In *ICCV*.

Wang, X., Zhao, X., Prakash, V., Shi, W., & Gnawali, O. (2013). Computerized-eyewear based face recognition system for improving social lives of prosopagnosics. In *International conference on pervasive computing technologies for healthcare*.

Wang, Q., Zhao, L., Yuan, L., Liu, T., & Peng, X. (2023c). Learning from semantic alignment between unpaired multiviews for egocentric video recognition. In *ICCV*.

Wen, H., Liu, Y., Huang, J., Duan, B., & Yi, L. (2022). Point primitive transformer for long-term 4D point cloud video understanding. In *ECCV*.

Wen, Y., Pan, H., Yang, L., Pan, J., Komura, T., & Wang, W. (2023a). Hierarchical temporal transformer for 3d hand pose estimation and action recognition from egocentric RGB videos. In *CVPR*.

Wen, Y., Tang, Z., Pang, Y., Ding, B., & Liu, M. (2023b). Interactive spatiotemporal token attention network for skeleton-based general interactive action recognition. In *IROS*.

Wieczorek, M., Rychalska, B., & Dąbrowski, J. (2021). On the unreasonable effectiveness of centroids in image retrieval. In *NeurIPS*.

Wijmans, E., Datta, S., Maksymets, O., Das, A., Gkioxari, G., Lee, S., Essa, I., Parikh, D., & Batra, D. (2019). Embodied question answering in photorealistic environments with point cloud perception. In *CVPR*.

Wolf, W. (1996) Key frame selection by motion analysis. In *ICASSP*.

Wong, B., Chen, J., Wu, Y., Lei, S.W., Mao, D., Gao, D., & Shou, M.Z. (2022). AssistQ: Affordance-centric question-driven task completion for egocentric assistant. In *ECCV*.

Wu, C.Y., Li, Y., Mangalam, K., Fan, H., Xiong, B., Malik, J., & Feichtenhofer, C. (2022a). MeMViT: Memory-augmented multi-scale vision transformer for efficient long-term video recognition. In *CVPR*.

Wu, G., Lin, J., & Silva, C.T. (2022b). IntentVizor: Towards generic query guided interactive video summarization. In *CVPR*.

Wu, J.Z., Zhang, D.J., Hsu, W., Zhang, M., & Shou, M.Z. (2023). Label-efficient online continual object detection in streaming video. In *ICCV*.

Xia, L., Gori, I., Aggarwal, J.K., & Ryoo, M.S. (2015). Robot-centric activity recognition from first-person RGB-D videos. In *WACV*.

Xiong, B., & Grauman, K. (2014). Detecting snap points in egocentric video with a web photo prior. In *ECCV*.

Xiong, X., Arnab, A., Nagrani, A., & Schmid, C. (2022). M&M Mix: A multimodal multiview transformer ensemble. arXiv:2206.09852

Xiong, B., Kim, G., & Sigal, L. (2015). Storyline representation of egocentric videos with an applications to story-based search. In *ICCV*.

Xu, M., Li, Y., Fu, C. Y., Ghanem, B., Xiang, T., & Pérez-Rúa, J. M. (2023). Where is my Wallet? Modeling object proposal sets for egocentric visual query localization. In *CVPR*.

Xu, J., Mukherjee, L., Li, Y., Warner, J., Rehg, J.M., & Singh, V. (2015). Gaze-enabled egocentric video summarization via constrained submodular maximization. In *CVPR*.

Xu, W., Yu, Z., Xue, H., Ye, R., Yao, S., & Lu, C. (2023b). Visual-tactile sensing for in-hand object reconstruction. In *CVPR*.

Xu, W., Chatterjee, A., Zollhoefer, M., Rhodin, H., Fua, P., Seidel, H. P., & Theobalt, C. (2019). Mo2Cap2: Real-time mobile 3D motion capture with a cap-mounted fisheye camera. *Transactions on Visualization and Computer Graphics, 25*(5), 2093–2101.

Xue, Z., & Grauman, K. (2023). Learning fine-grained view-invariant representations from unpaired ego-exo videos via temporal alignment. In *NeurIPS*.

Xue, Z., Song, Y., Grauman, K., & Torresani, L. (2023). Egocentric video task translation. In *CVPR*.

Yagi, T., Mangalam, K., Yonetani, R., & Sato, Y. (2018). Future person localization in first-person videos. In *CVPR*.

Yamada, K., Sugano, Y., Okabe, T., Sato, Y., Sugimoto, A., & Hiraki, K. (2011). Can saliency map models predict human egocentric visual attention? In *ACCV Workshop*.

Yamada, K., Sugano, Y., Okabe, T., Sato, Y., Sugimoto, A., & Hiraki, K. (2012). Attention prediction in egocentric video using motion and visual saliency. In *Pacific-rim symposium on image and video technology*.

Yan, S., Xiong, X., Arnab, A., Lu, Z., Zhang, M., Sun, C., & Schmid, C. (2022). Multiview transformers for video recognition. In *CVPR*.

Yang, J., Bhalgat, Y., Chang, S., Porikli, F., & Kwak, N. (2022). Dynamic iterative refinement for efficient 3D hand pose estimation. In *WACV*.

Yang, X., Chu, F.J., Feiszli, M., Goyal, R., Torresani, L., & Tran, D. (2023). Relational space-time query in long-form videos. In *CVPR*.

Yang, M., Du, Y., Ghasemipour, K., Tompson, J., Schuurmans, D., & Abbeel, P. (2024). Learning interactive real-world simulators. In *ICLR*.

Yang, J.A., Lee, C.H., Yang, S.W., Somayazulu, V.S., Chen, Y.K., & Chien, S.Y. (2016). Wearable social camera: Egocentric video summarization for social interaction. In *International conference on multimedia and expo workshop*.

Yao, T., Mei, T., & Rui, Y. (2016). Highlight detection with pairwise deep ranking for first-person video summarization. In *CVPR*.

Ye, Z., Li, Y., Fathi, A., Han, Y., Rozga, A., Abowd, G.D., & Rehg, J.M. (2012). Detecting eye contact using wearable eye-tracking glasses. In *International joint conference on pervasive and ubiquitous computing*.

Ye, Y., Li, X., Gupta, A., Mello, S.D., Birchfield, S., Song, J., Tulsiani, S., & Liu, S. (2023b). Affordance diffusion: Synthesizing hand-object interactions. In *CVPR*.

Ye, Z., Li, Y., Liu, Y., Bridges, C., Rozga, A., & Rehg, J. M. (2015). Detecting bids for eye contact using a wearable camera. In *International conference and workshops on automatic face and gesture recognition*.

Ye, V., Pavlakos, G., Malik, J., & Kanazawa, A. (2023a). Decoupling human and camera motion from videos in the wild. In *CVPR*.

Yonetani, R., Kitani, K.M., & Sato, Y. (2015). Ego-surfing first person videos. In *CVPR*.

Yonetani, R., Kitani, K.M., & Sato, Y. (2016). Recognizing micro-actions and reactions from paired egocentric videos. In *CVPR*.

Yu, L., Chen, X., Gkioxari, G., Bansal, M., Berg, T.L., & Batra, D. (2019). Multi-target embodied question answering. In *CVPR*.

Yu, J., Li, X., Zhao, X., Zhang, H., & Wang, Y.X. (2023). Video state-changing object segmentation. In *ICCV*.

Yu, X., Rao, Y., Zhao, W., Lu, J., & Zhou, J. (2021). Group-aware contrastive regression for action quality assessment. In *ICCV*.

Yu, F., Seff, A., Zhang, Y., Song, S., Funkhouser, T., & Xiao, J. (2015). LSUN: Construction of a large-scale image dataset using deep learning with humans in the loop. arXiv:1506.03365

Yuan, Y., & Kitani, K. (2018). 3D ego-pose estimation via imitation learning. In *ECCV*.

Yuan, Y., & Kitani, K. (2019). Ego-pose estimation and forecasting as real-time PD control. In *ICCV*.

Yuan, S., Ye, Q., Stenger, B., Jain, S., & Kim, T.K. (2017). Big-Hand2.2M benchmark: Hand pose dataset and state of the art analysis. In *CVPR*.

Zatsarynna, O., & Gall, J. (2023). Action anticipation with goal consistency. In *ICIP*.

Zhang, K., Chao, W.L., Sha, F., & Grauman, K. (2016). Video summarization with long short-term memory. In *ECCV*.

Zhang, Z., Dong, Y., Liu, Y., & Yi, L. (2023d). Complete-to-partial 4D distillation for self-supervised point cloud sequence representation learning. In *CVPR*.

Zhang, C., Gupta, A., & Zisserman, A. (2023a) Helping Hhnds: An object-aware ego-centric video recognition model. In *ICCV*.

Zhang, S., Ma, Q., Zhang, Y., Aliakbarian, S., Cosker, D., & Tang, S. (2023c). Probabilistic human mesh recovery in 3D scenes from egocentric views. In *ICCV*.

Zhang, S., Ma, Q., Zhang, Y., Qian, Z., Kwon, T., Pollefeys, M., Bogo, F., & Tang, S. (2022c). EgoBody: Human body shape and motion of interacting people from head-mounted devices. In *ECCV*.

Zhang, M., Teck Ma, K., Hwee Lim, J., Zhao, Q., & Feng, J. (2017). Deep future gaze: Gaze anticipation on egocentric videos using adversarial networks. In *CVPR*.

Zhang, G., Zhang, Y., Zhang, T., Li, B., & Pu, S. (2023b). PHA: Patch-wise high-frequency augmentation for transformer-based person Re-identification. In *CVPR*.

Zhang, L., Zhou, S., Stent, S., & Shi, J. (2022a). Fine-grained egocentric hand-object segmentation: Dataset, model, and applications. In *ECCV*.

Zhang, L., Zhou, S., Stent, S., & Shi, J. (2022b). Fine-grained egocentric hand-object segmentation: Dataset, model, and applications. In *ECCV*.

Zhang, Y., Cao, C., Cheng, J., & Lu, H. (2018). EgoGesture: A new dataset and benchmark for egocentric hand gesture recognition. *Transactions on Multimedia, 20*(5), 1038–1050.

Zhao, B., & Xing, E.P. (2014). Quasi real-time summarization for consumer videos. In *CVPR*.

Zhao, Y., Ma, H., Kong, S., & Fowlkes, C. (2023b). Instance tracking in 3D scenes from egocentric videos. arXiv:2312.04117.

Zhao, Y., Misra, I., Krähenbühl, P., & Girdhar, R. (2023c). Learning video representations from large language models. In *CVPR*.

Zhao, Q., Wang, S., Zhang, C., Fu, C., Do, M.Q., Agarwal, N., Lee, K., & Sun, C. (2023a). AntGPT: Can large language models help long-term action anticipation from videos? arXiv:2307.16368.

Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., & Tian, Q. (2015). Scalable person Re-identification: A benchmark. In *ICCV*.

Zheng, X., Wen, C., Xue, Z., & Wang, J. (2023b). HaMuCo: Hand pose estimation via multiview collaborative self-supervised learning. In *ICCV*.

Zheng, Y., Yang, Y., Mo, K., Li, J., Yu, T., Liu, Y., Liu, C.K., & Guibas, L.J. (2022). GIMO: Gaze-informed human motion prediction in context. In *ECCV*.

Zheng, J., Zheng, Q., Fang, L., Liu, Y., & Yi, L. (2023a). CAMS: CAnonicalized manipulation spaces for category-level functional hand-object manipulation synthesis. In *CVPR*.

Zhong, Z., Schneider, D., Voit, M., Stiefelhagen, R., & Beyerer, J. (2023). Anticipative feature fusion transformer for multi-modal action anticipation. In *WACV*.

Zhou, X., Arnab, A., Sun, C., & Schmid, C. (2023). How can objects help action recognition? In *CVPR*.

Zhu, K., Guo, H., Yan, T., Zhu, Y., Wang, J., & Tang, M. (2022). PASS: Part-aware self-supervised pre-training for person Re-identification. In *ECCV*.

Zhu, H., Kapoor, R., Min, S.Y., Han, W., Li, J., Geng, K., Neubig, G., Bisk, Y., Kembhavi, A., & Weihs, L. (2023b). EXCALIBUR: Encouraging and evaluating embodied exploration. In *CVPR*.

Zhu, C., Xiao, F., Alvarado, A., Babaei, Y., Hu, J., El-Mohri, H., Culatana, S., Sumbaly, R., & Yan, Z. (2023a). EgoObjects: A large-scale egocentric dataset for fine-grained object understanding. In *ICCV*.

Zhu, S., Yang, L., Chen, C., Shah, M., Shen, X., & Wang, H. (2023c). R^2 Former: Unified retrieval and reranking transformer for place recognition. In *CVPR*.