International Journal of Computer Vision (2026) 134:62
https://doi.org/10.1007/s11263-025-02608-y

Bridging Perspectives: A Survey on Cross-view Collaborative Intelligence with Egocentric-Exocentric Vision

Yuping He[1] · Yifei Huang[2] · Guo Chen[1] · Lidong Lu[1] · Baoqi Pei[3] · Jilan Xu[4] · Tong Lu[1] · Yoichi Sato[2]

Received: 6 June 2025 / Accepted: 12 November 2025

Abstract

Perceiving the world from both egocentric (first-person) and exocentric (third-person) perspectives is fundamental to human cognition, enabling rich and complementary understanding of dynamic environments. In recent years, allowing the machines to leverage the synergistic potential of these dual perspectives has emerged as a compelling research direction in video understanding. In this survey, we provide a comprehensive review of video understanding from both exocentric and egocentric viewpoints. We first ground our review in key application domains, from healthcare to embodied intelligence, to establish the practical value of ego-exo collaboration. From the needs of these applications, we derive a set of core research tasks. We then systematically organize recent advancements into three primary research directions: (1) leveraging egocentric data to enhance exocentric understanding, (2) utilizing exocentric data to improve egocentric analysis, and (3) joint learning frameworks that unify both perspectives. We also provide a detailed overview of relevant datasets and conclude by discussing current limitations and promising future directions. By synthesizing insights from both perspectives, our goal is to inspire advancements in video understanding and artificial intelligence, bringing machines closer to perceiving the world in a human-like manner.

Keywords Video understanding · Egocentric video · Exocentric video · Datasets and benchmarks.

1 Introduction

Perceiving the world from both egocentric (first-person) and exocentric (third-person) perspectives is a fundamental ability in human intelligence. The mirror neuron theory (Rizzolatti & Craighero, 2004) posits that the same neural mechanisms are activated when an individual performs an action and when they observe another performing the same action. This biological insight underscores the intrinsic connection between first- and third-person viewpoints, inspiring efforts to emulate this capability. By enabling machines to integrate and leverage information across these perspectives, we can advance video understanding and move closer to human-like perception.

The exocentric (third-person) and egocentric (first-person) perspectives offer complementary views of human activity, akin to two sides of the same coin. The egocentric view provides an actor-centered perspective (Sigurdsson et al., 2018), capturing rich human-object interactions and reflecting the wearer's intentions and goals (Damen, 2021; Grauman, 2024a; Song et al., 2024a). Unlike the exocentric view, egocentric videos are inherently more dynamic, featuring continuous motion and shifting backgrounds, which pose challenges such as partial visibility of the wearer (Khirodkar et al., 2023; Fan et al., 2017). Still, the release of large-scale egocentric datasets (Damen, 2021; Grauman, 2024a, b; Perrett et al., 2025) has spurred substantial progress in egocentric video understanding (Lin, 2022; Pramanick, 2023; Zhang et al., 2023; Chen et al., 2022; Pei et al., 2024, 2025).

In contrast, the exocentric view offers an observer-like perspective (Sigurdsson et al., 2018), providing a broader context of the scene. It also enables the capture of the subject's full-body pose (Choi et al., 2021; Li et al., 2024; Kocabas et al., 2021), which is difficult to obtain from ego-

Communicated by Dima Damen.

Y. He, Y. Huang, and G. Chen have equal contributions

✉ Yifei Huang
hyf015@gmail.com

[1] State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023, China

[2] University of Tokyo, Tokyo, Japan

[3] Zhejiang University, Zhejiang 310027, China

[4] The University of Oxford, Oxford, UK

Published online: 13 January 2026

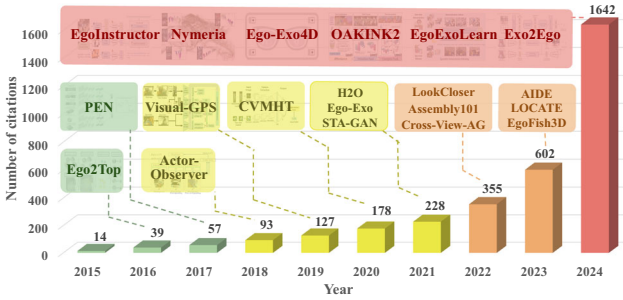

Fig. 1 Number of citations to egocentric-exocentric related papers from 2015 to 2024. Citation data was collected from Google Scholar. The statistics are computed based on papers and datasets discussed in Sections 4 and 5, all of which utilize both egocentric and exocentric perspectives.

centric recordings. Different from egocentric videos, these videos are usually recorded from a stable, fixed position, covering a wide field of view and capturing detailed scene context. These videos can be easily captured using devices such as smartphones and surveillance cameras, and their widespread availability on the Internet has led to the creation of diverse large-scale datasets, for example, (Miech et al., 2019; Gu, 2018; Soomro et al., 2012; Carreira & Zisserman, 2017; Wang et al., 2023; Chen et al., 2024). These datasets have driven significant advancements in third-person video understanding (Simonyan & Zisserman, 2014; Carreira & Zisserman, 2017; Arnab et al., 2021; Bertasius et al., 2021; Chen et al., 2022; Chen, 2024; Wang et al., 2023).

While egocentric and exocentric perspectives have distinct characteristics, they are inherently complementary (Grauman, 2024b). The ego-view provides details from the actor's perspective, while the exo-view offers a broader contextual understanding of the scene. Researchers can unlock new opportunities to advance video understanding by integrating these perspectives. This synergy has led to a growing body of work exploring cross-view learning, as demonstrated in Fig. 1.

Despite these advancements, most surveys in video understanding (Liu, 2022; Stergiou & Poppe, 2024; Jiao, 2022; Ramachandra et al., 2022) focus on specific tasks and primarily concentrate on exocentric videos. In egocentric vision, Plizzari (2024) review advancements across multiple tasks. A recent survey by Thatipelli et al. (2025) represents an important step by providing an overview of combined egocentric and exocentric methods. Our work takes a step forward by introducing a systematic taxonomy of existing research, structuring the field into three primary directions: (1) leveraging egocentric data to enhance exocentric understanding, (2) utilizing exocentric data to improve egocentric analysis, and (3) joint learning frameworks for cross-view video understanding.

The overall structure of this survey is illustrated in Fig. 2. Inspired by Plizzari (2024), we also adopt a "future-to-present" approach. Specifically, we start by highlighting the transformative potential of integrating egocentric and exocentric perspectives (Grauman, 2024b), demonstrating how cross-view collaboration can benefit various domains (Section 2). We then identify key research tasks to realize these applications (Section 3). In addition to the systematic review of existing research works (Section 4), we also analyze benchmark datasets that support both perspectives (Section 5), evaluate their diversity and applicability. Finally, we discuss the limitations of current approaches and propose promising research directions (Section 6).

2 Applications

In this section, we highlight the practical value of integrating egocentric and exocentric video understanding. We explore eight representative application scenarios that have a significant demand for ego-exo collaboration. For each scenario, we first provide examples of how egocentric and exocentric techniques are already applied in real-world systems. Following this, we delve into the domain-specific advantages of each viewpoint, analyzing the complementary properties that make their integration particularly valuable. Since most current systems are limited to a single view or are in the early stages of collaboration, we conclude each discussion by exploring a near-term synergistic application, as demonstrated in Fig. 3.

2.1 Cooking

Vision-based kitchen assistants have recently emerged, with systems like the Samsung Family Hub refrigerator (Lavars, 2016) and the June (Oven, 2018) using exocentric cameras for food recognition and task-specific automation. However, these systems are limited in scope and lack holistic cooking support.

These two perspectives offer complementary information that is highly relevant to cooking tasks. The exocentric view offers a global understanding of the kitchen workspace, including the spatial layout of tools and ingredients, the user's overall workflow. This context is crucial for tracking the sequence of cooking procedures. In contrast, the egocentric view delivers a fine-grained account of hand–object interactions, such as knife handling or the exact texture of an ingredient during preparation. This close-up perspective captures indispensable details for assessing the precision and quality of each individual step.

Looking forward, a feasible near-term application can be envisioned in the form of a cooking training assistant. This builds upon existing vision-based kitchen assistants and

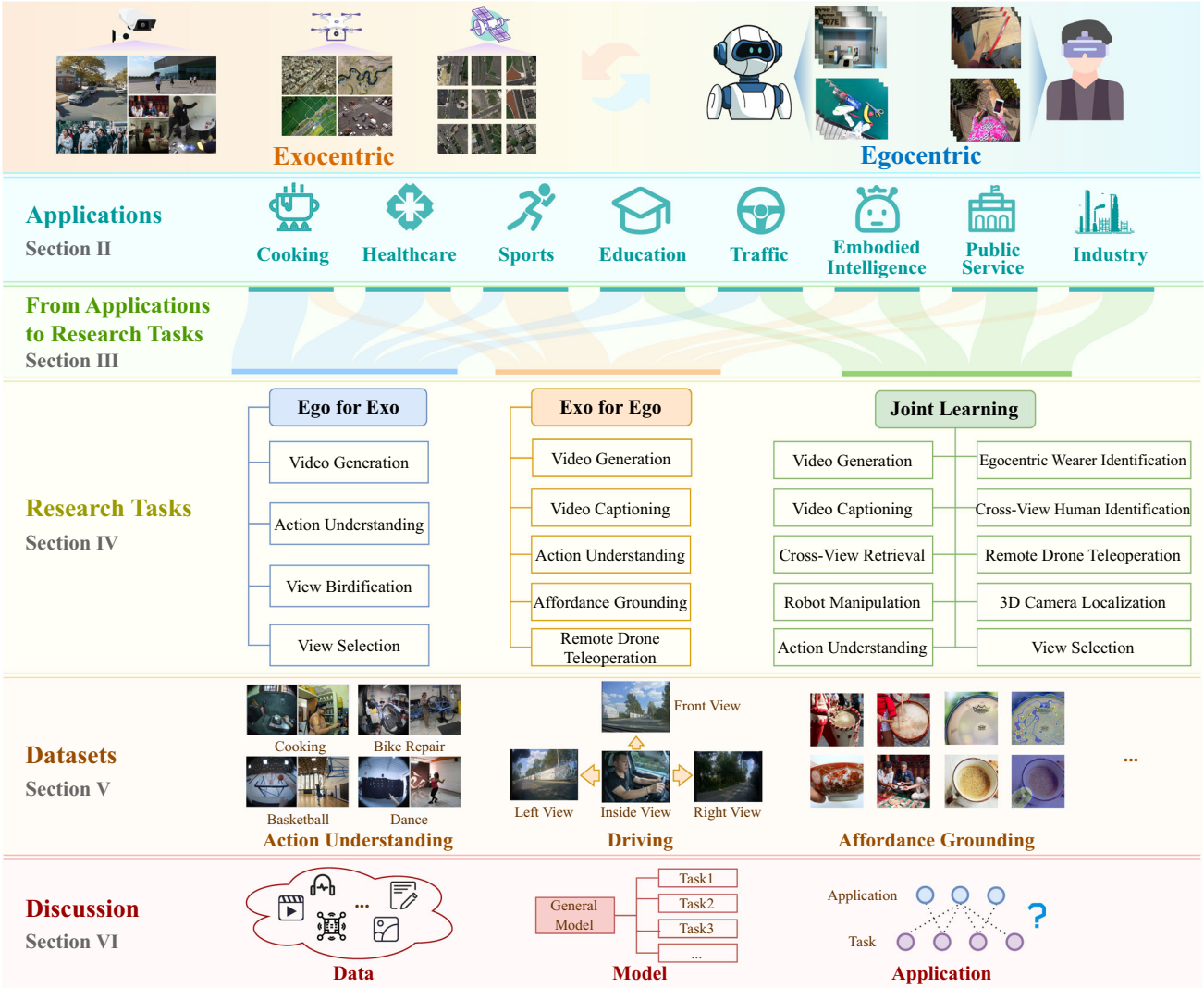

Fig. 2 The overall structure of the survey. We first highlight the application value of egocentric and exocentric collaboration (Section 2). We then identify critical research tasks for each application (Section 3). Next, we provide a comprehensive overview of the current research advancements (Section 4). This section is divided into: ego for exo, exo for exo, and joint learning, each covering various research tasks. Additionally, we examine datasets that encompass both perspectives (Section 5). Finally, we discuss limitations and future directions (Section 6).

the development of smart wearable devices (e.g., Microsoft HoloLens (Microsoft, 2025), Apple Vision Pro (Apple, 2023)). This assistant would integrate both perspectives to provide personalized guidance. For example, by leveraging past egocentric recordings, the system can recommend appropriate exocentric demonstration videos tailored to a learner's proficiency. During cooking, exocentric cameras can capture spatial trajectories and overall task progression (e.g., detecting that the user has moved to the countertop and is performing the "placing items into the bowl" step), while egocentric cameras capture fine-grained execution details (e.g., identifying the available ingredients). By correlating high-level procedural stages with low-level manipulation details, the system can deliver real-time, targeted feedback, such as reminding the user of missing ingredients.

2.2 Sports

Exocentric vision systems currently dominate sports analysis, with applications such as sports tracking systems (McDonald , 2013) and referee assistance system (Winter, 2024). For broadcasting, Fox Sports' "Be The Player" (Dachman, 2017) generates egocentric replays from exocentric views. However, as wearing cameras can hinder players' movements, the use of egocentric perspectives and multi-view collaboration remains limited.

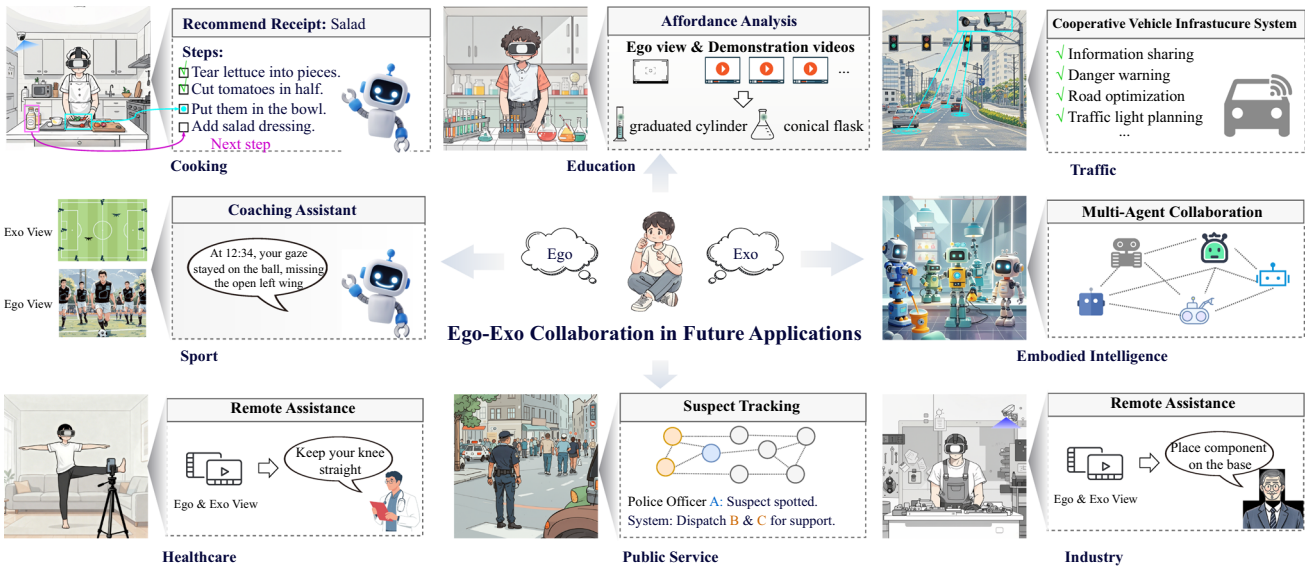

Fig. 3 Examples of the potential collaboration of egocentric and exocentric vision in diverse applications. We illustrate how integrating egocentric and exocentric video understanding techniques can enhance these applications.

In sports, the two perspectives demonstrate clear complementarity. The exocentric view provides a direct visualization of team-level dynamics, including trajectories, formations, and spatial distributions. Combined with computer vision techniques, it can also yield statistics such as running speed and body posture. By contrast, the egocentric view offers a subjective window into the athlete's perceptual–cognitive process. It reveals what the athlete is actually perceiving and gazing. This information provides a valuable source for analyzing their decision-making process.

Looking forward, an important frontier in sports analytics lies in the integration of field-wide multi-camera infrastructures (exocentric) with increasingly advanced wearable egocentric sensors. A promising near-future application is an intelligent coaching assistant that connects athletes' physical actions with their perceptual–cognitive processes. For example, during a striker's drill, exocentric cameras could capture an open passing lane, while egocentric footage shows the striker focusing only on the ball. The system could then highlight the missed opportunity that even coaches might overlook. This ability to semantically align macro-level tactical events with micro-level attentional patterns provides personalized feedback.

2.3 Healthcare

Currently, both egocentric and exocentric videos are widely used in healthcare settings. Exocentric cameras are extensively deployed in hospitals, allowing real-time monitoring of patients' health conditions. In remote assistance (Asia, 2022a), on-site doctors use egocentric cameras to transmit patients' conditions to remote experts for real-time guidance. In emergency services (Asia, 2022b), body cameras worn by ambulance workers transmit live video to doctors, aiding in patient preparation. However, current applications often rely on a single view, lacking multi-view collaboration.

In this domain, egocentric and exocentric views complement each other. Exocentric cameras are typically fixed in hospitals or rehabilitation settings, providing continuous and stable coverage of the patient's overall environment. This allows the monitoring of movement trajectories and general activity patterns over time. In surgical contexts, exocentric recording systems capture the entire procedure, supporting postoperative assessment. In contrast, mobile egocentric cameras can follow the perspective of a doctor, nurse, or patient, delivering first-hand information. During surgery or complex care tasks, they provide a direct view of the operator's fine-grained manipulation. In rehabilitation scenarios, egocentric vision can capture detailed execution of exercises, such as subtle joint angles, that may be unclear from a distant third-person view.

Looking ahead, a near-future application is a rehabilitation assistant that supports remote therapists in providing personalized care. Patients' training can be recorded using a smartphone as an exocentric camera, while wearable devices capture egocentric views of hand or limb movements. The exocentric video allows therapists to assess overall posture and exercise completion, whereas the egocentric view provides fine-grained insights into detailed movements. For instance, the system may reveal that the leg posture is generally correct (exo) but the knee of the supporting foot is slightly bent and requires adjustment (ego). By integrating these multi-view observations, therapists can adapt exercise

sequences or difficulty levels, ensuring that patients follow the rehabilitation program safely and effectively.

2.4 Education

Nowadays, cameras are widely installed on classroom ceilings to track student movements and enhance safety (Rahman, 2024). Additionally, class recording systems capture lectures, supporting both review sessions and online learning (Reolink, 2024; Huang et al., 2025). However, these systems currently operate as passive recording tools, lacking the ability to actively contribute to teaching activities.

Data from the two perspectives can contribute in complementary ways. Exocentric views provide a comprehensive overview of the classroom, capturing group dynamics and teacher–student interactions. This information can support teachers in reviewing overall class performance after the session. In contrast, egocentric recordings from students reveal individual learning activities, such as attentional focus or signs of distraction, enabling more personalized feedback. Moreover, online third-person demonstration videos, which are increasingly used as teaching resources, can be further integrated with these classroom recordings to enrich the learning experience.

Building upon large-scale online third-person demonstration videos and AR techniques, a feasible near-future application is an interactive laboratory assistant. A key challenge for students lies in translating abstract third-person instructions into concrete first-person actions. The assistant would address this by leveraging egocentric AR glasses to recognize the specific instrument a student is currently focusing on at the lab bench. It would then perform a cross-view association, linking the real-world object from the student's perspective with its corresponding entry in a database of third-person expert demonstrations. Finally, the system would overlay intuitive, actionable guidance directly within the student's field of view. Such a system could offer students an immersive learning experience and deliver tailored guidance.

2.5 Traffic

In the traffic domain, the integration of egocentric in-vehicle cameras with exocentric roadside monitoring systems is beginning to take shape. Several automotive companies have already deployed vehicle-to-infrastructure solutions (Allsup, 2023; Musulin, 2019), where in-car devices communicate with roadside units to exchange information. These systems support practical functions such as traffic light visualization, green-wave guidance, and hazard warnings, allowing drivers to anticipate changes in road conditions more effectively.

Within these solutions, egocentric and exocentric data serve indispensable yet complementary roles. Egocentric data from onboard cameras provide each vehicle with exclusive, real-time observations of its immediate surroundings. They capture fine-grained details, such as a car suddenly braking or a pedestrian stepping into the road, which roadside cameras may miss due to distance or occlusion. In contrast, exocentric roadside cameras offer wide-area coverage of intersections and highways. They can monitor overall traffic flows, identify congestion patterns, and detect hazards beyond the field of view of individual vehicles.

Building on existing solutions, a promising near-term direction is to achieve deeper coordination among private vehicles, public transport, and roadside infrastructure. For example, in emergency scenarios involving ambulances, today's traffic management still relies heavily on manual coordination by police officers. A more integrated system could automatically combine the ambulance's egocentric camera feed with roadside surveillance, dynamically clearing routes and providing real-time guidance to other vehicles. Such coordination illustrates the practical value of linking egocentric and exocentric perspectives for safer and more efficient urban mobility.

2.6 Embodied Intelligence

Currently, both egocentric and exocentric data are actively used across a range of real-world embodied intelligence systems. Modern robots leverage both egocentric and exocentric vision for applications like space exploration, medical assistance, customer service, and security (harkiran78, 2024). Autonomous wheelchairs rely on onboard cameras to perceive their immediate surroundings (Hellogard, 2025). Similarly, AR headsets are employed by frontline workers to support assembly or training tasks [53], (Gupta, 2024).

In embodied intelligence, exocentric data are usually collected from external sensors, such as fixed surveillance cameras or the viewpoint of one robot observing another. Exocentric views provide a stable, global perspective that is valuable for situational awareness in cluttered environments, where an agent's onboard sensors may be occluded or limited in range. They enable reliable tracking of multiple agents and long-term monitoring of spatial layouts, which are essential for coordination and navigation. By contrast, egocentric data are typically captured by cameras mounted directly on the agent or user's own device, such as a robot's onboard camera or AR/VR headsets (Microsoft, 2025; Apple, 2023). This perspective captures the fine-grained details of manipulation and the agent's actual line of sight. Such first-person cues are particularly important for understanding intention and action execution, which cannot be inferred reliably from exocentric views alone.

With the advancement of multi-robot localization and mapping technologies, multi-robot cooperative systems have begun to see preliminary applications in tasks such as inspec-

tion and mine surveying (Wang et al., 2025). Building on these early developments, we anticipate that, in the near future, such systems will progressively extend to operate in large-scale complex environments. For instance, in disaster response scenarios, the environment is often continuously changing and more complex than controlled settings. In these contexts, the collaboration of egocentric and exocentric perspectives can play a critical role. Specifically, frontline robots can leverage egocentric cameras to capture local details of the disaster site, including the identification of survivors and assessment of immediate hazards, while aerial drones or fixed cameras providing exocentric views monitor the positions of multiple robots and overall environment. By integrating information from both perspectives, the command system can maintain an up-to-date map of the disaster area and optimize operational planning.

2.7 Public Service

Egocentric and exocentric videos are increasingly applied in public service domain. Surveillance cameras aid in locating criminals and missing persons, while body-worn cameras capture on-site scenes for law enforcement [56]. In search and rescue, aerial drone footage complements ground-level views for timely response (Singh, 2023). However, these systems typically operate in isolation, limiting their effectiveness.

In this domain, data from different perspectives offer complementary advantages. Fixed exocentric surveillance networks form the backbone of urban monitoring. Cameras installed on streets or mounted on aerial drones provide stable, continuous coverage of large areas, making them indispensable for tracking population movements or detecting suspicious activities. In contrast, egocentric views capture mobile, context-rich information directly from the perspective of personnel on the scene. This immediacy is particularly valuable for documenting evidence and supporting decision-making in rapidly evolving events.

Despite their complementary roles, current body-worn devices mainly upload footage to isolated systems, but remain disconnected from urban surveillance networks. This gap highlights the need for a unified framework that integrates these perspectives. Building on existing systems, the next step is to establish synergistic links between egocentric footage and exocentric surveillance. For instance, during a suspect pursuit, the system could automatically associate an officer's body-worn recording with nearby street cameras, enabling cross-view tracking of both the suspect and the officer. Such integration would aid command center to dispatch forces accordingly and provide real-time tactical guidance.

2.8 Industry

In modern manufacturing, ceiling-mounted cameras are widely employed for safety monitoring (Fogsphere, 2023). On automated assembly lines, cameras on robotic arms help precisely locate and assemble parts (Begg, 2024). During quality inspection, multi-view scans accurately identify product defects (Jobit, 2024). However, current industrial vision systems operate largely in isolated viewpoints, limiting their ability to provide comprehensive process monitoring.

First-person data captured from workers or devices provide close-up details that fixed cameras cannot capture, such as the precise positioning of tools or the status of partially assembled components. This information enables a more accurate analysis of each worker's actions and state. In contrast, exocentric cameras, widely distributed across the factory floor, offer a global perspective of the workspace. From elevated viewpoints, they capture the positions and movements of all workers, robotic arms, and equipment, allowing automated systems to monitor the overall production process. By integrating these two perspectives, it becomes possible to simultaneously track fine-grained local actions and global workspace dynamics, thereby enhancing task coordination and resource management across the production line.

We envision a feasible, near-term application of ego-exo collaboration to enhance current remote expert systems. Existing solutions primarily allow frontline workers to transmit first-person video to remote experts. However, experts can only observe the local view captured by the worker, lacking a global understanding of the surrounding environment. In complex industrial settings, this limited perspective may result in safety blind spots when providing guidance. Building on the existing factory surveillance network, we propose that future remote assistance systems could simultaneously incorporate video streams from multiple third-person cameras deployed in the workshop, especially those closest to the operator. This approach enables the visualization of the worker's full-body movements and operational trajectory. By integrating both views, remote experts can deliver more precise instructions, thereby improving task efficiency and safety.

3 From Applications To Research Tasks

The previous section outlines how egocentric and exocentric perspectives can collaborate to enable a wide range of applications. However, realizing these envisioned applications requires addressing several fundamental research challenges. In this section, we identify key research tasks that demand egocentric-exocentric collaboration and review existing efforts that contribute to their development.

A foundational capability for almost all ego–exo collaboration is the ability to establish correspondence between the two viewpoints. This gives rise to a key research task: cross-view retrieval, which aims to identify semantically matching videos across perspectives. The importance of this task has been highlighted in training scenarios, such as cooking, where a user's first-person attempt can be related with corresponding third-person demonstrations. It is also critical in domains like public service and traffic management, where footage from body-worn or onboard cameras could be linked with surveillance streams.

When corresponding videos from another viewpoint are not directly available, the need for video generation across perspectives arises. This task focuses on synthesizing the target viewpoint given observations from a different one. The need for this capability is particularly evident in instructional scenarios, observed in domains like industrial assembly and remote healthcare therapy. In these cases, third-person demonstrations could be transformed into first-person views to provide an immersive perspective.

Beyond visual matching or synthesis, a higher level of semantic abstraction is often required for analysis, documentation, and long-term recording. This need gives rise to the task of cross-view video captioning, which aims to generate a structured, textual description by jointly reasoning over both egocentric and exocentric video streams. Such a capability is demanded in sports analysis to generate comprehensive performance reports for coaches, and in education to create summaries of classroom interactivity.

Analyzing actions through cross-view understanding is of significant importance. Effectively integrating fine-grained motion cues from the first-person view with full-body posture from the third-person view is a key requirement in domains such as sports coaching, industrial remote assistance, and healthcare therapy. In complex manipulation scenarios, such as those encountered in cooking, or industrial assembly, affordance grounding emerges as a crucial task. This enables the use of large-scale third-person demonstrations to guide the execution of unfamiliar tools in the user's egocentric view. Identifying the egocentric camera wearer in the third-person view gives rise to the task of egocentric wearer identification, which is essential in scenarios where locating specific individuals is required. Beyond identifying the wearer, projecting the trajectory of target group observed in the egocentric view into a top-down view defines the task of view birdification, which is particularly relevant for coordinated tracking or monitoring in dynamic environments.

In environments populated with numerous cameras, a set of foundational tasks emerges related to managing the sensor network itself. One such task is best view selection, which focuses on identifying the most informative camera stream for a given purpose. This requirement becomes evident in industrial remote assistance, where the system selects the best exocentric view to capture a field worker, or in public service, where it is necessary to determine which surveillance camera can best track an officer. In contrast to a fixed viewpoint, remote drones provide a dynamic, on-demand third-person perspective. The need for such aerial views is apparent in multi-robot systems for maintaining team-wide situational awareness and in public service for rapid situational assessment at disaster sites. Beyond selecting a single best view, understanding the spatial relationship between all sensors gives rise to the task of multi-camera localization. This is critical in scenarios with multiple mobile agents, such as in multi-robot systems or large-scale factory floors. Complementing the localization of cameras is the task of cross-view human identification. This aims to identify and correlate individuals as they move between different viewpoints, a capability that is emphasized in applications like suspect tracking in the public service domain.

Having established the bridge from application-level needs to a set of core research tasks, we now turn to the mapping of existing academic efforts. As illustrated in Fig. 4, current studies can be positioned according to their targeted application domains and the research tasks they address. Collectively, these works provide a promising foundation for advancing ego–exo collaboration across diverse scenarios. However, current progress remains insufficient to meet the increasing demands of real-world deployment. It is also worth noting that, although potential opportunities in domains such as education and public service were highlighted in the previous section, we did not identify corresponding research that explicitly explores ego–exo collaboration in these areas. This absence further underscores the research gap and motivates future investigation. Building on this, the next section presents a detailed review of research tasks and their associated advancements, emphasizing the capabilities and limitations of existing research.

4 Research Tasks

The previous section introduces research progress tailored to specific applications. Building on this foundation, this section provides a comprehensive review of advancements in cross-view collaboration with both egocentric and exocentric perspectives. We organize the research directions into three categories, which are defined as follows:

1) **Exocentric for Egocentric:** This direction focuses on leveraging knowledge from the exocentric domain to enhance egocentric video understanding.
2) **Egocentric for Exocentric:** Inversely, this direction emphasizes utilizing knowledge from the egocentric domain to improve exocentric video understanding.

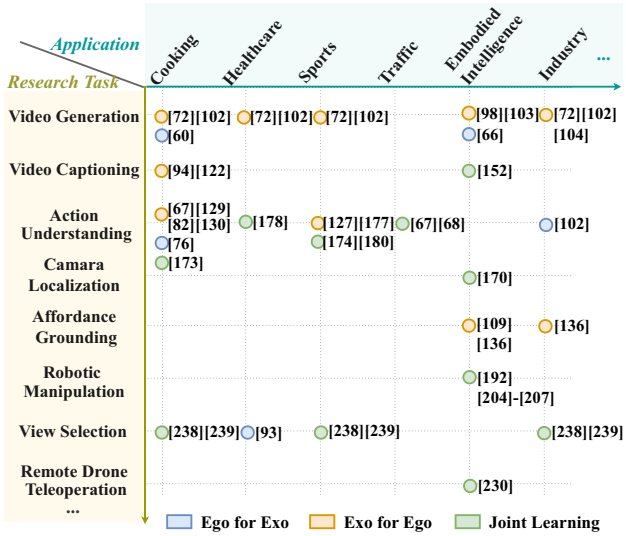

Fig. 4 Mapping relevant research works to applications and research tasks.

3) **Joint Learning:** This direction aims to integrate egocentric and exocentric perspectives to address cross-view video understanding tasks.

For each direction, we cover various research tasks and review the existing work. An overview is illustrated in Fig. 5.

4.1 Egocentric for Exocentric

The unique viewpoints of egocentric videos provide rich details that are often missing from exocentric perspectives. This subsection reviews research efforts that leverage egocentric perspectives to enhance exocentric tasks.

Video Generation. Ego-to-exo video generation involves generating an exocentric video from an egocentric one, offering a different perspective of the same environment. It offers significant research value across various fields. For instance, in virtual touring, travelers can review their routes from the third-person perspective to plan their trips effectively.

Video generation has made significant progress in recent years (Hu et al., 2023; Blattmann et al., 2023; Yang et al., 2024). However, ego-to-exo video generation poses unique challenges. Egocentric view often includes obscured regions, making it difficult to reconstruct the broader scene of the exocentric perspective. Additionally, maintaining consistency across views is challenging due to their significant disparity. Recent studies in video generation use depth maps (Esser et al., 2023), poses (Hu et al., 2023), and other conditional inputs (Blattmann et al., 2023; Yin et al., 2023) to provide spatial-temporal constraints. However, acquiring such cues in both egocentric and exocentric settings remains difficult.

For ego-to-exo video generation, IDE (Luo et al., 2024) introduces a novel framework that leverages human intention to maintain consistency across perspectives. It proposes that human intention is view-independent and can be used to establish connections between views. Specifically, it represents human intention through human movement and action descriptions, which serve as conditional inputs for the diffusion model, as illustrated in Fig. 6. Different from IDE (Luo et al., 2024), another work (Abdullah et al., 2024) investigates this task for underwater vehicles. Although onboard cameras provide a first-person view, this limited perspective restricts the operator's ability to maneuver in complex underwater environments. To address this, this approach uses past egocentric views and camera poses to create an eye-on-the-back view. This synthesis exocentric views provide broader scene context and enhance operational efficiency.

• Discussion: Despite prior efforts, ego-to-exo video generation introduces a set of challenges that are distinct from generic video synthesis tasks. First, egocentric views often have a limited field of view and are dominated by hand-object interactions (Plizzari, 2024; Dou et al., 2024). This makes it challenging to reconstruct occluded objects and the human body in the exocentric view. To address this, future work could draw inspiration from methods designed to recover occluded object regions (Ke et al., 2021; Chen et al., 2025) and human full-body structures (Qiu et al., 2020; Wan et al., 2024). Second, egocentric cameras typically follow the wearer's movements, resulting in abrupt background shifts, as noted in Luo et al. (2024); Liu et al. (2024). This is especially pronounced in dynamic activities such as sports. Such instability makes it challenging to synthesize a coherent exocentric background. IDE (Luo et al., 2024) attempts to mitigate this problem by leveraging co-occurring objects across views to mine consistency. However, this strategy is insufficient for capturing large-scale background structures. To overcome these challenges, future research could build on advances in novel view synthesis that model dynamic scenes (Shao et al., 2023; Pumarola et al., 2021; Song et al., 2023).

Action Understanding. Human action analysis is widely studied with third-person data (Carreira & Zisserman, 2017; Miech et al., 2019; Gu, 2018; Soomro et al., 2012). The exocentric perspective captures the full body movements but often misses action details. In contrast, egocentric videos excel at capturing detailed human-object and human-human interactions, which offer a complementary viewpoint to enhance exocentric action understanding.

To leverage complementary egocentric perspectives, Reilly et al. (2025) propose a distillation approach, as illustrated in Fig. 7. This approach employs projectors to align video features with large language models embeddings, followed by knowledge distillation to transfer egocentric cues into exocentric representations. It highlights the potential of egocentric cues in improving exocentric activity understanding for large vision-language models.

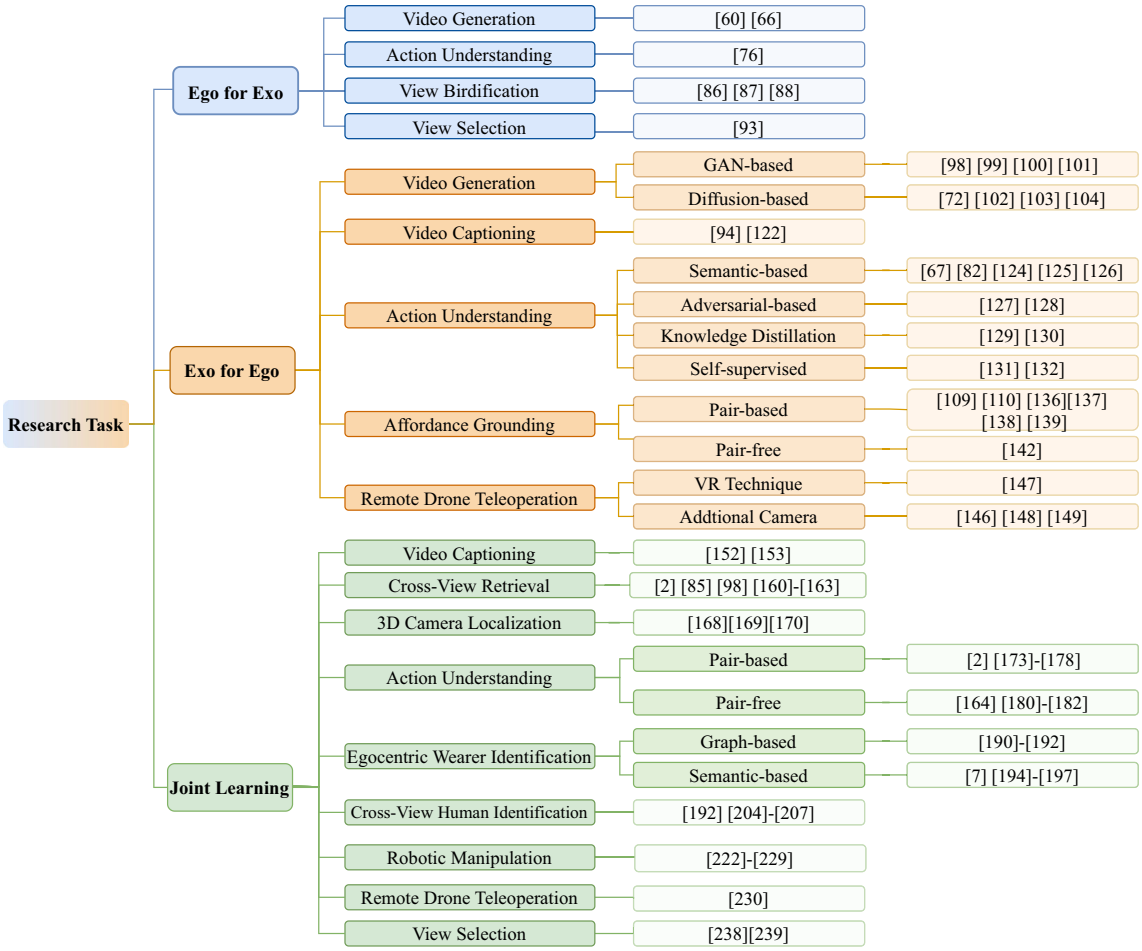

Fig. 5 Overall structure of Section Research Tasks. We discuss the research task from three aspects: Egocentric for Exocentric, Exocentric for Egocentric, and Joint Learning. Each subsection reviews a variety of tasks and their existing works.

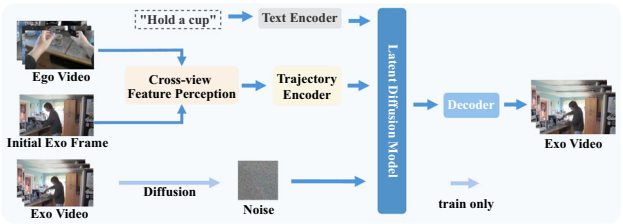

Fig. 6 Illustration of a diffusion-based framework for ego-to-exo video generation, adapted from (Luo et al., 2024).

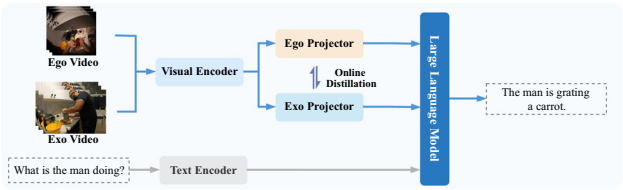

Fig. 7 Illustration of a typical method for ego-for-exo action understanding, adapted from (Reilly et al., 2025). This method distills egocentric cues into exocentric representations.

• Discussion: While recent work shows promising progress, the usefulness of egocentric data is task-dependent. For actions such as walking or cycling, key body parts are rarely visible in first-person view, offering little information for action recognition (Li et al., 2024). By contrast, for fine-grained manipulations, egocentric cues can be critical. For example, tightening and loosening a screw may appear similar in third-person videos but are distinguishable from the egocentric viewpoint. Therefore, a key open problem is to automatically determine how to integrate egocentric data into exocentric action understanding. Besides, from an application perspective, egocentric data can enhance action analysis in domains requiring precise instrument handling, such as industrial assembly and surgery. It can provide unique benefits for monitoring skill acquisition (Bertasius et al., 2017; Doughty et al., 2019; Huang, 2024; Grauman, 2024b) and detecting execution errors (Lee et al., 2024; Wang et al., 2023a). Despite these advantages, several domain-specific challenges remain. In these environments, multiple visually similar tools often coexist. Egocentric views mainly cap-

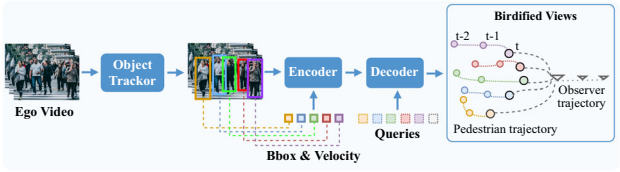

Fig. 8 Illustration of a typical method for view birdification, adapted from (Nishimura et al., 2023a). This task aims to estimate the trajectories of a crowd in a bird's-eye view from an observer's egocentric perspective.

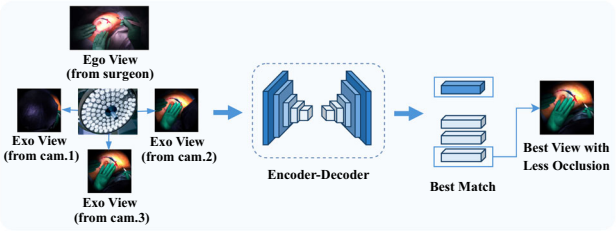

Fig. 9 Illustration of a typical method for view selection in surgical recording, adapted from (Saito et al., 2021). It aims to identify the exocentric view with minimal occlusion by using the surgeon's egocentric perspective as a selection criterion.

ture the actively manipulated tool, while exocentric views show the broader spatial arrangement. This mismatch makes it difficult to translate fine-grained ego cues into higher-level action semantics. Future research may leverage text descriptions (Luo et al., 2024; Wang et al., 2023b), gaze information (Kasahara et al., 2022), or shared attention regions (Yu et al., 2019, 2023) to bridge this gap.

View Birdification. This task aims to estimate the trajectories of a crowd from a bird-eye's view from egocentric videos captured by an observer. It recovers the global movements of people from the observations of the observer. This task has a wide range of applications such as crowd behavior analysis and surveillance.

In Nishimura et al. (2023b), a cascaded optimization based method is proposed to alternate between estimating the displacements of the egocentric camera and its surrounding pedestrians. However, this iterative approach incurs high computational cost. To address this issue, ViewBirdiformer (Nishimura et al., 2023a) proposes a transformer-based architecture that performs view birdification in a single forward pass. As illustrated in Fig. 8, it first utilizes a multi-object tracking algorithm to extract pedestrian movements, including bounding box coordinates and velocity vectors. These features are then encoded via a transformer encoder to model pedestrian interactions. Subsequently, the transformer decoder leverages camera queries and pedestrian trajectory queries from the previous timestep to predict pedestrian trajectories for the next timestep. In subsequent work, InCrowdFormer (Nishimura et al., 2023) addresses uncertainties caused by unknown pedestrian heights and simultaneously predicts pedestrian trajectories along with their associated uncertainty probabilities.

• Discussion: view birdification has promising applications in crowd management and security monitoring (Kratz & Nishino, 2009). These scenarios mainly rely on fixed surveillance cameras, which are often hindered by limited coverage. In contrast, mobile egocentric cameras can effectively capture blind spots and dynamically track targets. However, their restricted field of view introduces task-specific challenges. In crowded scenes, the egocentric view may only reveal a subset of individuals, resulting in severe and dynamic occlusions. Therefore, ego-based birdification must infer global trajectories from fragmented observations. To address this issue, future research could leverage cues from human intention and group dynamics. Since human movement is goal-directed and continuous, methods such as trajectory anticipation and social behavior modeling (Alahi et al., 2016; Huang et al., 2019; Yue et al., 2022) may help recover paths of individuals occluded in the egocentric view.

View Selection. This task leverages egocentric video to guide the automatic selection of the most informative exocentric view. A key application arises in the surgical domain, where recording systems typically employ multiple cameras mounted on the surgical lamp. To minimize occlusion and fully capture the surgical field, it is essential to automatically determine the optimal camera view at each moment.

A pioneering study in this area is Saito et al. (2021). Building on the observation that the surgeon's perspective is often the most effective to capture surgical targets, this method selects the exocentric camera that best aligns with the surgeon's egocentric view, as demonstrated in Fig. 9. However, it relies solely on image similarity and ignores the temporal dynamics of the surgical workflow.

• Discussion: Beyond the surgical domain, ego-guided exo view selection shows promise in applications such as industrial training video production. An expert's egocentric stream can guide a multi-camera system to automatically select the most informative static angle that highlights fine-grained manipulations, thereby reducing post-production effort. However, realizing such applications requires addressing task-specific challenges. A key issue is that the egocentric stream is not always a reliable reference for exocentric selection. Rapid head movements may introduce motion blur or capture irrelevant background. Naive feature matching can propagate these distortions to the exocentric view or trigger distracting, unnecessary switches. To overcome this, future work should explore models capable of deciding when to trust the egocentric cue. Future work could incorporate procedural knowledge, such as textual instructions (Xu, 2024) or task graphs (Lee et al., 2024; Seminara et al., 2024), to better align selection with high-level human intent. Moreover, temporal sequence modeling (Bansal et al., 2022; Lu &

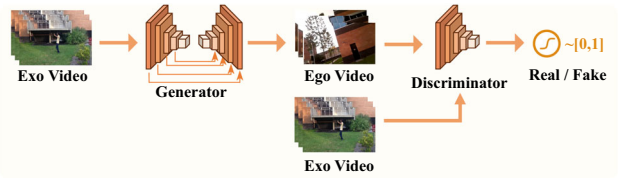

Fig. 10 Illustration of the general GAN-based framework for exo-to-ego video generation. The generator uses exocentric images to synthesize egocentric views, while the discriminator distinguishes between real and synthesized egocentric images.

Elhamifar, 2024) offers a way to suppress short-term motion noise while preserving longer-horizon task context.

4.2 Exocentric for Egocentric

Exocentric perspectives can complement egocentric analysis by providing a broader view of the environment. Additionally, large-scale exocentric video datasets (Carreira & Zisserman, 2017; Miech et al., 2019; Gu, 2018; Soomro et al., 2012) has driven significant progress in exocentric video understanding (Simonyan & Zisserman, 2014; Carreira & Zisserman, 2017; Arnab et al., 2021; Bertasius et al., 2021). Building on these advancements, recent studies have investigated leveraging data and models from the exocentric domain to enhance egocentric analysis. This subsection reviews key approaches that utilize exocentric video techniques to improve egocentric tasks.

Video Generation. Exo-to-ego generation aims to create a first-person view from third-person recordings. This task benefits various fields. For example, in VR and AR applications, exo-to-ego generation can help the users understand procedures by converting third-person videos into their own perspectives. Similarly, the embodied agents can leverage exo-to-ego generation to better understand their surrounding environment.

Current exo-to-ego generation approaches can be categorized into *GAN-based* (Elfeki et al., 2018; Liu et al., 2021, 2022; Garello et al., 2022) and *diffusion-based* (Luo et al., 2025; Liu et al., 2024; Cheng et al., 2024; Spisak et al., 2024) methods. In Elfeki et al. (2018); Garello et al. (2022), exocentric images are used as conditional inputs to GAN for synthesizing egocentric images. Fig. 10 illustrates the general framework of GAN-based approaches. Liu et al. (2022) proposes a two-parallel-GANs architecture to transform images from one viewpoint to another. However, these works (Elfeki et al., 2018; Garello et al., 2022; Liu et al., 2022) are limited to image generation. For video generation, STA-GAN (Liu et al., 2021) proposes a bi-directional GAN to learn both spatial and temporal information. However, it relies on semantic maps for guidance to overcome generation ambiguities. More recent work (Luo et al., 2025; Liu et al., 2024; Cheng et al., 2024; Spisak et al., 2024) leverages diffusion models. Exo2Ego (Luo et al., 2025) and Exo2Ego-V (Liu et al., 2024) focus on synthesizing videos of human activities, while (Spisak et al., 2024) targets robot manipulation scenarios.

• Discussion: Despite ongoing research, transforming exocentric videos into egocentric views holds promise for real-world applications, particularly in educational settings such as industrial and surgical training (Halim et al., 2024; Neuwirth & Ros, 2021). In these domains, first-person perspectives can provide trainees with immersive experiences and highlight key object–action interactions. However, applying this technique faces challenges beyond generic video generation. First, exo-to-ego generation must simulate naturalistic, intention-driven human motion and gaze behavior (Li et al., 2024). These cues are vital in instructional contexts, as they direct learners' attention to relevant objects and actions. Future work may explore IMU signals (Zhang et al., 2025) for head dynamics, gaze prediction modules (Kasahara et al., 2022) for attention focus, or affordance grounding (Luo et al., 2022; Li et al., 2023) to inject priors on hand–object interaction. Second, unlike generating alternative static viewpoints (Mildenhall et al., Dec. 2021; Barron et al., 2021, 2022), egocentric synthesis requires modeling dynamic and coherent background changes (Liu et al., 2024). While Exo2Ego-V (Liu et al., 2024) makes progress by employing PixelNeRF (Yu et al., 2021) to learn view-translation priors, it depends on multiple exocentric views and known camera poses, which are often unavailable in real-world scenarios. Future research could explore single-view NeRF methods (Deng et al., 2023; Gu et al., 2023) or dynamic 3D Gaussian Splatting (Luiten et al., 2024; Wu et al., 2024) for scene reconstruction.

Video Captioning. This task involves generating descriptive textual narratives for videos, aiming to produce coherent sentences that describe the actions, objects, and interactions in the video.

Traditionally, video captioning has been extensively studied in the context of third-person videos (Liu et al., 2018; Pan, 2020; Munusamy & Sekhar, 2020), supported by large-scale exocentric video datasets. In contrast, egocentric video captioning has received less attention due to the limited availability of large-scale, high-quality egocentric datasets.

Currently, a promising direction for egocentric video captioning is leveraging large-scale third-person data. To mitigate domain shift, Ohkawa et al. (2023) introduce an intermediate ego-like view to gradually adapt from exocentric to egocentric views. On the other hand, EgoInstructor (Xu, 2024) is a retrieval-augmented captioning model that uses semantically relevant exocentric videos as references for egocentric video captioning, as shown in Fig. 11.

• Discussion: Despite recent progress, adapting video captioning from exocentric to egocentric perspectives remains challenging. First, the semantic focus differs: exocentric

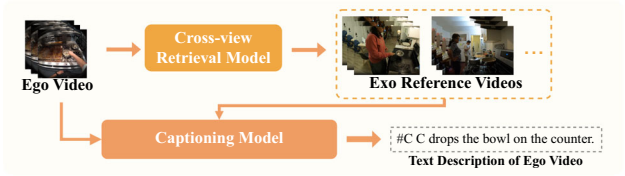

Fig. 11 Illustration of a typical method for exocentric for egocentric video captioning, adapted from Xu (2024). This method retrieves relevant exocentric videos to serve as references for captioning egocentric videos.

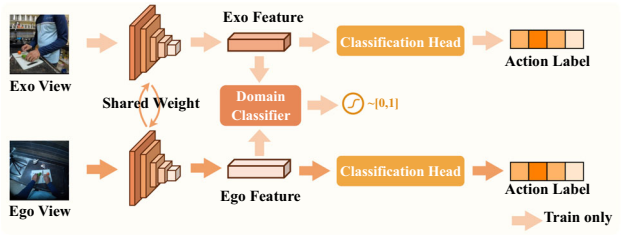

Fig. 12 Illustration of a general adversarial-based approach for exo-for-ego action understanding. During training, the domain classifier differentiates between egocentric and exocentric features, while the model is optimized to deceive it. During inference, only egocentric videos are used.

captions often emphasize global scene and social context, while egocentric captions must highlight hand–object interactions and fine-grained manipulations (Grauman, 2024a). This mismatch makes direct transfer prone to missing task-relevant details. Second, egocentric videos inherently include wearer's attention through gaze and hand–eye coordination (Li et al., 2013), which are absent in most third-person data, leading to temporal misalignment and insufficient action granularity.

Action Understanding. Due to the availability of large-scale exocentric datasets (Carreira & Zisserman, 2017; Miech et al., 2019; Gu, 2018; Soomro et al., 2012), exocentric action understanding has been extensively studied. Consequently, a body of research explores leveraging knowledge from the exocentric domain to improve understanding of egocentric action.

Semantic-based methods focus on leveraging shared semantics between egocentric and exocentric videos to bridge the gap between the two domains. Existing studies have explored the use of activity sounds (Zhang et al., 2022), geometric correlations (Truong & Luu, 2023), skeleton poses (Rocha et al., 2023), and narrations (Wang et al., 2023b) to establish relationships between egocentric and exocentric perspectives. In addition, EMBED (Dou et al., 2024) utilizes hand-object interactions to transform exocentric video-language datasets into egocentric style.

Adversarial-based methods employ adversarial strategies to minimize the discrepancy between the exocentric domain (source domain) and the egocentric domain (target domain). In Choi et al. (2020); Wang et al. (2022), a domain classifier is utilized to differentiate whether the feature originates from egocentric or exocentric videos. During training, the model is optimized to generate features to fool the domain classifier, thereby aligning egocentric features with exocentric features. Fig. 12 demonstrates the general adversarial-based framework.

Knowledge distillation methods seek to distill knowledge from exocentric models to improve egocentric action understanding. In Quattrocchi et al. (2023); Li et al. (2021), the model is first trained on exocentric videos. Subsequently, knowledge distillation losses are applied to adapt the model for egocentric videos.

Self-supervised methods address the challenge of requiring large-scale labeled egocentric data. Egofish3D (Liu et al., 2023) utilizes 3D poses estimated by an exocentric pose estimator as supervision signals to train an egocentric pose estimator without 3D ground truth annotations. Ex2Eg-MAE (Tran et al., 2025) first learns to reconstruct exocentric frontal facial videos using synthesized multi-view data that emulate egocentric environments and then evaluates on egocentric social role understanding tasks.

• Discussion: The landscape of exo-for-ego action understanding highlights complementary but still fragmented progress across different methodological streams. Semantic-based methods provide intuitive bridges between perspectives, yet their reliance on auxiliary signals such as sounds, skeletons, or narrations makes them sensitive to noise and incomplete coverage, especially when crucial manipulative cues are absent from exocentric recordings (Zhang et al., 2022; Rocha et al., 2023; Wang et al., 2023b). Adversarial approaches (Choi et al., 2020; Wang et al., 2022) achieve cross-domain alignment in feature space, but training is notoriously unstable; mode collapse and over-regularization often dilute fine action boundaries, limiting downstream interpretability. Knowledge distillation methods (Quattrocchi et al., 2023; Li et al., 2021) face the "teacher–student bias" problem: distilled egocentric models inherit not only strengths but also blind spots of exocentric teachers, hindering adaptation to viewpoint-specific cues. Meanwhile, self-supervised techniques such as Egofish3D (Liu et al., 2023) and Ex2Eg-MAE (Tran et al., 2025) reduce annotation dependence but raise new questions of cross-modal consistency, since proxy signals (e.g., estimated 3D poses or synthetic multi-view data) may not faithfully reflect real-world egocentric dynamics. Beyond these method-level issues, future applications require moving from coarse recognition to structured reasoning, such as procedural order (Song et al., 2024a) and error modes (Sener, 2022). Progress will likely depend on integrating richer multi-modal supervision, such as combining visual, audio, and textual signals, to capture the full spectrum of egocentric interactions.

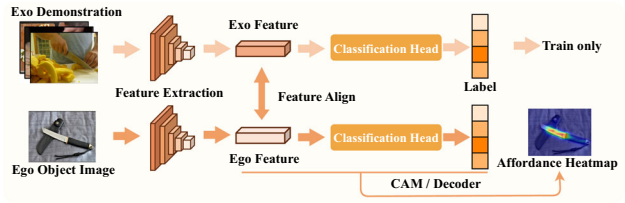

Fig. 13 Illustration of the exo-for-ego affordance grounding framework. During training, egocentric object images are aligned with exocentric demonstration images with the same affordance label. During inference, only the egocentric image is inputted to identify the affordance region.

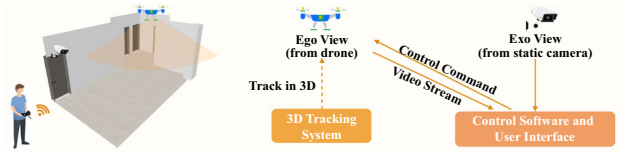

Fig. 14 Illustration of remote drone teleoperation with additional cameras, adapted from Li et al. (2020). To overcome the limited view of egocentric cameras on drones, exocentric cameras are used to capture the surrounding environment.

Affordance Grounding. This task aims to identify and localize the interaction regions of objects based on given instructions. In this task, the exocentric view captures the interactions between human and object while the egocentric view refers to the object only images. Affordance grounding plays a critical role in applications such as embodied intelligence (Song et al., 2024b; Ju et al., 2025), where robots must not only recognize objects but also understand how to interact with them.

Exo-for-Ego affordance grounding methods can be categorized into two types based on training data: *pair-based* and *pair-free*. Fig. 13 presents a general framework for this task.

Pair-based method (Luo et al., 2022; Li et al., 2023; Yang et al., 2024; Xu et al., 2024; Zhang et al., 2024; Rai et al., 2024) learn from a group of exocentric images and the corresponding egocentric object image that share the same affordance label. During inference, only the egocentric object image is used. Luo et al. (2022) introduce Cross-View-AG based on Class Activation Mapping (CAM) Zhou et al. (2016), which has served as a foundational paradigm for many subsequent studies. However, CAM is only used in post-processing during inference and lacks effective supervision for the generated affordance map. To address this, LOCATE (Li et al., 2023) replaces the vanilla CAM with a learnable module to enable supervision of the CAM-generated map. Furthermore, GAAF-Dex (Yang et al., 2024) enhances (Li et al., 2023) by applying concentration loss to make the affordance map more compact.

With advances in large language models (LLMs), a number of studies (Xu et al., 2024; Zhang et al., 2024; Rai et al., 2024) integrate language signals into affordance grounding learning. WSMA uses CLIP (Radford et al., 2021) to encode affordance labels and fuses them with egocentric image embeddings. However, it does not address the issue of action ambiguity, where an object may support multiple actions. To address this limitation, Zhang et al. (2024) enable the model to predict both affordance region and object-action descriptions. In contrast to Zhang et al. (2024), Rai et al. (2024) utilize world knowledge from LLMs to generate more detailed captions that include information about object parts and attributes.

Unlike previous work, *pair-free methods* do not require paired inputs. Instead, they learn from a group of exocentric images and their affordance information. INTRA (Jang et al., 2024) uses contrastive learning as a weakly supervised objective to extract shared knowledge from different affordance labels.

• Discussion: Current pair-based methods rely on explicit exo–ego alignment but mostly operate on 2D image correspondences. This makes it difficult to capture how 3D hand–object interactions in exocentric views translate into object-centric affordances in egocentric views. Pair-free methods relax the dependence on paired data but sacrifice fine-grained supervision, especially when affordances evolve dynamically during actions. A central challenge is that affordance is inherently spatio-temporal: the same object may afford different regions across phases of manipulation, e.g., grasping and pouring in the case of a cup. Since such transitions are often more evident in exocentric views, future research could focus on learning 3D temporal affordance representations (Delitzas et al., 2024) and transferring them into egocentric object views. This cross-view spatio-temporal reasoning can better disambiguate action–object relations, which is crucial for applications such as robotic manipulation and AR/VR training.

Remote Drone Teleoperation. Drones can navigate challenging environments or locations impassable for humans. It has a wide range of applications such as disaster investigations (Kyrkou & Theocharides, 2020) and product delivery (Cauchard et al., 2021). Typically, drone control systems offer an egocentric view through an on-board camera. However, this limited field of view fails to fully capture the surroundings.

To address the limitations of egocentric views, previous research has explored using *VR technique* (Erat et al., 2018) or *additional cameras* (Li et al., 2020; Temma et al., 2019; Inoue et al., 2023) to provide exocentric views. Fig. 14 illustrates using overhead camera to provide exocentric views for drone teleoperation. In Erat et al. (2018), VR technique provides a 3D model of the environment, allowing pilots to perceive the drone's surroundings. Another line of works (Li

et al., 2020; Temma et al., 2019; Inoue et al., 2023) utilize additional cameras to capture the environment of the drone. StarHopper (Li et al., 2020) uses a fixed overhead camera while Temma et al. (2019) uses a secondary drone that semi-automatically flies around the primary drone. Inspired by Temma et al. (2019), BirdViewAR (Inoue et al., 2023) further uses AR overlays to highlight the primary drone's spatial status and proposes an automatic framing method to ensure the secondary drone follows the primary drone in fast-moving scenarios.

• Discussion: The use of exocentric views in drone teleoperation provides a crucial remedy to the limited situational awareness offered by egocentric cameras. Prior works demonstrate the benefits of fixed overhead cameras (Li et al., 2020), auxiliary drones (Temma et al., 2019), and augmented overlays (Inoue et al., 2023), each enhancing global perception and orientation. Nevertheless, these solutions raise challenges regarding viewpoint synchronization and the operator's ability to seamlessly integrate heterogeneous perspectives. An open research direction is to design fusion strategies that can dynamically switch or combine egocentric and exocentric feeds based on environmental complexity and task demands. Such systems would reduce cognitive burden while preserving both fine-grained local details and global context, improving task efficiency in real-world drone teleoperation.

4.3 Joint Learning

Joint learning aims to leverage both egocentric and exocentric perspectives to address cross-view video understanding tasks. It requires both egocentric and exocentric views as input during both training and inference. This contrasts with unidirectional paradigms (e.g., exo-for-ego or ego-for-exo), where often one view serves as auxiliary information during training, but only a single view is utilized at test time. In joint learning, however, it emphasizes bidirectional collaboration to resolve cross-view tasks. Below, we systematically review advancements in cross-view tasks, highlighting diverse strategies for effectively integrating the complementary nature of egocentric and exocentric perspectives.

Video Captioning. In daily life, video captioning can document a wide range of human activities in natural language. This capability can enhance the development of smart assistants (Huang et al., 2024a, b) to help humans memorize and retrieve items.

Current research (Nakashima et al., 2018, 2020) investigates captioning lifelog videos in multi-view settings. The logging system comprises a first-person view from an individual, a second-person view from a service robot, and a third-person view from a fixed camera, as demonstrated in Fig. 15. In Nakashima et al. (2018), multi-view images are

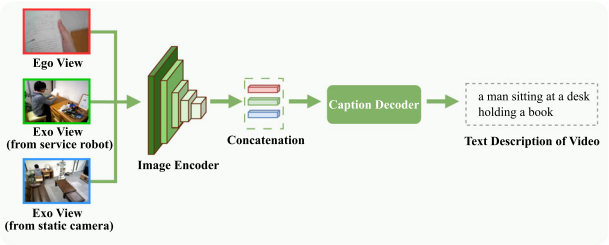

Fig. 15 Illustration of video captioning from three perspectives: egocentric view from human, exocentric view from a service robot, and exocentric view from a static camera.

independently processed into image features, which are then concatenated and projected into a unified feature space. The unified features are subsequently input into a caption decoder to generate captions. In contrast, Nakashima et al. (2020) employ attention mechanisms for feature fusion. This method first uses Faster R-CNN (Ren et al., 2015) to detect salient regions from each view. To address redundant cross-view information, the detected features are clustered into several groups and then fused via attention mechanisms.

• Discussion: For ego-exo video captioning, several challenges remain for future research. One key issue is balancing description granularity. Due to the different fields of view, egocentric and exocentric videos may emphasize different visual elements. This requires models to reconcile these disparities to generate consistent captions. Additionally, as discussed in Chang et al. (2024); Kuribayashi et al. (2025), users may prefer different levels of detail. Future research should enable model to adjust description granularity to align with user-specific needs. Another challenge is managing redundant and complementary information across views. While prior work (Nakashima et al., 2020) addresses this by clustering features at frame-level, it overlooks action-level correspondences. For example, an egocentric view might depict "hand pulls a lever", while an exocentric view captures "doors open". To generate coherent captions, models must integrate cross-view action dependencies. To achieve this, future work can integrate techniques like action segmentation (Quattrocchi et al., 2023; Sarfraz et al., 2021) and action relation (He, Mar. 2024; Qian et al., 2022). Beyond technical challenges, joint video captioning holds significant promise for smart assistant (Huang et al., 2024a, b). By integrating multiple perspectives, such systems can generate comprehensive activity logs, enabling assistants to memorize historical events and support downstream tasks like temporal grounding and visual question answering.

Cross-View Retrieval. This task focuses on identifying and retrieving corresponding visual elements, such as videos (Ardeshir et al., 2016), frames (Elfeki et al., 2018; Sigurdsson et al., 2018; Yu et al., 2023), and moments (Sigurdsson

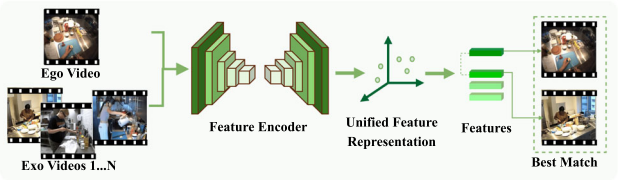

Fig. 16 Illustration of the general cross-view retrieval framework. Exocentric and egocentric videos are encoded into a shared representation space to retrieve the best match from the alternate view.

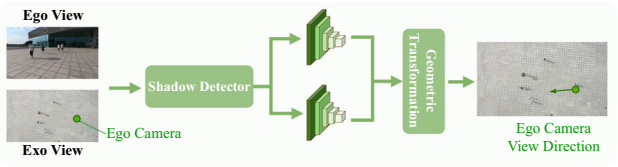

Fig. 17 Illustration of a typical method for egocentric camera localization, adapted from Han et al. (2023). This method uses shadows to relate egocentric and exocentric views(top views), and estimates the egocentric camera direction in the top view.

et al., 2018; Yu et al., 2023), from different viewpoints, as demonstrated in Fig. 16.

Early work (Ardeshir et al., 2016) explores linear and nonlinear mappings to transform motion features between two views. More recent approaches (Elfeki et al., 2018; Sigurdsson et al., 2018) first utilize separate branches to extract features from different views and then employ contrastive learning to align representations. Furthermore, T-JANet (Yu et al., 2023) leverages overlapping attention regions between views to guide representation learning. VIEWPOINTROSETTA (Luo et al., 2025) introduces a translator module that is first trained on paired data to map features across viewpoints. This translator is then used to synthesize features in another viewpoint, forming pseudo ego–exo pairs. By combining real and pseudo pairs in a unified contrastive learning framework, this method enhances view-invariant representation learning. However, these works mainly address cross-view correspondence at the video level. Recently, Ego-Exo4D (Grauman, 2024b) introduces a cross-view object correspondence task, which aims to predict object masks in one view given queries from another view. PSALM (Zhang et al., 2025) demonstrates zero-shot capability for this task. It first utilizes LLM to process visual and textual prompts, followed by a general segmentation model to generate object masks. Building on PSALM (Zhang et al., 2025), ObjectRelator (Fu et al., 2024) generates descriptive language prompts for query objects to exploit the LLM's reasoning ability. To address object appearance disparities across views, ObjectRelator (Fu et al., 2024) further introduces a cross-view object alignment module to project masks from different views into unified space.

• Discussion: Current approaches primarily learn shared representations across views, but inherent view disparities complicate feature alignment. A more promising direction is to disentangle features according to their role in the retrieval task (Huang et al., 2022). View-invariant features can capture high-level task semantics and serve as anchors for coarse-grained retrieval, while view-specific features model perspective-dependent cues. Retrieval can first use view-invariant features for initial matching, then leverage view-specific features for fine-grained re-ranking. Beyond technical challenges, the unique value of ego–exo retrieval can be illustrated in smart assistive systems (Huang et al., 2024a). For instance, a user wearing AR glasses may query a recipe tutorial based on their real-time view of a kitchen counter. The system must personalize recommendations to user's proficiency. This goes beyond generic retrieval systems that typically return popular tutorials. To support such applications, several task-specific challenges must be addressed. The continuous egocentric video stream demands long-term video understanding techniques (Islam et al., 2024; Ye et al., 2024) to align them with scripted exocentric demonstrations. In addition, memory mechanisms (Huang et al., 2024a; Yang et al., 2025) and skill assessment (Bertasius et al., 2017; Doughty et al., 2019) are needed to infer the user's characteristics from past egocentric videos. This inferred skill level information can guide the retrieval system to provide tutorials with suitable guidance.

3D Camera Localization. This task aims to determine the position and orientation of a camera in the environment.

Han et al. (2023) and Qian et al. (2022) propose to localize egocentric cameras from a global top-down view. Han et al. (2023) leverage shadow to relate egocentric and top views and propose a shadow detection model to predict shadow direction, as shown in Fig. 17. Furthermore, Qian et al. (2022) utilize the spatial distribution of subjects in the 3D environment to estimate egocentric camera poses in a virtual top-down view. In contrast to Han et al. (2023); Qian et al. (2022), YOWO (Yang et al., 2024) introduces a novel approach to localize ceiling-mounted cameras (CMCs). Previous methods (Ataer-Cansizoglu et al., 2014; Yi, 2023) typically use SLAM for scene reconstruction and subsequently employ visual localization to estimate camera poses. However, the perspective disparity between egocentric and exocentric views poses challenge for cross-view localization. Moreover, the static nature of CMCs prevents using motion information to correct localization errors. To address these limitations, YOWO jointly optimizes scene reconstruction and CMC registration. It employs a mobile agent to navigates the environment to generate both agent trajectories and scene layout. Meanwhile, CMCs capture the agent to provide pseudo trajectories. By correlating these trajectories, YOWO aligns CMC poses with the scene layout.

• Discussion: A critical limitation of current egocentric camera localization work (Han et al., 2023; Qian et al., 2022) is

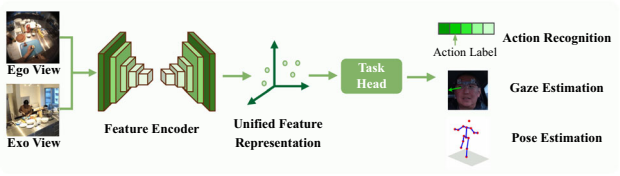

Fig. 18 Illustration of cross-view action understanding. This involves action recognition, gaze estimation and pose estimation tasks.

its reliance on synthetic data, where virtual agents perform random walks in controlled environments. Such settings fail to reflect the dynamics of real-world, unscripted egocentric videos. A similar concern has also been raised in Plizzari (2024). In natural human behavior, camera motion is not random but instead driven by intention and goals (Damen, 2021; Grauman, 2024a; Song et al., 2024a). This highlights an emerging direction toward semantics-driven localization. Rather than relying solely on low-level geometric features, future systems could exploit high-level intent inferred from video streams. Semantic cues can act as powerful priors to regularize geometric estimation. For example, recognizing the action of opening a door implies a predictable arc-like trajectory with limited vertical displacement. By leveraging intention-driven motion priors, a system could improve localization accuracy, especially in visually ambiguous environments where purely geometric methods would fail.

Action Understanding. As discussed in Lin (2022); Pramanick (2023), models predominantly trained on exocentric videos exhibit poor performance in egocentric data. Cross-view action understanding has emerged as a promising approach to enable a single model to achieve viewpoint-invariant action analysis. This field encompasses multiple key tasks, including action recognition, gaze estimation, and pose estimation, as illustrated in Fig. 18. Current research in this area can be broadly classified into two categories based on training data: *pair-based* and *pair-free*.

Paired-based methods use synchronized egocentric and exocentric video pairs. For action recognition task, Soran et al. (2014); Ardeshir and Borji (2018); Sigurdsson et al. (2018); Yonetani et al. (2016) leverage paired videos to learn a unified feature across different views. Soran et al. (2014) jointly predict action labels and assess each camera's importance. In Ardeshir and Borji (2018); Sigurdsson et al. (2018), egocentric and exocentric videos are encoded by separate branches and subsequently aligned into a unified feature space. Yonetani et al. (2016) use a pair of egocentric videos from two individuals to recognize micro-actions and reactions. In driving scenarios, LBW (Kasahara et al., 2022) utilizes both the driver's face and the forward road scene for gaze estimation. Similarly, Yang (2023) integrate in-vehicle and out-vehicle views to recognize the driver's state. In the field of pose estimation, Dhamanaskar et al. (2023) map multi-view RGB frames and optical flow into a joint embedding space, while Hein et al. (2023) evaluate multi-view methods (Haugaard & Iversen, 2023) for the pose estimation of surgical instruments.

While effective, paired-based approaches are limited by the expense of obtaining synchronized paired data. To address this limitation, recent research (Xue & Grauman, 2023; Kalluri et al., Jul 2024; Xu et al., 2023; Huang et al., 2022) has shifted towards leveraging unpaired videos.

Paired-free approaches aim to learn shared action representations from unpaired egocentric and exocentric video data. During inference, pair-free models demonstrate flexibility by accepting either egocentric or exocentric video inputs for action analysis. This line of work can easily utilize existing large-scale third-person and first-person datasets. To align unpaired data, AE2 (Xue & Grauman, 2023) introduces a temporal alignment strategy. Based on the assumption that aligning egocentric and exocentric videos is inherently easier than aligning them when one sequence is temporally reversed, this approach employs reversed frames as negative samples for contrastive learning. In contrast to AE2 (Xue & Grauman, 2023), LaGTran (Kalluri et al., Jul 2024) leverages language descriptions to mitigate the domain gap between egocentric and exocentric videos. The method is based on the premise that text descriptions exhibit a smaller domain discrepancy compared to the original videos. POV (Xu et al., 2023) incorporates learnable prompts to video tokens to learn view-agnostic representations. Unlike previous work, Huang et al. (2022) highlight the importance of view-specific information and disentangle features into view-invariant and view-specific components.

• Discussion: The recent research shift from paired to unpaired data opens up the opportunity to leverage large-scale egocentric and exocentric datasets. However, as noted in Huang et al. (2022); Luo et al. (2025), videos that are only coarsely matched at the semantic level may still at the temporal or viewpoint level. Such discrepancies make strict frame-level alignment impractical, posing a core challenge for learning consistent cross-view action representations. Future work could design selective modules (Zhu et al., 2025; Zhang et al., 2025) to emphasize key action moments while filtering out blurred egocentric segments or exocentric frames dominated by background. Another promising direction is to incorporate auxiliary modalities, such as audio (Jia, 2024; Jia et al., 2024) or narrations (Xu, 2024; Wang et al., 2023b), to capture implicit action similarities when visual alignment is weak. Furthermore, current research is confined to basic tasks like action recognition, while advanced tasks like action assessment remain underexplored. Integrating both perspectives can offer a holistic understanding of action proficiency. To enable practical deployment, a key challenge is effectively integrating the dynamic granularity of action information across views. Future work could draw on hierarchical archi-

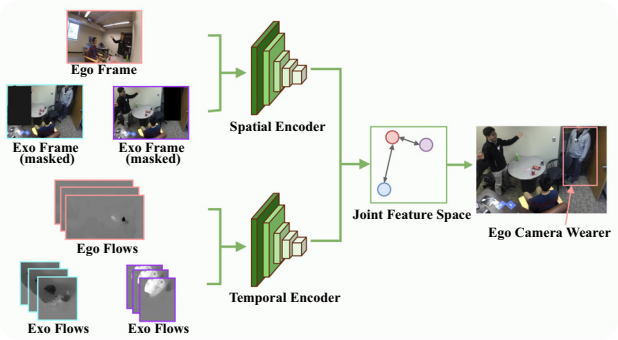

Fig. 19 Illustration of a typical method for egocentric wearer identification, adapted from Fan et al. (2017). This method uses spatial and temporal information to learn view-invariant features and identify the egocentric wearer in exocentric images.

tectures (Fan et al., 2021; Li et al., 2022) to combine these complementary cues into a unified action representation. In addition, incorporating perspective-specific strengths with domain knowledge (Grauman, 2024b; Rao et al., 2024) could support more adaptive assessment. For instance, in sports training, egocentric views are well suited for evaluating hand grip, while exocentric views provide better signals for assessing body posture.

Egocentric Wearer Identification. Given both third-person and first-person videos captured in the same environment, this task aims to identify the egocentric camera wearer in third-person videos. It is similar to person re-identification across different views but is more challenging since the camera wearer seldom appears in the egocentric view.

Early researches (Ardeshir & Borji, 2016, 2019, 2018) employ graph-based techniques to identify the camera holder of egocentric videos in top-view videos. Ardeshir and Borji (2016) models each video view as a graph and proposes a spectral graph matching technique. Building on this, Ardeshir and Borji (2019) extends the work of Ardeshir and Borji (2016) by considering time delays across videos. Furthermore, Ardeshir and Borji (2018) employs visual, geometric, and spatiotemporal reasoning to generate candidates and then use graph cuts (Fulkerson et al., 2009) to evaluate candidates.

More recent approaches (Fan et al., 2017; Xu et al., 2018; Zhao et al., 2024; Yang et al., 2019; Wen et al., 2021) leverage shared semantic across views. Fan et al. (2017) leverage spatial (RGB frames) and temporal (optical flow) similarities to relate two views, as shown in Fig. 19. It employs contrastive learning to predict the camera wearer, utilizing first-person videos paired with third-person videos (masking the correct wearer) as positive samples, and third-person videos (masking a random person) as negative samples. However, this approach primarily focuses on appearance similarity across views, overlooking the dynamic nature of the environment. To address this limitation, Visual-GPS (Yang et al., 2019) leverages motion and action information to improve robustness, as these features are less sensitive to environmental variations. Subsequent work (Wen et al., 2021) proposes a more challenging setting: predicting the camera wearer's location and pose in a third-person scene frame, where the wearer is absent. Furthermore, Xu et al. (2018) and Zhao et al. (2024) jointly address person identification and segmentation and prove that solving these two problems simultaneously is mutually beneficial.

• Discussion: Current appearance-based methods (Fan et al., 2017; Xu et al., 2018; Zhao et al., 2024) may fail when the wearer is partially visible in the exocentric view. In such cases, even motion cues may struggle if critical body parts are occluded. Furthermore, in crowded scenarios, similar appearances (e.g., shared clothing) or similar actions (e.g., group sports) further hinder discriminative feature extraction. To address these limitations, future work could incorporate additional cues, such as object interactions (Xu, 2023; Shiota et al., 2024; Yang et al., 2025) or person-person interactions (Jia et al., 2024; Yonetani et al., 2016), to provide more distinctive information. Beyond technical challenges, egocentric wearer identification shows value in domains such as law enforcement. After a crime event, authorities may possess an anonymous first-person video capturing a critical moment (e.g., a suspect's face) alongside wide-angle surveillance footage of the scene. The task is to locate the individual who recorded the egocentric video in a dense crowd. In such settings, simple spatial proximity is unreliable, as multiple individuals may occupy similar positions. Future research could leverage cues like head motion (Thapar et al., 2020) and hand gestures (Thapar et al., 2020; Tsutsui et al., 2021) to infer ego-wearer identity. Crucially, these cues must be extracted from the egocentric stream and then associated with exocentric observations. Furthermore, unlike studies focusing on limited-person environments (Fan et al., 2017; Xu et al., 2018; Zhao et al., 2024), real-world scenarios require performing a "needle-in-a-haystack" search over large populations. A potential solution is a coarse-to-fine pipeline: coarse pruning first leverages spatial layout or salient objects to restrict candidate regions, and the remaining candidates are then progressively refined through fine-grained matching.

Cross-View Human Identification. This task aims to detect and identify the same individuals across views. Current approaches (Ardeshir & Borji, 2018; Han et al., 2020a, b, 2022a, b) study this task on top-view and side-view. The top view, captured by drones at high altitudes, covers large areas and displays human spatial distribution. In contrast, side views from mounted cameras provide more details. Ardeshir and Borji (2018) propose a graph-based technique while Han et al. (2020a) use a multi-view human association algorithm to match individuals across different views. However, these work (Ardeshir & Borji, 2018; Han et al., 2020a) are limited to human identification across views and does not address

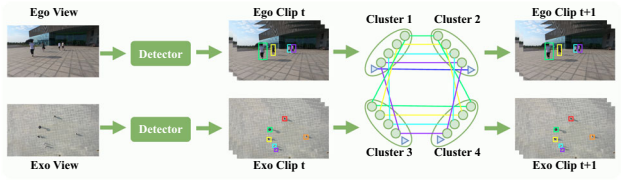

Fig. 20 Illustration of a typical method for cross-view human tracking and association, adapted from Han et al. (2020b). This method segments video pairs into clips and tracks individuals across clips and views.

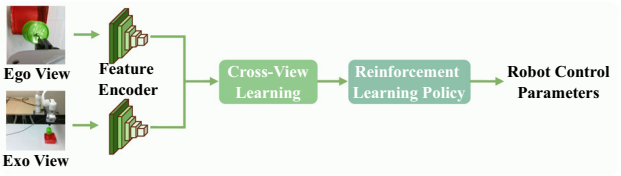

Fig. 21 Illustration of a typical framework of multi-view robotic manipulation, adapted from Jangir et al. (2022). Multi-view data is integrated via cross-view learning module.

tracking. Han et al. (2020b) propose a joint optimization model for identifying and tracking. This approach first segments video pairs into clips and tracks individuals across clips and views, as demonstrated in Fig. 20. Additionally, Han et al. (2022a) extends this work by incorporating spatial distribution for cross-view association and introducing a new approach for appearance reasoning. Previous approaches (Han et al., 2020a, b, 2022a) rely on offline detection models (Redmon et al., 2015) to detect human bounding boxes, which may hinder association performance. To address this, Han et al. (2022b) propose a joint method for cross-view multi-human detection and association.

• Discussion: The dynamic and unconstrained nature of egocentric cameras introduces challenges that extend beyond those in traditional person re-identification (Re-ID). Existing studies (Ardeshir & Borji, 2018; Han et al., 2020a, b, 2022a, b) are mostly limited to single environments, resembling the traditional Re-ID setting (Zhao et al., 2019; Fu et al., 2019; Hou et al., 2019) where a fixed camera network is assumed. In contrast, in mobile ego–exo scenarios, the set of relevant exocentric cameras changes dynamically as the wearer moves through different environments. Real-world examples include a security guard patrolling a campus or a first responder navigating a disaster zone. This setting poses a hierarchical problem: the system must first address a coarse-grained localization task to identify the "active zone" of relevant exocentric cameras before performing fine-grained person identification. A promising direction is to leverage visual place recognition (Berton et al., 2022; Zhu et al., 2023; Ali-Bey et al., 2023) on egocentric backgrounds to estimate the wearer's approximate location. This localization signal can then guide the dynamic selection of exocentric cameras, restricting the Re-ID process to a relevant subset. Such a coarse-to-fine strategy enables an efficient, context-aware identification process.

Robotic Manipulation. This task involves controlling robots to interact with objects and perform actions, such as grasping or moving, to achieve specific goals.

Multi-view robot manipulation has been widely studied. However, most approaches simply concatenate multi-view observations at the image level (Zhan et al., 2022) or feature level (Seo et al., 2023; Brohan et al., 2022; Zhao et al., 2023; Bharadhwaj et al., 2024; Chi et al., 2023; Zhao et al., 2024), without fully exploiting their complementary characteristics. We focus on approaches that explore integrating the complementary strengths of different perspectives.

Lookcloser (Jangir et al., 2022) utilizes cross-view attention mechanisms to integrate egocentric and exocentric perspectives, as shown in Fig. 21. In Hsu et al. (2022), a variational information bottleneck is applied to third-person representations to mitigate their impact on out-of-distribution generalization. Acar et al. (2024) utilize multi-view data to train a teacher policy, which then guides a single-view student policy through knowledge distillation. Sharma et al. (2019) first use third-person human demonstration videos to generate task goal in robot's perspective, which are then combined with robot's current observation to predict actions. Similarly, Shang and Ryoo (2021) leverage synchronized first-person and third-person demonstrations to learn viewpoint-agnostic representations and then use third-person demonstrations for policy learning. Both MV-MWM (Seo et al., 2023) and MFSC (Wang et al., 2025) introduce multi-view masked reconstruction strategies to learn representations from multi-view observations. Unlike previous approaches, MVD (Dunion & Albrecht, 2024) introduces a robust method that supports varying numbers of cameras in inference.

• Discussion: In ego–exo robotic manipulation, the complementary nature of first-person and third-person views presents unique challenges. Due to the limited field of view of the egocentric camera and potentially suboptimal placement of exocentric cameras, some critical information may not be captured across views. This can make it difficult for MVD (Dunion & Albrecht, 2024) to extract shared features and can significantly impact approaches that rely heavily on egocentric observations (Hsu et al., 2022). Besides, considering the inherent characteristics of each perspective, third-person views are better suited for tasks requiring navigation to target locations, whereas first-person views are more effective for executing fine-grained manipulation. To fully exploit these complementary strengths, future work could explore task-driven feature integration, where the relative weighting of each view is dynamically adjusted according to the specific sub-task. Such an approach would enable more efficient

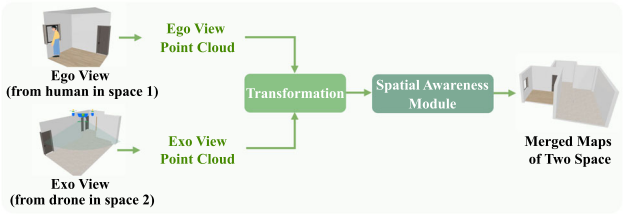

Fig. 22 Illustration of a typical method for remote drone teleoperation with human-drone collaboration, adapted from Morando and Loianno (2024). This method combines user(ego) and drone(exo) perspectives into a unified environmental representation.

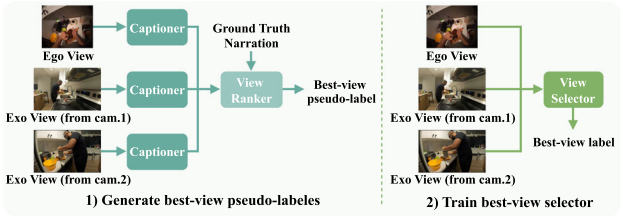

Fig. 23 Illustration of a typical ego-exo view selection method, adapted from Majumder et al. (2024a). This approach leverages video captions as weak supervision for selecting the best view.

multi-view representation learning tailored to the distinct contributions of egocentric and exocentric views.

Remote Drone Teleoperation. Traditional drone manipulation primarily focuses on unidirectional collaboration, where humans send commands to control drones. In contrast, joint learning emphasizes bidirectional information exchange, allowing drones to access the human's perspective for decision-making. This enhanced interaction supports a wider range of collaborative tasks. For instance, in a rescue mission, if a human operator identifies a potential victim through a wearable camera, the drone can autonomously navigate to the location to provide assistance. Such bidirectional communication improves operational efficiency.

A notable work in this field is presented in Morando and Loianno (2024). In this study, point cloud data from the drone and the user's wearable device are merged into a unified environmental representation, as demonstrated in Fig. 22. Then, this approach provides visualizations of the environment from both the user's and the drone's perspectives, ensuring mutual awareness of the surroundings between the user and the drone.

• Discussion: While (Morando & Loianno, 2024) establishes a foundation for mutual spatial awareness, a truly collaborative system must build on this shared perception to enable bidirectional information exchange. On the user side, current system requires operators to switch between user and drone views, which may increase cognitive load and lead to errors (Inoue et al., 2023). To mitigate this, future work could draw inspiration from human–computer interaction research (Vertegaal, 1999; Chen et al., 2023; Lee et al., 2022). One promising direction is the use of gaze-contingent displays, where the operator's focal view is rendered in high resolution while peripheral views are dimmed, thereby reducing visual clutter. On the drone side, the opportunity lies in evolving from a passive tool into a proactive partner, as highlighted in the rescue case. To achieve this, drone systems should be capable of performing semantic and object recognition on the user's egocentric stream and mapping the detected 2D location in the human video to a 3D world coordinate within the shared spatial representation. This capability would allow the drone to autonomously generate actionable navigation goals, thereby complementing human guidance and realizing bidirectional collaboration.

View Selection. The task of selecting the optimal viewpoint from multi-view videos has been widely studied. Prior work explores determining the best camera angles and positions in panoramic 360° views (Hu et al., 2017; Su & Grauman, 2017), and automating viewpoint selection in multi-view systems (Yus et al., 2015; Chen et al., 2019). However, these methods typically address egocentric or exocentric views separately, ignoring scenarios where both views are available.

Unlike previous work, recent work (Majumder et al., 2024a, b) propose to address view selection in instructional videos, which incorporate both egocentric and exocentric perspectives. Majumder et al. (2024a) utilize language descriptions as weak supervision, as shown in Fig. 23. Specifically, the approach generates captions for each view via video captioning models. These captions are scored against ground-truth narration and ranked to produce best-view pseudo-labels, which are utilized to train the view selection model. Another work (Majumder et al., 2024b) proposes a pretext task to detect view switches in instructional videos with varying viewpoints. The model trained for this task is subsequently repurposed to train a view selection model.

• Discussion: In ego–exo settings, the challenge for view selection extends beyond simply choosing the clearest shot. It requires models to balance information granularity: egocentric views capture fine-grained actions, whereas exocentric views provide spatial context and full-body posture. A promising direction is to ground view selection in procedural task semantics, allowing models to decide whether the current phase requires a detailed "zoomed-in" or a contextual "far-view." While (Majumder et al., 2024a) uses narrations for guidance, such annotations often emphasize the present step and neglect broader workflow. One promising approach is to leverage large-scale key-step demonstrations (Song et al., 2024a; Grauman, 2024b, a) to acquire procedural knowledge. Complementary to this, incorporating action anticipation (Wang et al., 2023; Furnari & Farinella, 2019) can further enhance decision-making. By predicting

the user's immediate next step, the system can switch views in advance to ensure that upcoming actions are clearly captured. For example, if the model anticipates a "lifting a heavy box" action, it could proactively select the exocentric view that best frames the full-body motion. This capability allows the system to generate smoother and more informative view transitions.

5 Datasets

We introduce publicly available datasets offering both egocentric and exocentric perspectives. We categorize these datasets based on domain and describe their intended purposes, views, annotations, and unique features. This overview helps researchers select suitable datasets for their studies.

Table 1 provides a summary of the datasets. For datasets that provide synchronized videos, we list the number of first-person and third-person viewpoints. Most datasets cover multiple activity types, while others focus on activities in specialized scenario. Additionally, datasets (Khirodkar et al., 2023; Jia et al., 2020; Guo et al., 2023; Kong et al., 2024; Ardeshir & Borji, 2016; Fan et al., 2017; Xu et al., 2018; Han et al., 2020b; Qian et al., 2022; Yang et al., 2024; Hein et al., 2023) include multi-agent settings, involving multiple participants in a video. This facilitates the analysis of human interactions and collaboration in complex activities. Furthermore, datasets (Huang, 2024; Kasahara et al., 2022; Kong et al., 2024; Grauman, 2024b; Ilaslan et al., 2023; Hein et al., 2023) provide egocentric eye gaze information, offering valuable insights into human intention and decision-making process. Below, we provide a detailed description of each dataset.

A. Action Understanding. Most ego-exo action understanding datasets focus on activities in specific scenarios or controlled environments. **CMU-MMAC** (la Torre Frade et al., 2008) records videos of individuals cooking recipes in a lab kitchen. **H2O** (Kwon et al., 2021), **Assembly101** (Sener, 2022), **ARCTIC** (Fan et al., 2023) and **OAKINK2** (Zhan, 2024) focus on hand-object manipulation on the tabletop. **Homage** (Rai, 2021) captures daily life activities in two houses. **LEMMA** (Jia et al., 2020) features multi-agent goal-directed daily activities in living room and kitchen scenarios. **FT-HID** (Guo et al., 2023) focuses on multi-person interactions and includes 30 human interaction action classes. **EgoExo-Fitness** (Li et al., 2024) focuses on full-body action understanding in natural fitness scenarios. **Charades-Ego** (Sigurdsson et al., 2018) leverages scripts from the Charades (Sigurdsson et al., 2016) and self-collected data, recording multi-view videos of participants performing these scripts. **CORE4D-Real** (Zhang et al., 2024) uniquely captures multi-person and object interactions in household object rearrangement.

More recent datasets involve diverse activities in multiple environments. **Ego-Exo4D** (Grauman, 2024b) is a large-scale multi-view dataset focused on skilled human activities. It offers multimodal annotations, including audio, eye gaze, 3D point clouds, and detailed language descriptions. Both **EgoExoLearn** (Huang, 2024) and **EgoMe** (Qiu et al., 2025) include exocentric demonstration videos and corresponding egocentric recordings of individuals performing the tasks based on the demonstrations. These datasets offer valuable resources for studying how humans interpret and adapt actions from an external perspective to their own.

To analyze human motion, **EgoPW** (Wang et al., 2022), **First2Third-Pose** (Dhamanaskar et al., 2023) and **ECHP** (Liu et al., 2023) are designed for egocentric human full-body pose estimation with support from third-person cameras. Specifically, egocentric videos in ECHP (Liu et al., 2023) are recorded using a head-mounted fisheye camera. **AssemblyHands** (Ohkawa et al., 2023) and **ThermoHands** (Ding et al., 2024) focus on hand-object interaction and provide hand pose annotations. **EgoHumans** (Khirodkar et al., 2023) features 3D pose estimation and tracking. **Nymeria** (Ma et al., 2024) is a large-scale motion dataset collected in the wild, featuring multimodal egocentric data and a third-person view by an observer. In the surgical domain, Hein et al. (2023) propose a multi-view dataset for the pose estimation of surgical instruments.

OVR Dwibedi et al. (2024) is the first multi-view dataset for temporal repetition counting. This task aims to identify repetitive events in a video. Videos in OVR (Dwibedi et al., 2024) are sourced from exocentric dataset Kinetics propose a multi-view dataset kinetic and egocentric dataset Ego4D (Grauman, 2024a). Annotations include the start and end times of repetitions, the number of repetitions, and action descriptions. The open-vocabulary semantics of OVR (Dwibedi et al., 2024) support text-conditioned repetition counting.

B. Driving. Integrating both in-vehicle and out-vehicle views can provide a comprehensive understanding of the driver's behavior. **LBW** (Kasahara et al., 2022) is a multi-view driving dataset for driver's attention estimation. It includes gaze data from eye-tracking glasses and the forward road scene. **AIDE** (Yang, 2023) is designed for assistive driving perception, capturing naturalistic driving from four views: three external (front, left, right) and one internal (driver's state). Annotations cover facial expressions, body postures, gestures, and vehicle conditions. **WTS** (Kong et al., 2024) provides not only vehicle and infrastructure perspectives, but also pedestrian perspectives. It can advance fine-grained video events detection.

C. Affordance Grounding. **AGD20K** (Luo et al., 2022) is the earliest image-level multi-view affordance grounding dataset. It classifies the collected data into seen and unseen sets to evaluate the model's generalization abil-

ity. It has become a widely used benchmark for numerous methods. To advance dexterous manipulation research, **FAH** (Yang et al., 2024) identifies multi-finger grasping regions through detailed hand movement categorization. **PAD** (Luo et al., 2021) provides pixel-level annotations, enabling precise affordance grounding through semantic segmentation models.

D. Generation. **ThirdtoFirst** (Li et al., 2021) is designed for exocentric to egocentric image synthesis. It consists of 531 temporally aligned video pairs. Video collectors perform various actions in front of the exocentric camera (side or top-view), while a body-worn camera captures their motion from the first-person perspective.

E. Scene Understanding. **360+x** (Chen et al., 2024) is a multi-view, multi-modal panoptic scene understanding dataset. It includes third-person panoramic and front views, as well as first-person monocular and binocular views. The dataset also offers audio, location data, and textual scene descriptions. Benchmarks include video scene classification, temporal action localization, and cross-modality retrieval.

F. Video Question Answering. **GazeVQA** (Ilaslan et al., 2023) is designed for task-oriented video question answering. It features collaboration between an instructor and a novice in assembling or disassembling an industrial product. A key feature of GazeVQA (Ilaslan et al., 2023) is the inclusion of egocentric eye gaze information, which aids in understanding human intention.

G. Egocentric Wearer Identification. **Ego2Top** (Ardeshir & Borji, 2016), **IUShareView** (Fan et al., 2017) , and **TF2023** (Zhao et al., 2024) utilize a fixed exocentric camera and multiple egocentric cameras mounted on different individuals in the environment. In IUShareView (Fan et al., 2017) and TF2023 (Zhao et al., 2024), each person is annotated with a unique ID. Additionally, TF2023 (Zhao et al., 2024) provides segmentation masks for individuals in third-person views.

H. Cross-View Human Identification. **CVMHT** (Han et al., 2020b) comprises over 23K frames of top-view and horizontal-view videos from five different locations. Annotations include bounding boxes and cross-view ID numbers for subjects. **DMHA** (Han et al., 2022b) is a synthetic dataset featuring top-view and side-view videos from common outdoor surveillance scenes. Compared to CVMHT (Han et al., 2020b), it also includes the side-view camera's location and view direction in the top-view.

I. Camera Localization. **CSRD-II** (Qian et al., 2022) and **CSRD-V** (Qian et al., 2022) are synthetic datasets for egocentric camera localization. Annotations include subject positions and camera poses in the bird's-eye view. **YOWO** (Yang et al., 2024) is a synthetic dataset for exocentric camera localization. An agent with an egocentric camera traverses the scene, collaborating with ceiling-mounted cameras for scene reconstruction and camera localization.

Table 1 Overview of ego-exo datasets: 'Frames' shows frame statistics. 'Ego/Exo Views' lists viewpoints for synchronized datasets. 'Multi-Activities' indicates varied activities. 'Multi-Agents' denotes interactions among multiple people.

Dataset	Year	Domain	Frames	Exo Views	Ego Views	Multi-Activities	Multi-Agents	Gaze
CMU-MMAC (la Torre Frade et al., 2008)	2008	Action Understanding	0.2M	3	2	✗	✗	✗
Charades-Ego (Sigurdsson et al., 2018)	2018	Action Understanding	7.4M	1	1	✓	✗	✗
LEMMA (Jia et al., 2020)	2020	Action Understanding	4.1M	2	1	✓	✓	✗
H2O (Kwon et al., 2021)	2021	Action Understanding	571K	4	1	✗	✗	✗
HOMAGE (Rai, 2021)	2021	Action Understanding	2.7M	1-4	1	✓	✗	✗
Assembly101 (Sener, 2022)	2022	Action Understanding	110M	8	4	✗	✗	✗
EgoPW (Wang et al., 2022)	2022	Action Understanding	318K	1	1	✓	✗	✗
ARCTIC (Fan et al., 2023)	2023	Action Understanding	2.1M	8	1	✗	✗	✗
FT-HID (Guo et al., 2023)	2023	Action Understanding	6.4M	3	2	✓	✓	✗
EgoHumans (Khirodkar et al., 2023)	2023	Action Understanding	571K	8-15	1	✓	✓	✗

Table 1 (continued)

Dataset	Year	Domain	Frames	Exo Views	Ego Views	Multi-Activities	Multi-Agents	Gaze
AssemblyHands (Ohkawa et al., 2023)	2023	Action Understanding	3.03M	8	4	✗	✗	✗
First2Third-Pose (Dhamanaskar et al., 2023)	2023	Action Understanding	190K	2-3	1	✓	✗	✗
ECHP (Liu et al., 2023)	2023	Action Understanding	75K	2	1	✓	✗	✗
Hein et al. (2023)	2023	Action Understanding	1.7M	5	2	✗	✓	✓
OAKINK2 (Zhan, 2024)	2024	Action Understanding	4.0M	3	1	✓	✗	✗
EgoExo-Fitness (Li et al., 2024)	2024	Action Understanding	3.4M	3	3	✓	✗	✗
CORE4D-Real (Zhang et al., 2024)	2024	Action Understanding	1.4M	4	1	✗	✓	✗
Ego-Exo4D (Grauman, 2024b)	2024	Action Understanding	236.5M	4	1	✓	✗	✓
EgoExoLearn (Huang, 2024)	2024	Action Understanding	10.3M	-	-	✓	✗	✓
ThermoHands (Ding et al., 2024)	2024	Action Understanding	96K	1	1	✓	✗	✗
Nymeria (Ma et al., 2024)	2024	Action Understanding	201M	1	1	✓	✗	✓
OVR (Dwibedi et al., 2024)	2024	Action Understanding	5.4M	-	-	✓	✗	✗
EgoMe (Qiu et al., 2025)	2025	Action Understanding	8.5M	1	1	✓	✗	✓
LBW (Kasahara et al., 2022)	2022	Driving	123K	1	2	✗	✗	✓
AIDE (Yang, 2023)	2023	Driving	521.6K	1	3	✗	✗	✗
WTS (Kong et al., 2024)	2024	Driving	52.8K	18	2	✗	✓	✓
PAD (Luo et al., 2021)	2021	Affordance Grounding	4K	1	1	-	-	-
AGD20K (Luo et al., 2022)	2022	Affordance Grounding	20K	1	1	-	-	-
FAH (Yang et al., 2024)	2024	Affordance Grounding	6K	1	1	-	-	-
Thirdtofirst (Li et al., 2021)	2021	Generation	334.6K	1	1	✓	✗	✗
360+x (Chen et al., 2024)	2024	Scene Understanding	8.5M	2	2	✓	✗	✗
GazeVQA (Ilaslan et al., 2023)	2023	Video Question Answering	12.6M	2	1	✗	✗	✓
Ego2Top (Ardeshir & Borji, 2016)	2016	Egocentric Wearer Identification	225K	1	1-6	✓	✓	✗
IUShareView (Fan et al., 2017)	2017	Egocentric Wearer Identification	11.2K	1	2	✓	✓	✗
TF2023 (Zhao et al., 2024)	2024	Egocentric Wearer Identification	49.8K	1	2	✓	✓	✗
CVMHT (Han et al., 2020b)	2020	Cross-View Human Identification	23K	1	2-3	✓	✓	✗
DMHA (Han et al., 2022b)	2022	Cross-View Human Identification	84.8K	1	1	✗	✓	✗
CSRD-II (Qian et al., 2022)	2022	Camera Registration	2K	1	2	✓	✓	✗
CSRD-V (Qian et al., 2022)	2022	Camera Registration	5K	1	5	✓	✓	✗
YOWO (Yang et al., 2024)	2024	Camera Registration	-	5-17	1	-	-	-

6 Discussion

This section discusses the limitations of current research and offers insights into future directions from the perspectives of data, model, and application.

Insights from Data. Most existing datasets focus on daily life activities, resulting in a scarcity of data tailored to specific scenarios such as public service, healthcare, and education. This limitation hinders the development of approaches for specialized applications. Additionally, most datasets use sophisticated multi-camera setups to record synchronized egocentric and exocentric videos. This significantly increases costs and limits the scalability of data collection. Future research could investigate transforming existing unpaired egocentric (Grauman, 2024a; Damen, 2021) and exocentric (Carreira & Zisserman, 2017; Gu, 2018; Soomro et al., 2012) datasets to enable collaboration between these perspectives. Furthermore, integrating video data with other modalities, such as audio (Jia, 2024) and IMU sensors (Zhang et al., 2025), could enrich the captured information, providing a more comprehensive understanding of complex scenarios.

Insights from Model. Most existing models are designed for specific tasks and lack generalizability. In contrast, recent advancements in vision-language models (VLMs) (Chen, 2024; Chen et al., 2023; Li et al., 2025; Liu et al., 2024) highlight their effectiveness to handle diverse tasks. Future research could explore equipping VLMs with the capability to integrate egocentric and exocentric perspectives, facilitating unified cross-view tasks in a single framework. Moreover, current methods often rely on synchronized egocentric and exocentric data. However, the limited scale of such paired datasets hinders the effective training of large models. To overcome this limitation, promising directions include leveraging alignment strategies or retrieval-augmented methods (Luo et al., 2024) to better utilize unpaired data.

Insights from Application. Current research are primarily centered on daily life contexts, with limited attention to specialized application domains. For instance, while affordance grounding has been well-studied for everyday objects (Li et al., 2023; Luo et al., 2022; Rai et al., 2024; Xu et al., 2024; Zhang et al., 2024), predicting affordance regions for surgical tools or industrial components receives less attention. Extending egocentric and exocentric collaboration techniques to domains such as medicine and industry could unlock new opportunities in these fields.

7 Conclusion

This survey presents a comprehensive review of cross-view collaboration with egocentric and exocentric vision. We begin by discussing the practical value of egocentric and exocentric collaboration across various applications. We then link these applications to key research tasks required to realize them. Current research advancements are categorized into three directions: egocentric for exocentric, exocentric for egocentric, and joint learning, with a detailed overview of progress in each area. In addition, we review relevant datasets that support both perspectives. Finally, we provide a discussion on data, models, and applications, and outline future research directions. We hope this review inspires deeper exploration into egocentric-exocentric collaboration, paving the way for artificial intelligence to perceive the world with human-like vision.

Acknowledgements This work is supported by the National Natural Science Foundation of China (No. 62372223 and U24A20330), and Nanjing University-China Mobile Communications Group Co., Ltd. Joint Institute, and JST ASPIRE Grant Number JPMJAP2303 and JSPS KAKENHI Grant Numbers JP25K24384 and JP24K02956.

Funding Open Access funding provided by The University of Tokyo

References

Abdullah, A., Chen, R., Rekleitis, I., & Islam, M. (2024)"Ego-to-exo: Interfacing third person visuals from egocentric views in real-time for improved rov teleoperation," arXiv:2407.00848.

Acar, C., Binici, K., Tekirdağ, A., & Wu, Y. (2024). Visual-policy learning through multi-camera view to single-camera view knowledge distillation for robot manipulation tasks. *IEEE Transactions on Robotics and Automation, 9*(1), 691–698.

Alahi, A., Goel, K., Ramanathan, V., Robicquet, A., Fei-Fei, L., & Savarese, S. (2016). Social lstm: Human trajectory prediction in crowded spaces. In *Proceedings of the IEEE conference on computer vision and pattern recognition* pp. 961–971.

Ali-Bey, A., Chaib-Draa, B., & Giguere, P. (2023). Mixvpr: Feature mixing for visual place recognition. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision* pp. 2998–3007.

Allsup, M. (2023). "Pg&e kicks off v2x pilot program with ford, sunrun," Latitude Media, 11 2023. [Online]. Available: https://www.latitudemedia.com/news/pg-e-kicks-off-v2x-pilot-program-with-ford-sunrun/

Apple, (2023)."Apple vision pro," Apple. [Online]. Available: https://www.apple.com.cn/apple-vision-pro/

Ardeshir, S., & Borji, A. (2016). Ego2top: Matching viewers in egocentric and top-view videos. In *European conference on computer vision* pp. 253–268.

Ardeshir, S., & Borji, A. (2018). "Integrating egocentric videos in top-view surveillance videos: Joint identification and temporal alignment," In *Proceedings of the European Conference on Computer Vision*.

Ardeshir, S., & Borji, A. (2018). An exocentric look at egocentric actions and vice versa. *Comput Vision Image Understanding,* pp. 61–68.

Ardeshir, S., Regmi, K., & Borji, A. (2016). "Egotransfer: Transferring motion across egocentric and exocentric domains using deep neural networks," arXiv:1612.05836.

Ardeshir, S., & Borji, A. (2019). Egocentric meets top-view. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 41*(6), 1353–1366.

Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., & Schmid, C. (2021). Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision* pp. 6816–6826.

Asia, H. M. (2022) "Here's how asean's first 5g smart hospital is using ai to usher in a new era of healthcare," HMA, 02 2022. [Online]. Available: https://www.hospitalmanagementasia.com/tech-innovation/heres-how-aseans-first-5g-smart-hospital-is-using-ai-to-usher-in-a-new-era-of-healthcare/

Asia, H. M. (2022). "Smart glasses in hospitals: Viewing care delivery through a new lens," HMA, 03. [Online]. Available: https://www.hospitalmanagementasia.com/tech-innovation/smart-glasses-in-hospitals-viewing-care-delivery-through-a-new-lens/

Ataer-Cansizoglu, E., Taguchi, Y., Ramalingam, S., & Miki, Y. (2014). Calibration of non-overlapping cameras using an external slam system. *2014 2nd International Conference on 3D Vision,1*, 509–516.

Bansal, S., Arora, C., & Jawahar, C. (2022). My view is the best view: Procedure learning from egocentric videos. In *European Conference on Computer Vision* Springer, pp. 657–675.

Barron, J. T., Mildenhall, B., Tancik, M., Hedman, P., Martin-Brualla, R., & Srinivasan, P. P. (2021). Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *Proceedings of the IEEE/CVF international conference on computer vision* pp. 5855–5864.

Barron, J. T., Mildenhall, B., Verbin, D., Srinivasan, P. P., & Hedman, P. (2022). Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* pp. 5470–5479.

Begg, R. (2024). "A vision-guided robotic system designed to grab any object,". [Online]. Available: https://www.machinedesign.com/markets/robotics/video/55131589/cynlr-a-vision-guided-robotic-system-designed-to-grab-any-object

Bertasius, G., & Wang, H.,& Torresani, L. (2021). Is space-time attention all you need for video understanding? in *Int. Conf. Mach. Learn.*

Bertasius, G., Soo Park, H., Yu, S. X., & Shi, J. (2017)."Am i a baller? basketball performance assessment from first-person videos," In *Proceedings of the IEEE international conference on computer vision*, pp. 2177–2185.

Berton, G., Masone, C., & Caputo, B. (2022). Rethinking visual geo-localization for large-scale applications. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* pp. 4878–4888.

Bharadhwaj, H., Vakil, J., Sharma, M., Gupta, A., Tulsiani, S., & Kumar, V. (2024). Roboagent: Generalization and efficiency in robot manipulation via semantic augmentations and action chunking. In *In 2024 IEEE International Conference on Robotics and Automation (ICRA)* pp. 4788–4795.

Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., & Rombach, R. (2023). "Stable video diffusion: Scaling latent video diffusion models to large datasets," arXiv:2311.15127.

Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., Dabis, J., Finn, C., & Zitkovich, B. (2022). "Rt-1: Robotics transformer for real-world control at scale," arXiv:2212.06817.

Carreira, J., & Zisserman, A. (2017). Quo vadis, action recognition? a new model and the kinetics dataset. *In proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* pp. 4724–4733.

Cauchard, J. R., Khamis, M., Garcia, J., Kljun, M., & Brock, A. M. (2021). Toward a roadmap for human-drone interaction. *Interactions, 28*(76–81), 03.

Chang, R.-C., Liu, Y., & Guo, A. (2024). Worldscribe: Towards context-aware live visual descriptions. *ACM Symp. User Interface Softw. Technol.,* pp. 1–18.

Chen, G. (2024). *et al.*, "Video mamba suite: State space model as a versatile alternative for video understanding," arXiv: 2403.09626.

Chen, H., Hou, Y., Qu, C., Testini, I., Hong, X., & Jiao, J. (2024). $360 + x$: A panoptic multi-modal scene understanding dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* pp. 19373–19382.

Chen, G., Liu, Y., Huang, Y., He, Y., Pei, B., Xu, J., & Wang, L. (2024). "Cg-bench: Clue-grounded question answering benchmark for long video understanding," 2024, arXiv: 2412.12075.

Chen, J., Lu, K., Tian, S., & Little, J. (2019). Learning sports camera selection from internet videos, in IEEE Winter Conf. *2019 IEEE Winter Conference on Applications of Computer Vision* IEEE, pp. 1682–1691.

Chen, K., Ramanan, D., & Khurana, T. (2025). Using diffusion priors for video amodal segmentation. In *Proceedings of the Computer Vision and Pattern Recognition Conference* pp. 22890–22900.

Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., & Dai, J. (2024). Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* pp. 24185–24198.

Chen, G., Xing, S., Chen, Z., Wang, Y., Li, K., Li, Y.,& Qiao, Y. (2022). "Internvideo-ego4d: A pack of champion solutions to ego4d challenges,", arXiv:2211.09529.

Chen, G., Zheng, Y. D., Wang, J., Xu, J., Huang, Y., Pan, J., & Wang, L. (2023). "Videollm: Modeling video sequence with large language models," arXiv: 2305.13292.

Cheng, F., Luo, M., Wang, H., Dimakis, A., Torresani, L., Bertasius, G., & Grauman, K. (2024)."4diff: 3d-aware diffusion model for third-to-first viewpoint translation," In *European Conference on Computer Vision*.

Chen, H., Zendehdel, N., Leu, M. C., & Yin, Z. (2023). Real-time human-computer interaction using eye gazes. *Manufacturing Letters, 35*, 883–894.

Chen, G., Zheng, Y.-D., Wang, L., & Lu, T. (2022). Dcan: improving temporal action detection via dual context aggregation. *Proceedings of the AAAI conference on artificial intelligence, 36*(1), 248–257.

Chi, C. (2023). Diffusion policy: Visuomotor policy learning via action diffusion, in *Proc. Robot.: Sci. Syst.*

Choi, H., Moon, G., Chang, J. Y., & Lee, K. M. (2021). Beyond static features for temporally consistent 3d human pose and shape from a video. *In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* pp. 1964–1973.

Choi, J., Sharma, G., Chandraker, M., & Huang, J.-B. (2020). Unsupervised and semi-supervised domain adaptation for action recognition from drones. In *IEEE Winter Conf. Appl. Comput. Vis.,* pp. 1717–1726.

Dachman, J. (2017). "Super bowl li preview: Inside fox sports' "be the player" first-person pov replay tech,". [Online]. Available: https://www.sportsvideo.org/2017/01/13/super-bowl-li-preview-inside-fox-sports-be-the-player-360-pov-replay-technology/

Damen, D., et al. (2021). The epic-kitchens dataset: Collection, challenges and baselines. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 43*(11), 4125–4141.

De la Torre, F., Hodgins, J., Bargteil, A., Martin, X., Macey, J., Collado, A., & Beltran, P. (2008)."Guide to the carnegie mellon university multimodal activity (cmu-mmac) database," Carnegie Mellon University, Pittsburgh, PA, Tech. Rep. CMU-RI-TR-08-22

Delitzas, A., Takmaz, A., Tombari, F., Sumner, R., Pollefeys, M., & Engelmann, F. (2024). "SceneFun3D: Fine-Grained Functionality and Affordance Understanding in 3D Scenes," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*

Deng, C., Jiang, C., Qi, C. R., Yan, X., Zhou, Y., Guibas, L., Anguelov, D., et al. (2023). Nerdi: Single-view nerf synthesis with language-guided diffusion as general image priors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* pp. 20637–20647.

Dhamanaskar, A., Dimiccoli, M., Corona, E., Pumarola, A., & Moreno-Noguer, F. (2023). Enhancing egocentric 3d pose estimation with third person views. *Pattern Recognition,138*, Article 109358.

Ding, F., Zhu, Y., Wen, X., Liu, G., & Lu, C. X. (2024). "Thermohands: A benchmark for 3d hand pose estimation from egocentric thermal image," arXiv:2403.09871.

Dou, Z. Y., Yang, X., Nagarajan, T., Wang, H., Huang, J., Peng, N., & Chu, F. J. (2024). "Unlocking exocentric video-language data for egocentric video representation learning," arXiv:2408.03567.

Doughty, H., Mayol-Cuevas, W., & Damen, D. (2019). The pros and cons: Rank-aware temporal attention for skill determination in long videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* pp. 7862–7871.

Dunion, M., & Albrecht, S. V. (2024) "Multi-view disentanglement for reinforcement learning with multiple cameras," arXiv:2404.14064.

Dwibedi, D., Aytar, Y., Tompson, J., & Zisserman, A. (2024). "Ovr: A dataset for open vocabulary temporal repetition counting in videos," arXiv:2407.17085.

Elfeki, M., Regmi, K., Ardeshir, S., & Borji, A. (2018)."From third person to first person: Dataset and baselines for synthesis and retrieval," arXiv:1812.00104.

Erat, O., Isop, W. A., Kalkofen, D., & Schmalstieg, D. (2018). Drone-augmented human vision: Exocentric control for drones exploring hidden areas. *IEEE transactions on visualization and computer graphics, 24*(4), 1437–1446.

Esser, P., Chiu, J., Atighehchian, P., Granskog, J., & Germanidis, A. (2023). Structure and content-guided video synthesis with diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision* pp. 7312–7322.

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 IEEE Conf. Comput. Vis. Pattern Recog.,

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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*

Fan, H., Xiong, B., Mangalam, K., Li, Y., Yan, Z., Malik, J., & Feichtenhofer, C. (2021). Multiscale vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision* pp. 6824–6835.

Fogsphere, (2023)."Empowering health & safety monitoring in manufacturing - fogsphere,". [Online]. Available: https://fogsphere.com/industries-served/manufacturing/

Fu, Y., Wang, R., Fu, Y., Paudel, D. P., Huang, X., & Van Gool, L. (2024). "Objectrelator: Enabling cross-view object relation understanding in ego-centric and exo-centric videos," arXiv:2411.19083.

Fulkerson, B., Vedaldi, A., & Soatto, S. (2009). Class segmentation and object localization with superpixel neighborhoods. In *2009 IEEE 12th international conference on computer vision* pp. 670–677.

Furnari, A., & Farinella, G. M. (2019). What would you expect? anticipating egocentric actions with rolling-unrolling lstms and modality attention. In *Proceedings of the IEEE/CVF International conference on computer vision* pp. 6252–6261.

Fu, Y., Wang, X., Wei, Y., & Huang, T. (2019). Sta: Spatial-temporal attention for large-scale video-based person re-identification. *Proceedings of the AAAI conference on artificial intelligence, 33*(01), 8287–8294.

Garello, L., Rea, F., Noceti, N., & Sciutti, A. (2022). Towards third-person visual imitation learning using generative adversarial networks. In *2022 IEEE International Conference on Development and Learning* pp. 121–126.

Grauman, K., et al. (2024). Ego4d: Around the world in 3,000 hours of egocentric video. *IEEE Trans. Pattern Anal. Mach. Intell.,* pp. 1–32.

Grauman, K., et al. (2024). Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* pp. 19383–19400.

Gu, C., Sun, C., Ross, D. A., Vondrick, C., Pantofaru, C., Li, Y., & Malik, J. (2018). Ava: A video dataset of spatio-temporally localized atomic visual actions. *In Proceedings of the IEEE conference on computer vision and pattern recognition* pp. 6047–6056.

Gu, J., Trevithick, A., Lin, K.-E., Susskind, J. M., Theobalt, C., Liu, L., & Ramamoorthi, R. (2023). Nerfdiff: Single-image view synthesis with nerf-guided distillation from 3d-aware diffusion. *International Conference on Machine Learning* pp. 11808–11826. PMLR,

Guo, Z., Hou, Y., Wang, P., Gao, Z., Xu, M., & Li, W. (2023). Ft-hid: a large-scale rgb-d dataset for first-and third-person human interaction analysis. *Neural Computing and Applications,* pp. 2007–2024.

Gupta, R. (2024). "Hololens in manufacturing: Use cases and future of ar," Kompanions, 08. [Online]. Available: https://www.kompanions.com/blog/hololens-in-manufacturing/

Halim, F., Widysanto, A., Wahjoepramono, P. O. P., Candrawinata, V. S., Budihardja, A. S., Irawan, A., Sudirman, T., Christina, N., Koerniawan, H. S., Tobing, J. F. L., et al. (2024). Objective comparison of the first-person-view live streaming method versus face-to-face teaching method in improving wound suturing skills for skin closure in surgical clerkship students: Randomized controlled trial. *JMIR medical education,10*, Article e52631.

Han, R., Gan, Y., Li, J., Wang, F., Feng, W., & Wang, S. (2022). Connecting the complementary-view videos: Joint camera identification and subject association. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern* pp. 2406–2415.

Han, R., Zhao, J., Feng, W., Gan, Y., Wan, L., & Wang, S. (2020). Complementary-view co-interest person detection. In *Proceedings of the 28th ACM international conference on multimedia* p. 2746–2754.

Han, R., Feng, W., Zhang, Y., Zhao, J., & Wang, S. (2022). Multiple human association and tracking from egocentric and complementary top views. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 44*(9), 5225–5242.

Han, R., Feng, W., Zhao, J., Niu, Z., Zhang, Y., & Wan, L. (2020). Complementary-view multiple human tracking, in AAAI Conf. *Artificial Intelligence, 34*, 02.

Han, R., Gan, Y., Wang, L., Li, N., Feng, W., & Wang, S. (2023). Relating view directions of complementary-view mobile cameras via the human shadow. *International Journal of Computer Vision, 131*(5), 1106–1121.

harkiran78, (2024).“Top 10 applications of robotics in 2024,” geeksforgeeks. [Online]. Available: https://www.geeksforgeeks.org/applications-of-robotics

Haugaard, R. L., & Iversen, T. M. (2023). Multi-view object pose estimation from correspondence distributions and epipolar geometry. *Int. Conf. Robot. Autom.,* pp. 1786–1792.

He, T., et al. (2024). Collaborative weakly supervised video correlation learning for procedure-aware instructional video analysis. *AAAI Conf. Artif. Intell., 38*(3), 2112–2120.

Hein, J., Cavalcanti, N., Suter, D., Zingg, L., Carrillo, F., Calvet, L., & F2023rnstahl, P. (2023). “Next-generation surgical navigation: Marker-less multi-view 6dof pose estimation of surgical instruments,” arXiv:2305.03535.

Hellogard, (2025). “Autonomous wheel chair,” HelloGard.com, 2025. [Online]. Available: https://www.hellogard.com/autonomouswheelchair

Hou, R., Ma, B., Chang, H., Gu, X., Shan, S., & Chen, X. (2019). Vrstc: Occlusion-free video person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* pp. 7183–7192.

Hsu, K., Kim, M. J., Rafailov, R., Wu, J., & Finn, C. (2022) “Vision-based manipulators need to also see from their hands,” arXiv:2203.12677.

Hu, L., Gao, X., Zhang, P., Sun, K., Zhang, B., & Bo, L. (2023). Animate anyone: Consistent and controllable image-to-video synthesis for character animation. *IEEE Conf. Comput. Vis. Pattern Recog.,* pp. 8153–8163.

Hu, H.-N., Lin, Y.-C., Liu, M.-Y., Cheng, H.-T., Chang, Y.-J., & Sun, M. (2017). Deep 360 pilot: Learning a deep agent for piloting through 360 sports videos. In *2017 IEEE Conference on Computer Vision and Pattern Recognition* pp. 1396–1405.

Huang, Y., Bi, H., Li, Z., Mao, T., & Wang, Z. (2019).“Stgat: Modeling spatial-temporal interactions for human trajectory prediction,” In *Proceedings of the IEEE/CVF international conference on computer vision*

Huang, Y., Chen, G., Xu, J., Zhang, M., Yang, L., Pei, B., & Qiao, Y. (2024). Egoexolearn: A dataset for bridging asynchronous ego- and exo-centric view of procedural activities in real world. In *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* pp. 22072–22086.

Huang, D. A., Liao, S., Radhakrishnan, S., Yin, H., Molchanov, P., Yu, Z., & Kautz, J. (2024). Lita: Language instructed temporal-localization assistant, In *European Conference on Computer Vision*

Huang, Y., Xu, J., Pei, B., He, Y., Chen, G., Yang, L., & Wang, L. (2024). “Vinci: A real-time embodied smart assistant based on egocentric vision-language model,” arXiv: 2412.21080.

Huang, Y., Yan, T., Gong, S., Gao, X., Kang, C., Liu, R., Lu, H., & Zheng, B. (2025). “Living the novel: A system for generating self-training timeline-aware conversational agents from novels,” arXiv preprint arXiv:2512.07474.

Huang, Y., Yang, X., Gao, J., & Xu, C. (2022). Holographic feature learning of egocentric-exocentric videos for multi-domain action recognition. *IEEE Transactions on Multimedia, 24*, 2273–2286.

Ilaslan, M., Song, C., Chen, J., Gao, D., Lei, W., Xu, Q., & Shou, M. (2023). Gazevqa: A video question answering dataset for multiview eye-gaze task-oriented collaborations,. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 2023*, 10462–10479.

Inoue, M., Takashima, K., Fujita, K., & Kitamura, Y. (2023) “Birdviewar: Surroundings-aware remote drone piloting using an augmented third-person perspective,” in *Conf Hum Fact Comput Syst Proc*

Islam, M. M., Ho, N., Yang, X., Nagarajan, T., Torresani, L., & Bertasius, G. (2024). Video recap: Recursive captioning of hour-long videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* pp. 18198–18208.

Jang, J. H., Seo, H., & Chun, S. Y. (2024) “Intra: Interaction relationship-aware weakly supervised affordance grounding,” arXiv:2409.06210.

Jangir, R., Hansen, N., Ghosal, S., Jain, M., & Wang, X. (2022). Look closer: Bridging egocentric and third-person views with transformers for robotic manipulation. *IEEE Transactions on Robotics and Automation, 7*(2), 3046–3053.

Jia, B., Chen, Y., Huang, S., Zhu, Y., & Zhu, S.-C. (2020). Lemma: A multi-view dataset for le arning m ulti-agent m ulti-task a ctivities. In *European Conference on Computer Vision* Springer, pp. 767–786.

Jia, W., Liu, M., Jiang, H., Ananthabhotla, I., Rehg, J. M., Ithapu, V. K., & Gao, R. (2024). The audio-visual conversational graph: From an egocentric-exocentric perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* pp. 26386–26395.

Jia, W., Liu, M., Jiang, H., Ananthabhotla, I., Rehg, J. M., Ithapu, V. K., & Gao, R. (2024). The audio-visual conversational graph: From an egocentric-exocentric perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* pp. 26396–26405.

Jiao, L., et al. (2022). New generation deep learning for video object detection: A survey. *IEEE Transactions on Neural Networks and Learning Systems, 33*(8), 3195–3215.

Jobit, (2024). “How visual ai transforms assembly line operations in factories,” [Online]. Available: https://randomwalk.ai/blog/how-visual-ai-transforms-assembly-line-operations-in-factories/

Ju, Y., Hu, K., Zhang, G., Zhang, G., Jiang, M., & Xu, H. (2025). Robo-abc: Affordance generalization beyond categories via semantic correspondence for robot manipulation. In *European Conference on Computer Vision* pp. 222–239.

Kalluri, T., Majumder, B. P., & Chandraker, M. (2024). Tell, don’t show: Language guidance eases transfer across domains in images and videos. *Int. Conf. Mach. Learn.,* pp. 22879–22894.

Kasahara, I., Stent, S., & Park, H. S. (2022). Look both ways: Self-supervising driver gaze estimation and road scene saliency. In *European Conference on Computer Vision* Springer, pp. 126–142.

Ke, L., Tai, Y.-W., & Tang, C.-K. (2021). Occlusion-aware video object inpainting. In *Proceedings of the IEEE/CVF international conference on computer vision* pp. 14468–14478.

Khirodkar, R., Bansal, A., Ma, L., Newcombe, R., Vo, M., & Kitani, K. (2023). Ego-humans: An ego-centric 3d multi-human benchmark. *In Proceedings of the IEEE/CVF International Conference on Computer Vision* pp. 19807–19819.

Kocabas, M., Huang, C.-H.P., Tesch, J., Müller, L., Hilliges, O., & Black, M. J. (2021). Spec: Seeing people in the wild with an estimated camera. *In Proceedings of the IEEE/CVF International Conference on Computer Vision* pp. 11035–11045.

Kong, Q., Kawana, Y., Saini, R., Kumar, A., Pan, J., Gu, T., & Kobori, N. (2024). “Wts: A pedestrian-centric traffic video dataset for fine-grained spatial-temporal understanding,” arXiv:2407.15350.

Kratz, L., & Nishino, K. (2009). Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models. *IEEE conference on computer vision and pattern recognition, IEEE, 2009*, 1446–1453.

Kuribayashi, M., Uehara, K., Wang, A., Morishima, S., & Asakawa, C. (2025) “Wanderguide: Indoor map-less robotic guide for exploration by blind people,” arXiv:2502.08906.

Kwon, T., Tekin, B., Stühmer, J., Bogo, F., & Pollefeys, M. (2021). H2o: Two hands manipulating objects for first person interaction recognition. In *Proceedings of the IEEE/CVF international conference on computer vision* pp. 10118–10128.

Kyrkou, C., & Theocharides, T. (2020). Emergencynet: Efficient aerial image classification for drone-based emergency monitoring using atrous convolutional feature fusion. IEEE J. *IEEE Journal of*

Selected Topics in Applied Earth Observations and Remote Sensing, 13, 1687–1699.

Lavars, N. (2016). "Samsung's new smart fridge lets you check in on its contents through internal cameras," . [Online]. Available: https://newatlas.com/samsung-family-hub-smart-fridge/41192/

Lee, G., Healey, J., & Manocha, D. (2022). Vrdoc: Gaze-based interactions for vr reading experience, . IEEE International Symposium on Mixed and Augmented Reality (ISMAR). *IEEE,2022*, 787–796.

Lee, S.-P., Lu, Z., Zhang, Z., Hoai, M., & Elhamifar, E. (2024). Error detection in egocentric procedural task videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* pp. 18655–18666.

Li, J., Balakrishnan, R., & Grossman, T. (2020). "Starhopper: A touch interface for remote object-centric drone navigation," in *Proc Graphics Interface*, ser. GI 2020, pp. 317 – 326.

Li, Z., Chen, G., Liu, S., Wang, S., V. S., V., Ji, Y., & Yu, Z. (2025)."Eagle 2: Building post-training data strategies from scratch for frontier vision-language models," arXiv: 2501.14818.

Li, Y., Fathi, A., & Rehg, J. M. (2013). Learning to predict gaze in egocentric video. In *Proceedings of the IEEE international conference on computer vision* pp. 3216–3223.

Li, Y.-M., Huang, W.-J., Wang, A.-L., Zeng, L.-A., Meng, J.-K., & Zheng, W.-S. (2024). "Egoexo-fitness: Towards egocentric and exocentric full-body action understanding,", arXiv:2406.08877.

Li, G., Jampani, V., Sun, D., & Sevilla-Lara, L. (2023). Locate: Localize and transfer object parts for weakly supervised affordance grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* pp. 10922–10931.

Li, Y., Nagarajan, T., Xiong, B., & Grauman, K. (2021). Ego-exo: Transferring visual representations from third-person to first-person videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* pp. 6943–6953.

Li, Y., Wu, C.-Y., Fan, H., Mangalam, K., Xiong, B., Malik, J., & Feichtenhofer, C. (2022). Mvitv 2: Improved multiscale vision transformers for classification and detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* pp. 4804–4814.

Li, G., Zhao, K., Zhang, S., Lyu, X., Dusmanu, M., Zhang, Y., Pollefeys, M., & Tang, S. (2024). Egogen: An egocentric synthetic data generator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* pp. 14497–14509.

Lin, K. Q., et al. (2022). Egocentric video-language pretraining. *Adv. Neural Inform. Process. Syst., 35*, 7575–7586.

Liu, H., et al. (2022). Video super-resolution based on deep learning: a comprehensive survey. *Artif Intell Rev,* pp. 5981–6035.

Liu, G., Latapie, H., Kilic, O., & Lawrence, A. (2022). Parallel generative adversarial network for third-person to first-person image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* pp. 1917–1923.

Liu, H., Li, C., Wu, Q., & Lee, Y. J. (2024). "Visual instruction tuning," *Adv. Neural Inform. Process. Syst.*, vol. 36

Liu, G., Tang, H., Latapie, H. M., Corso, J. J., & Yan, Y. (2021). Cross-view exocentric to egocentric video synthesis. In *Proceedings of the 29th ACM International Conference on Multimedia* pp. 974–982.

Liu, Y., Zhang, C., Xing, R., Tang, B., Yang, B., & Yi, L. (2024). "Core4d: A 4d human-object-human interaction dataset for collaborative object rearrangement," arXiv:2406.19353.

Liu, J.-W., Mao, W., Xu, Z., Keppo, J., & Shou, M. Z. (2024). Exocentric-to-egocentric video generation. *Advances in Neural Information Processing Systems, 37*, 136149–136172.

Liu, S., Ren, Z., & Yuan, J. (2018). Sibnet: Sibling convolutional encoder for video captioning. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 43*, 3259–3272.

Liu, Y., Yang, J., Gu, X., Chen, Y., Guo, Y., & Yang, G.-Z. (2023). Egofish3d: Egocentric 3d pose estimation from a fisheye camera via self-supervised learning. *IEEE Trans. Multimedia, 25*, 8880–8891.

Lu, Z., & Elhamifar, E. (2024). Fact: Frame-action cross-attention temporal modeling for efficient action segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,* pp. 18175–18185.

Luiten, J., Kopanas, G., Leibe, B., & Ramanan, D. (2024)"Dynamic 3d gaussians: Tracking by persistent dynamic view synthesis," in *2024 International Conference on 3D Vision (3DV)*. IEEE, 2024, pp. 800–809.

Luo, M., Xue, Z., Dimakis, A., & Grauman, K. (2025). Put myself in your shoes: Lifting the egocentric perspective from exocentric videos. In *European Conference on Computer Vision*, Springer pp. 407–425.

Luo, M., Xue, Z., Dimakis, A., & Grauman, K. (2025). Viewpoint rosetta stone: Unlocking unpaired ego-exo videos for view-invariant representation learning. In *Proceedings of the Computer Vision and Pattern Recognition Conference* pp. 15802–15812.

Luo, H., Zhai, W., Zhang, J., Cao, Y., & Tao, D. (2021). "One-shot affordance detection," arXiv:2106.14747.

Luo, H., Zhai, W., Zhang, J., Cao, Y., & Tao, D. (2022). Learning affordance grounding from exocentric images. In*Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* pp. 2252–2261.

Luo, Y., Zheng, X., Li, G., Yin, S., Lin, H., Fu, C., & Ji, R. (2024). "Video-rag: Visually-aligned retrieval-augmented long video comprehension," arXiv:2411.13093.

Luo, H., Zhu, K., Zhai, W.,& Cao, Y. (2024). "Intention-driven ego-to-exo video generation," arXiv:2403.09194.

Ma, L., Ye, Y., Hong, F., Guzov, V., Jiang, Y., Postyeni, R., & Newcombe, R. (2024). Nymeria: A massive collection of multimodal egocentric daily motion in the wild, in *European Conference on Computer Vision*

Majumder, S., Nagarajan, T., Al-Halah, Z., & Grauman, K. (2024)"Switch-a-view: Few-shot view selection learned from edited videos," arXiv: 2412.18386.

Majumder, S., Nagarajan, T., Al-Halah, Z., Pradhan, R., & Grauman, K. (2024). "Which viewpoint shows it best? language for weakly supervising view selection in multi-view videos," arXiv: 2411.08753.

McDonald, J. (2013). "Sportvu stats can be helpful, overwhelming," Nov 2013. [Online]. Available: https://www.expressnews.com/sports/spurs/article/SportVU-stats-can-be-helpful-overwhelming-4993731.php

Microsoft News (2019). "Hololens 'astounds' workers on toyota's factory floor in japan," Microsoft Stories Asia, 04. [Online]. Available: https://news.microsoft.com/apac/features/hololens-astounds-workers-on-toyotas-factory-floor-in-japan/

Microsoft, (2025)."Microsoft hololens," learn.microsoft.com. [Online]. Available: https://learn.microsoft.com/en-us/hololens/

Miech, A., Zhukov, D., Alayrac, J.-B., Tapaswi, M., Laptev, I., & Sivic, J. (2019). Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In Proceedings of the IEEE/CVF international conference on computer vision pp. 2630–2640.

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.

Morando, L., & Loianno, G. (2024). Spatial assisted human-drone collaborative navigation and interaction through immersive mixed reality. In *2024 IEEE International Conference on Robotics and Automation (ICRA)* pp. 8707–8713.

Munusamy, H., & Sekhar, C. C. (2020). Domain-specific semantics guided approach to video captioning. In *Proceedings of the*

IEEE/CVF winter conference on applications of computer vision pp. 1576–1585.

Musulin, K. (2019). "Ford to deploy c-v2x tech in all new vehicles in 2022," Smart Cities Dive. [Online]. Available: https://www.smartcitiesdive.com/news/ford-cv2x-tech-new-vehicles-2022/545412/

Nakashima, K., Iwashita, Y., & Kurazume, R. (2020). Lifelogging caption generation via fourth-person vision in a human-robot symbiotic environment. *Robomech J.,7*.

Nakashima, K., Iwashita, Y., Kawamura, A., & Kurazume, R. (2018). Fourth-person captioning: Describing daily events by unisupervised and tri-regularized training, in IEEE Trans. Syst. *Man, Cybern.,* pp. 2122–2127.

Neuwirth, L. S., & Ros, M. (2021). Comparisons between first person point-of-view 180 video virtual reality head-mounted display and 3d video computer display in teaching undergraduate neuroscience students stereotaxic surgeries. *Frontiers in Virtual Reality,2*, Article 706653.

Nishimura, M., Nobuhara, S., & Nishino, K. (2023)."Incrowdformer: On-ground pedestrian world model from egocentric views," arXiv:2303.09534.

Nishimura, M., Nobuhara, S., & Nishino, K. (2023). Viewbirdiformer: Learning to recover ground-plane crowd trajectories and ego-motion from a single ego-centric view. *IEEE Transactions on Robotics and Automation, 8*(1), 368–375.

Nishimura, M., Nobuhara, S., & Nishino, K. (2023). View birdification in the crowd: Ground-plane localization from perceived movements. *International Journal of Computer Vision, 131*(8), 2015–2031.

of Justice, N. I. (2022). "Research on body-worn cameras and law enforcement," National Institute of Justice. [Online]. Available: https://nij.ojp.gov/topics/articles/research-body-worn-cameras-and-law-enforcement

Ohkawa, T., He, K., Sener, F., Hodan, T., Tran, L., & Keskin, C. (2023). Assemblyhands: Towards egocentric activity understanding via 3d hand pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* pp. 12999–13008.

Ohkawa, T., Yagi, T., Nishimura, T., Furuta, R., Hashimoto, A., Ushiku, Y., & Sato, Y. (2023). "Exo2egodvc: Dense video captioning of egocentric procedural activities using web instructional videos," , arXiv:2311.16444.

Oven, J. (2018). "June oven,". [Online]. Available: https://firewireblog.com/2018/08/19/june-oven/

Pan, B., Cai, H., Huang, D. A., Lee, K. H., Gaidon, A., Adeli, E., & Niebles, J. C. (2020). Spatio-temporal graph for video captioning with knowledge distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* pp. 10867–10876.

Pei, B., Chen, G., Xu, J., He, Y., Liu, Y., Pan, K., & Qiao, Y. (2024). "Egovideo: Exploring egocentric foundation model and downstream adaptation,", arXiv:2406.18070.

Pei, B., Huang, Y., Xu, J., Chen, G., He, Y., Yang, L., & Wang, L. (2025). "Modeling fine-grained hand-object dynamics for egocentric video representation learning,", arXiv:2503.00986.

Perrett, T., Darkhalil, A., Sinha, S., Emara, O., Pollard, S., Parida, K. K., & Damen, D. (2025). Hd-epic: A highly-detailed egocentric video dataset, In Proceedings of the Computer Vision and Pattern Recognition Conference.

Plizzari, C., Goletto, G., Furnari, A., Bansal, S., Ragusa, F., Farinella, G. M., & Tommasi, T. (2024). An outlook into the future of egocentric vision. In *International Journal of Computer Vision,* pp. 1–57.

Pramanick, S., Song, Y., Nag, S., Lin, K. Q., Shah, H., Shou, M. Z., & Zhang, P. (2023). Egovlpv2: Egocentric video-language pre-training with fusion in the backbone. *Proceedings of the IEEE/CVF International Conference on Computer Vision* pp. 5262–5274.

Pumarola, A., Corona, E., Pons-Moll, G., & Moreno-Noguer, F. (2021). D-nerf: Neural radiance fields for dynamic scenes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* pp. 10318–10327.

Qian, Z., Han, R., Feng, W., Wang, F. F., & Wang, S. (2022). From a bird's eye view to see: Joint camera and subject registration without the camera calibration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* pp. 863–873.

Qian, Y., Luo, W., Lian, D., Tang, X., Zhao, P., & Gao, S. (2022). Svip: Sequence verification for procedures in videos. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* pp. 19858–19870.

Qiu, H., Shi, Z., Wang, L., Xiong, H., Li, X., & Li, H. (2025). "Egome: Follow me via egocentric view in real world," arXiv: 2501.19061.

Qiu, L., Zhang, X., Li, Y., Li, G., Wu, X., Xiong, Z., Han, X., & Cui, S. (2020). Peeking into occluded joints: A novel framework for crowd pose estimation. In *European conference on computer vision. Springer,* pp. 488–504.

Quattrocchi, C., Furnari, A., Di Mauro, D., Giuffrida, M. V., & Farinella, G. M. (2024)."Synchronization is all you need: Exocentric-to-egocentric transfer for temporal action segmentation with unlabeled synchronized video pairs," 2023, arXiv:2312.02638.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., & Sutskever, I. (2021). "Learning transferable visual models from natural language supervision," In *International conference on machine learning*

Rahman, K. (2024). "Cameras could be installed in classrooms in these states,". [Online]. Available: https://www.newsweek.com/cameras-installed-classrooms-1859098

Rai, A., Buettner, K., & Kovashka, A. (2024). Strategies to leverage foundational model knowledge in object affordance grounding, In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* pp. 1714–1723.

Rai, N., Chen, H., Ji, J., Desai, R., Kozuka, K., Ishizaka, S., & Niebles, J. C. (2021) Home action genome: Cooperative compositional action understanding. *IEEE Conf. Comput. Vis. Pattern Recog.,* pp. 11179–11188.

Ramachandra, B., Jones, M. J., & Vatsavai, R. R. (2022). A survey of single-scene video anomaly detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence,44*(5), 2293–2312.

Rao, J., Wu, H., Jiang, H., Zhang, Y., Wang, Y., & Xie, W. (2024)."Towards universal soccer video understanding," arXiv:2412.01820.

Redmon, J., Divvala, S. K., Girshick, R. B., & Farhadi, A. (2015). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* pp. 779–788.

Reilly, D., Govind, M. K., Xue, L., & Das, S. (2025). "From my view to yours: Ego-augmented learning in large vision language models for understanding exocentric daily living activities," arXiv:2501.05711.

Ren, S., He, K., Girshick, R. B., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 39*, 1137–1149.

Reolink. (2024). "Classroom camera: Transform education,". [Online]. Available: https://reolink.com/blog/classroom-camera

Rizzolatti, G., & Craighero, L. (2004). The mirror-neuron system. *Annual Review of Neuroscience, 27*, 169–92. 02.

Rocha, B., Moreno, P., & Bernardino, A. (2023). Cross-view generalisation in action recognition: Feature design for transitioning from exocentric to egocentric views. In *Iberian Robotics Conference,* pp. 155–166.

Saito, Y., Hachiuma, R., Saito, H., Kajita, H., Takatsume, Y., & Hayashida, T. (2021). Camera selection for occlusion-less surgery

recording via training with an egocentric camera. *IEEE Access, 9*, 138307–138322.

Sarfraz, M. S., Murray, N., Sharma, V., Diba, A., Gool, L. V., & Stiefelhagen, R. (2021). Temporally-weighted hierarchical clustering for unsupervised action segmentation. *IEEE Conf. Comput. Vis. Pattern Recog.,* pp. 11220–11229.

Seminara, L., Farinella, G. M., & Furnari, A. (2024). "Differentiable task graph learning: Procedural activity representation and online mistake detection from egocentric videos," arXiv preprint arXiv:2406.01486

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* pp. 21096–21106.

Seo, Y., Kim, J., James, S., Lee, K., Shin, J., & Abbeel, P. (2023) "Multi-view masked world models for visual robotic manipulation," In *nternational Conference on Machine Learning*

Seo, Y., Kim, J., James, S., Lee, K., Shin, J., & Abbeel, P. (Jul2023). Multi-view masked world models for visual robotic manipulation. In *nternational Conference on Machine Learning202*(23–29), 30613–30632.

Shang, J., & Ryoo, M. S. (2021). Self-supervised disentangled representation learning for third-person imitation learning. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* pp. 214–221.

Shao, R., Zheng, Z., Tu, H., Liu, B., Zhang, H., & Liu, Y. (2023). Tensor4d: Efficient neural 4d decomposition for high-fidelity dynamic reconstruction and rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* pp. 16632–16642.

Sharma, P., Pathak, D., & Gupta, A. (2019)."Third-person visual imitation learning via decoupled hierarchical controller," In *Advances in Neural Information Processing Systems*

Shiota, T., Takagi, M., Kumagai, K., Seshimo, H., & Aono, Y. (2024). Egocentric action recognition by capturing hand-object contact and object state. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* pp. 6541–6551.

Sigurdsson, G. A., Varol, G., Wang, X., Farhadi, A., Laptev, I., & Gupta, A. (2016) "Hollywood in homes: Crowdsourcing data collection for activity understanding," in *European conference on computer vision*,

Sigurdsson, G. A., Gupta, A., Schmid,C., Farhadi, A., & Alahari, K. (2018)."Actor and observer: Joint modeling of first and third-person videos," in *proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7396–7404.

Simonyan, K.,& Zisserman, A. (2014)."Two-stream convolutional networks for action recognition in videos," in *Adv. Neural Inform. Process. Syst.* Cambridge, MA, USA: MIT Press, 2014, p. 568–576.

Singh, I. (2023). "Over 1,000 people saved with drone search and rescue: Dji," DroneDJ. [Online]. Available: https://dronedj.com/2023/07/12/dji-drone-search-rescue-map/

Song, Y., Byrne, E., Nagarajan, T., Wang, H., Martin, M., & Torresani, L. (2024). *Ego4d goal-step: Toward hierarchical understanding of procedural activities* Adv: Neural Inform. Process. Syst., vol. 36

Song, Y., Sun, P., Jin, P., Ren, Y., Zheng, Y., Li, Z., & Gu, J. (2024). "Learning 6-dof fine-grained grasp detection based on part affordance grounding," arXiv: 2301.11564.

Song, L., Chen, A., Li, Z., Chen, Z., Chen, L., Yuan, J., Xu, Y., & Geiger, A. (2023). Nerfplayer: A streamable dynamic scene representation with decomposed neural radiance fields. *IEEE Transactions on Visualization and Computer Graphic, 29*(5), 2732–2742.

Soomro, K., Zamir, A. R., & Shah, M. (2012). "Ucf101: A dataset of 101 human actions classes from videos in the wild," arXiv:1212.0402.

Soran, B., Farhadi, A., & Shapiro, L. (2014)"Action recognition in the presence of one egocentric and multiple static cameras," in *Lect. Notes Comput. Sci.*

Spisak, J., Kerzel, M., & Wermter, S. (2024). "Diffusing in someone else's shoes: Robotic perspective taking with diffusion," arxiv.org/abs/2404.07735

Stergiou, A., & Poppe, R. (2024). "About time: Advances, challenges, and outlooks of action understanding," arXiv:2411.15106.

Su, Y.-C., & Grauman, K. (2017). Making 360 video watchable in 2d: Learning videography for click free viewing. In *2017 IEEE Conference on Computer Vision and Pattern Recognition* pp. 1368–1376.

Temma, R., Takashima, K., Fujita, K., Sueda, K., & Kitamura, Y. (2019). Third-person piloting: Increasing situational awareness using a spatially coupled second drone. *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology* pp. 507–519.

Thapar, D., Arora, C., & Nigam, A. (2020). Is sharing of egocentric video giving away your biometric signature? In *European Conference on Computer Vision* pp. 399–416.

Thapar, D., Nigam, A., & Arora, C. (2020)"Recognizing camera wearer from hand gestures in egocentric videos," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 2095–2103.

Thatipelli, A., Lo, S.-Y., & Roy-Chowdhury, A. K. (2025). "Egocentric and exocentric methods: A short survey," In *Computer Vision and Image Understanding*, p. 104371.

Tran, M., Kim, Y., Su, C.-C., Kuo, C.-H., Sun, M., & Soleymani, M. (2025). Ex2eg-mae: A framework for adaptation of exocentric video masked autoencoders for egocentric social role understanding. In *European Conference on Computer Vision* Springer, pp. 1–19.

Truong, T.-D. & Luu, K. (2023). "Cross-view action recognition understanding from exocentric to egocentric perspective," arXiv:2305.15699.

Tsutsui, S., Fu, Y., & Crandall, D. J. (2021). Whose hand is this? person identification from egocentric hand gestures. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* pp. 3399–3408.

Vertegaal, R. (1999). The gaze groupware system: mediating joint attention in multiparty communication and collaboration. *Proceedings of the SIGCHI conference on Human Factors in Computing Systems,* pp. 294–301.

Wan, X., Chen, Z., & Zhao, X. (2024). "Rsb-pose: Robust short-baseline binocular 3d human pose estimation with occlusion handling," In *IEEE Transactions on Image Processing*

Wang, J., Chen, G., Huang, Y., Wang, L., & Lu, T. (2023). Memory-and-anticipation transformer for online action understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* pp. 13824–13835.

Wang, Y., He, Y., Li, Y., Li, K., Yu, J., Ma, X., & Qiao, Y. (2023)."Internvid: A large-scale video-text dataset for multimodal understanding and generation," 2023, arXiv: 2307.06942.

Wang, X., Kwon, T., Rad, M., Pan, B., Chakraborty, I., Andrist, S., Bohus, D., Feniello, A., Tekin, B., Frujeri, F. V., et al. (2023). Holoassist: an egocentric human interaction dataset for interactive ai assistants in the real world. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* pp. 20270–20281.

Wang, Z., Li, Y. H., Li, X., Zang, H., Laroche, R., & Islam, R. (2025) "Learning fused state representations for control from multi-view observations," arXiv: 2502.01316.

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. *IEEE Conf. Comput. Vis. Pattern Recog.,* pp. 13157–13166.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* pp. 13157–13166.

Wang, Q., Zhao, L., Yuan, L., Liu, T., & Peng, X. (2023). Learning from semantic alignment between unpaired multiviews for egocentric video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* pp. 3284–3294.

Wang, C., Yu, C., Xu, X., Gao, Y., Yang, X., Tang, W., Yu, S., Chen, Y., Gao, F., Jian, Z. ,*et al.* (2025). "Multi-robot system for cooperative exploration in unknown environments: A survey," arXiv preprint arXiv:2503.07278.

Wen, Y., Singh, K. K., Anderson, M., Jan, W.-P., & Lee, Y. J. (2021). Seeing the unseen: Predicting the first-person camera wearer's location and pose in third-person scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* pp. 3446–3455.

Winter, D. (2024). "Hawk-eye's eagle eye on wimbledon tennis," 2024. [Online]. Available: https://www.redsharknews.com/hawk-eyes-eye-on-wimbledon

Wu, G., Yi, T., Fang, J., Xie, L., Zhang, X., Wei, W., Liu, W., Tian, Q., & Wang, X. (2024). 4d gaussian splatting for real-time dynamic scene rendering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* pp. 20310–20320.

Xu, M., Fan, C., Wang, Y., Ryoo, M. S., & Crandall, D. J. (2018). Joint person segmentation and identification in synchronized first- and third-person videos. In *Proceedings of the European Conference on Computer Vision* pp. 656–672.

Xu, J., Huang, Y., Hou, J., Chen, G., Zhang, Y., Feng, R., & Xie, W. (2024). Retrieval-augmented egocentric video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* pp. 13525–13536.

Xu, Y., Li, Y. L., Huang, Z., Liu, M. X., Lu, C., Tai, Y. W., & Tang, C. K. (2023). Egopca: A new framework for egocentric hand-object interaction understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* pp. 5273–5284.

Xu, B., Zheng, S., & Jin, Q. (2023). Pov: Prompt-oriented view-agnostic learning for egocentric hand-object interaction in the multi-view world. *Proceedings of the 31st ACM International Conference on Multimedia* pp. 2807–2816.

Xue, Z. S., & Grauman, K. (2023). Learning fine-grained view-invariant representations from unpaired ego-exo videos via temporal alignment. *Advances in Neural Information Processing Systems, 36*(55), 53688–53710.

Xu, L., Gao, Y., Song, W., & Hao, A. (2024). Weakly supervised multimodal affordance grounding for egocentric images. *Proceedings of the AAAI Conference on Artificial Intelligence, 38*(6), 6324–6332.

Yang, F., Chen, W., Yang, K., Lin, H., Luo, D., Tang, C., & Wang, Y. (2024)."Learning granularity-aware affordances from human-object interaction for tool-based functional grasping in dexterous robotics," arXiv:2407.00614.

Yang, D., Huang, S., Xu, Z., Li, Z., Wang, S., Li, M., & Zhang, L. (2023). Aide: A vision-driven multi-view, multi-modal, multi-tasking dataset for assistive driving perception. In*Proceedings of the IEEE/CVF International Conference on Computer Vision* pp. 20402–20413.

Yang, L., Jiang, H., Huo, Z., & Xiao, J. (2019). Visual-gps: Ego-downward and ambient video based person location association, in IEEE Conf. *Comput. Vis. Pattern Recog Workshops.,* pp. 371–380.

Yang, J., Liu, S., Guo, H., Dong, Y., Zhang, X., Zhang, S., Wang, P., Zhou, Z., Xie, B., Wang, Z., et al. (2025). Egolife: Towards egocentric life assistant. *Proceedings of the Computer Vision and Pattern Recognition Conference,* pp. 28885–28900.

Yang, Z., Teng, J., Zheng, W., Ding, M., Huang, S., Xu, J., & Tang, J. (2024). "Cogvideox: Text-to-video diffusion models with an expert transformer," arXiv:2408.06072.

Yang, F., Yamao, S., Kusajima, I., Moteki, A., Masui, S., & Jiang, S. (2024). Yowo: You only walk once to jointly map an indoor scene and register ceiling-mounted cameras. *IEEE Transactions on Circuits and Systems for Video Technology* pp. 1–1.

Yang, Y., Zhai, W., Wang, C., Yu, C., Cao, Y., & Zha, Z.-J. (2025). Egochoir: Capturing 3d human-object interaction regions from egocentric views. *Advances in Neural Information Processing Systems, 37*(55), 54529–54557.

Ye, H., Zhang, H., Daxberger, E., Chen, L., Lin, Z., Li, Y., & Yang, Y. (2024). "Mm-ego: Towards building egocentric multimodal llms for video qa," arXiv preprint arXiv:2410.07177.

Yi, X., Zhou, Y., Habermann, M., Golyanik, V., Pan, S., Theobalt, C., & Xu, F (2023). Egolocate: Real-time motion capture, localization, and mapping with sparse body-mounted sensors. *ACM Transactions on Graphics,42*(4).

Yin, S., Wu, C., Liang, J., Shi, J., Li, H., Ming, G., & Duan, N. (2023). "Dragnuwa: Fine-grained control in video generation by integrating text, image, and trajectory," arXiv:2308.08089.

Yonetani, R., Kitani, K. M., & Sato, Y. (2016). Recognizing micro-actions and reactions from paired egocentric videos. *IEEE Conf. Comput. Vis. Pattern Recog.,* pp. 2629–2638.

Yu, H., Cai, M., Liu, Y., & Lu, F. (2019). What i see is what you see: Joint attention learning for first and third person video co-analysis. In *Proceedings of the 27th ACM International Conference on Multimedia,* pp. 1358–1366.

Yu, A., Ye, V., Tancik, M., & Kanazawa, A. (2021). pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* pp. 4576–4585.

Yu, H., Cai, M., Liu, Y., & Lu, F. (2023). First- and third-person video co-analysis by learning spatial-temporal joint attention. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 45*(6), 6631–6646.

Yue, J., Manocha, D., & Wang, H. (2022). Human trajectory prediction via neural social physics. In *European conference on computer vision* pp. 376–394.

Yus, R., Mena, E., Ilarri, S., Illarramendi, A., & Bernad, J. (2015). Multicamba: a system for selecting camera views in live broadcasting of sport events using a dynamic 3d model. *Multimedia Tools and Applications, 74*, 4059–4090.

Zhan, X., Yang, L., Zhao, Y., Mao, K., Xu, H., Lin, Z., & Lu, C. (2024). Oakink2: A dataset of bimanual hands-object manipulation in complex task completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* pp. 445–456.

Zhan, A., Zhao, R., Pinto, L., Abbeel, P., & Laskin, M. (2022). Learning visual robotic control efficiently with contrastive pre-training and data augmentation. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* pp. 4040–4047.

Zhang, Y., Doughty, H., Shao, L., & Snoek, C. G. (2022). Audio-adaptive activity recognition across video domains. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* pp. 13791–13800.

Zhang, C., Gupta, A., & Zisserman, A. (2023). Helping hands: An object-aware ego-centric video recognition model. *In Proceedings of the IEEE/CVF International Conference on Computer Vision* pp. 13855–13866.

Zhang, M., Huang, Y., Liu, R., & Sato, Y. (2025). Masked video and body-worn imu autoencoder for egocentric action recognition. In *European Conference on Computer Vision* pp. 312–330.

Zhang, Z., Ma, Y., Zhang, E., & Bai, X. (2025). Psalm: Pixelwise segmentation with large multi-modal model. In *European Conference on Computer Vision* pp. 74–91.

Zhang, Z., Wei, Z., Sun, G., Wang, P., & Van Gool, L. (2024). "Self-explainable affordance learning with embodied caption," arXiv:2404.05603.

Zhang, S., Yang, J., Yin, J., Luo, Z., & Luan, J. (2025)."Q-frame: Query-aware frame selection and multi-resolution adaptation for video-llms," arXiv preprint arXiv:2506.22139, 2025.

Zhao, T. Z., Kumar, V., Levine, S., & Finn, C. (2023). "Learning Fine-Grained Bimanual Manipulation with Low-Cost Hardware," in *Proc. Robot.: Sci. Syst.*, Daegu, Republic of Korea

Zhao, Y., Shen, X., Jin, Z., Lu, H., & Hua, X.-S. (2019). Attribute-driven feature disentangling and temporal aggregation for video person re-identification. *IEEE Conf. Comput. Vis. Pattern Recog.,* pp. 4913–4922.

Zhao, T. Z., Tompson, J., Driess, D., Florence, P., Ghasemipour, K., Finn, C., & Wahid, A. (2024). "Aloha unleashed: A simple recipe for robot dexterity," arXiv:2410.13126.

Zhao, Z., Wang, Y., & Wang, C. (2024). Fusing personal and environmental cues for identification and segmentation of first-person camera wearers in third-person views. *IEEE Conf. Comput. Vis. Pattern Recog.,* pp. 16477–16487.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., & Torralba, A. (2016). Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition* pp. 2921–2929.

Zhu, W., Wang, Y., Li, H., Zhu, P., & Hu, Q. (2025)."Vtd-clip: Video-to-text discretization via prompting clip," arXiv preprint arXiv:2503.18407

Zhu, S., Yang, L., Chen, C., Shah, M., Shen, X., & Wang, H. (2023). R2former: Unified retrieval and reranking transformer for place recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* pp. 19370–19380.